

MME 2026

**Workshop on Multilingual and Multicultural Evaluation**

**Proceedings of the Workshop**

March 28, 2026

©2026 Association for Computational Linguistics

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)  
317 Sidney Baker St. S  
Suite 400 - 134  
Kerrville, TX 78028  
USA  
Tel: +1-855-225-1962  
[acl@aclweb.org](mailto:acl@aclweb.org)

ISBN 979-8-89176-368-5

## Introduction

The First Workshop on Multilingual and Multicultural Evaluation (MME 2026) will be held on March 28, 2026, in Rabat, Morocco, co-located with the 19th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2026).

While large language models (LLMs) are reshaping language technology across language and cultural barriers, the community lacks a comprehensive understanding of their performance across languages and cultural contexts. Simple approaches such as counting languages in training data or relying on translated evaluation benchmarks are not sufficient.

The MME workshop is established to tackle these issues by creating a dedicated venue for multilingual evaluation resources and metrics, standardizing practices to ensure accuracy and fairness, and integrating cultural and social perspectives into NLP and LLM evaluation.

We received 27 archival and 20 non-archival submissions. Archival papers underwent double-blind peer review, while non-archival submissions received a lightweight relevance check. The program committee accepted 15 archival papers for publication in the ACL Anthology, representing a 55.6% acceptance rate.

In total, the workshop comprises 43 presentations, including the 15 archival papers, 19 accepted non-archival submissions, and 9 papers accepted at other venues. The workshop also features 3 invited talks from experts in academia and industry.

We thank all authors for their scholarly contributions and the program committee for their reviewing efforts. We extend our gratitude to the EACL 2026 Organizing Committee, particularly the Workshop Chairs, for their support. Finally, we thank all participants for making the inaugural MME workshop a success.

MME 2026 Organizers

# Organizing Committee

## Organizers

Pinzhen Chen (Queen's University Belfast)

Vilém Zouhar (ETH Zurich)

Hanxu Hu (University of Zurich)

Simran Khanuja (Carnegie Mellon University)

Wenhao Zhu (ByteDance)

Barry Haddow (University of Edinburgh)

Alexandra Birch (University of Edinburgh)

Alham Fikri Aji (Mohamed bin Zayed University of Artificial Intelligence)

Rico Sennrich (University of Zurich)

Sara Hooker (Adaptable Intelligence)

# Program Committee

## Program Chairs

Pinzhen Chen (Queen's University Belfast)  
Vilém Zouhar (ETH Zurich)  
Hanxu Hu (University of Zurich)  
Simran Khanuja (Carnegie Mellon University)  
Wenhao Zhu (ByteDance)

## Reviewers

Niyati Bafna (Johns Hopkins University)  
Houda Bouamor (Carnegie Mellon University, Qatar)  
Laurie Burchell (Common Crawl Foundation)  
Yogen Vilas Chaudhari (PowerSchool)  
Pinzhen Chen (Queen's University Belfast)  
Koel Dutta Chowdhury (Saarland University)  
Peter Devine (University of Edinburgh)  
Ona de Gibert (University of Helsinki)  
Shehenaz Hossain (Munster Technological University)  
Songbo Hu (University of Cambridge)  
Xu Huang (Nanjing University)  
Aishwarya Jadhav (University of California, Berkeley)  
Ahrii Kim (Soongsil University)  
Mateusz Klimaszewski (Warsaw University of Technology)  
Junxiao Liu (Nanjing University)  
Bhavivya Malik (University of Edinburgh)  
Nikita Moghe (Amazon)  
Dayyán O'Brien (University of Edinburgh)  
Stefano Perrella (Sapienza University of Rome)  
Esther Ploeger (Aalborg University)  
Lorenzo Proietti (Sapienza University of Rome)  
Shenbin Qian (University of Oslo)  
A B M Ashikur Rahman (King Fahad University of Petroleum and Minerals)  
Jacqueline Rowe (University of Edinburgh)  
Sherrie Shen (University of Edinburgh)  
Ashok Urlana (International Institute of Information Technology, Hyderabad)  
Zhijun Wang (Nanjing University)  
Nikita Kiran Yeole (Virginia Tech)  
Zheng Zhao (University of Edinburgh)  
Vilém Zouhar (ETH Zurich)

# Invited Talks

## Invited Talks

Sebastian Ruder (Meta)

Freda Shi (University of Waterloo)

Wenda Xu (Google DeepMind)

## Table of Contents

<i>LLMs as Span Annotators: A Comparative Study of LLMs and Humans</i> Zdeněk Kasner, Vilém Zouhar, Patrícia Schmidtová, Ivan Kartáč, Kristýna Onderková, Ondrej Platek, Dimitra Gkatzia, Saad Mahamood, Ondrej Dusek and Simone Balloccu . . . . .	1
<i>On the Credibility of Evaluating LLMs using Survey Questions</i> Jindřich Libovický . . . . .	23
<i>An improved Code-Switching Detection System for some Indic Languages</i> Karan Bhanushali and Fritz Hohl . . . . .	35
<i>Vinclat: Evaluating Reasoning, Cognition and Culture in One Game</i> Marc Pàmies, Javier Aula-Blasco, Aitor Gonzalez-Agirre and Marta Villegas . . . . .	49
<i>Conceptual Cultural Index: A Metric for Cultural Specificity via Relative Generality</i> Takumi Ohashi and Hitoshi Iyatomi . . . . .	67
<i>The Anthropology of Food: How NLP can Help us Unravel the Food cultures of the World</i> Arij Riabi, Sougata Saha and Monojit Choudhury . . . . .	76
<i>LLM-as-a-qualitative-judge: automating error analysis in natural language generation</i> Nadezhda Chirkova, Tunde Oluwaseyi Ajayi, Seth Aycocock, Zain Muhammad Mujahid, Vladana Perlić, Ekaterina Borisova and Markarit Vartampetian . . . . .	99
<i>Cross-Lingual Stability of LLM Judges Under Controlled Generation: Evidence from Finno-Ugric Languages</i> Isaac Chung and Linda Freienthal . . . . .	133
<i>Cross-lingual and cross-country approaches to argument component detection: a comparative study</i> Cecilia Graiff, Chloé Clavel and Benoît Sagot . . . . .	149
<i>UNSC-Bench: Evaluating LLM Diplomatic Role-Playing Through UN Security Council Vote Prediction</i> Ayush Nangia, Aman Gokrani and Ruggero Marino Lazzaroni . . . . .	162
<i>Leveraging Wikidata for Geographically Informed Sociocultural Bias Dataset Creation: Application to Latin America</i> Yannis Karmim, Renato Pino, Hernan Contreras, Hernan Lira, Sebastian Cifuentes, Simon Escoffier, Luis Martí, Djamé Seddah and Valentin Barriere . . . . .	177
<i>Whom to Trust? Analyzing the Divergence Between User Satisfaction and LLM-as-a-Judge in E-Commerce RAG Systems</i> Arif Türkmen and Kaan Efe Keleş . . . . .	189
<i>Query-Following vs Context-Anchoring: How LLMs Handle Cross-Turn Language Switching</i> Kyuhee Kim, Chengheng Li Chen and Anna Sotnikova . . . . .	196
<i>Generating Difficult-to-Translate Texts</i> Vilém Zouhar, Wenda Xu, Parker Riley, Juraj Juraska, Mara Finkelstein, Markus Freitag and Daniel Deutsch . . . . .	204
<i>'A Woman is More Culturally Knowledgeable than A Man?': The Effect of Personas on Cultural Norm Interpretation in LLMs</i> Mahammed Kamruzzaman, Hieu Minh Nguyen, Nazmul Hassan and Gene Louis Kim . . . . .	220

# LLMs as Span Annotators: A Comparative Study of LLMs and Humans

Zdeněk Kasner<sup>1</sup> Vilém Zouhar<sup>2</sup> Patrícia Schmidtová<sup>1</sup>  
Ivan Kartáč<sup>1</sup> Kristýna Onderková<sup>1</sup> Ondřej Plátek<sup>1</sup>  
Dimitra Gkatzia<sup>3</sup> Saad Mahamood<sup>4</sup> Ondřej Dušek<sup>1</sup> Simone Balloccu<sup>5</sup>

<sup>1</sup>Charles University <sup>2</sup>ETH Zurich <sup>3</sup>Edinburgh Napier University

<sup>4</sup>trivago N.V. <sup>5</sup>TU Darmstadt, Germany

Contact: [kasner@ufal.mff.cuni.cz](mailto:kasner@ufal.mff.cuni.cz)

## Abstract

Span annotation – annotating specific text features at the span level – can be used to evaluate texts where single-score metrics fail to provide actionable feedback. Until recently, span annotation was done by human annotators or fine-tuned models. In this paper, we study whether large language models (LLMs) can serve as an alternative to human annotators. We compare the abilities of LLMs to skilled human annotators on three span annotation tasks: evaluating data-to-text generation, identifying translation errors, and detecting propaganda techniques. We show that overall, LLMs have only moderate inter-annotator agreement (IAA) with human annotators. However, we demonstrate that LLMs make errors at a similar rate as skilled crowdworkers. LLMs also produce annotations at a fraction of the cost per output annotation. We release the dataset of over 40k model and human span annotations for further research.<sup>1</sup>

## 1 Introduction

Fine-grained aspects of texts, such as semantic accuracy or coherence, depend on local lexical choices. To reflect these aspects in quality judgments of texts, techniques are needed that provide the appropriate amount of detail. However, most automatic evaluation metrics for Natural Language Generation (NLG) assign only singular scores for the whole text per each evaluated aspect (Gkatzia and Mahamood, 2015; Sai et al., 2023; Schmidtová et al., 2024). Although numerical values make it easy to rank systems, these metrics are too crude and susceptible to biases or miscalibration of the underlying models (Gehrmann et al., 2023; Liu et al., 2024; Wang et al., 2024; Gao et al., 2024).

The subject of our study, *span annotation* (Figure 1), offers an alternative approach. Instead of assigning a single score for each evaluated aspect, the

goal of span annotation is to localize text spans of interest and classify them according to task-specific guidelines. Span annotations are aligned to specific parts of the evaluated text, which makes them more explainable and actionable than numerical ratings.

Despite its advantages, span annotation has not yet been widely applied in automatic NLG evaluation. The method traditionally required human annotators, making it costly and difficult to scale (Da San Martino et al., 2019; Thomson and Reiter, 2020; Popovic, 2020; Kocmi et al., 2024c). The *LLM-as-a-judge* paradigm recently emerged as a promising solution to this problem (Zheng et al., 2023; Gu et al., 2024), allowing task-specific applications (Kocmi and Federmann, 2023; Hasanain et al., 2024). However, to our knowledge, no study has systematically compared span annotation performance between LLMs and human annotators.

The central focus of our investigation is comparing human annotators and state-of-the-art LLMs on span annotation tasks. We select three span annotation tasks (cf. Section 3.1): evaluating data-to-text generation (Thomson and Reiter, 2020), identifying errors in machine translation (Kocmi et al., 2024a), and detecting propaganda techniques in human-written texts (Da San Martino et al., 2019).

Our contributions are as follows:

1. We establish that with structured outputs and detailed annotation guidelines, LLMs can serve as robust span annotators, yielding relevant spans for all three annotation tasks we work with (Sections 3.2 and 5.1).
2. We show that LLMs have moderate inter-annotator agreement with human annotators overall, but can reach the agreement level among verified crowdworkers who passed a qualification task (Section 5.2).
3. We discover the sources of model errors by

<sup>1</sup>Project website: <https://llm-span-annotators.github.io>

Task	Text $Y$ with annotations $A$ (category, span, reason)	Categories $C$	Guidel. $\mathcal{G}$	Input $X$
D2T-Eval	Skies will be <b>mostly clear</b> , but <b>winds will remain strong</b> . <i>Rain on Mon &amp; Wed</i> <i>Wind speed data is missing.</i>	CONTRADICTION <b>C</b> NOT CHECKABLE <b>NC</b> (...)	Annotate semantic errors (...)	Mon Tue Wed 
MT-Eval	The quick brown fox <b>jump</b> over the lazy <b>fox</b> . <i>Third person singular</i> <i>'Hund' translates to 'dog'</i>	MAJOR <b>MJ</b> MINOR <b>MN</b>	Annotate translation errors (...)	Der schnelle braune Fuchs springt über den faulen Hund.
Propaganda	<b>Study Finds</b> That Driving Car Is More Efficient than Biking <i>Appeal to a 'study'</i>	APPEAL TO AUTHORITY <b>AA</b> (...)	Annotate propaganda techniques (...)	∅

Figure 1: Examples of span annotation tasks that we automate with LLMs. We unify the setup for evaluation tasks (D2T-EVAL, MT-EVAL) and text analysis tasks (PROPAGANDA).

manually analyzing a subset of LLM annotation outputs (Section 5.3).

4. We release a dataset of more than 40k human and model annotations, including annotations collected from crowdworkers and reasoning traces from reasoning LLMs.

## 2 Related Work

**LLMs for NLG Evaluation.** Automatic NLG metrics traditionally assess text quality by measuring similarity to human-written reference texts (Sai et al., 2023; Schmidová et al., 2024). As such, they are unable to quantify more fine-grained aspects (Gehrmann et al., 2023; Freitag et al., 2021) and do not correlate well with human judgments (Novikova et al., 2017; Reiter, 2018). With the emerging LLM-as-a-judge paradigm (Gu et al., 2024), LLMs have been applied as evaluators on various tasks, using simple numeric scoring (Bavaresco et al., 2025; Liu et al., 2023; Sottana et al., 2023; Leiter et al., 2023; Chiang and Lee, 2023), or free-form feedback (Li et al., 2024; Kim et al., 2024a,b; Kartáč et al., 2025). However, as these outputs are not firmly grounded in text, they tend to miss fine-grained aspects and are influenced by LLM biases (Stureborg et al., 2024; Koo et al., 2024; Wang et al., 2024).

**Span Annotation Protocol.** In machine translation (MT), span annotation is a long-standing component of protocols such as MQM or ESA (Lommel et al., 2014; Mariana, 2014; Popovic, 2020; Kocmi et al., 2024c), where human annotators mark erroneous spans in translations. In data-to-text (D2T) generation, span annotation was applied by Thomson and Reiter (2020), who introduced a span-based evaluation protocol for annotation of generated basketball match reports. Span

annotation is also used to judge intrinsic text qualities, such as coherence or use of rhetorical devices, in tasks such as propaganda detection (Da San Martino et al., 2019) and text summarization (Subbiah et al., 2024). Unlike our work, these works focus on span annotation with human annotators.

**Automatic Span Annotation.** Early attempts at automating span annotation with ad-hoc guidelines were based on fine-tuned pre-trained encoder models. That includes evaluation of MT (Guerreiro et al., 2024), D2T generation (Kasner et al., 2021) or text summarization (Goyal et al., 2022), as well as propaganda detection (Martino et al., 2020; Goffredo et al., 2023; Piskorski et al., 2023). Automating span annotation with LLMs is more flexible and benefits from increasing LLM capabilities. We build on work that applies LLMs as a task-specific evaluation tool (Kocmi and Federmann, 2023; Fernandes et al., 2023; Hasanain et al., 2024; Kasner and Dušek, 2024; Chang et al., 2024; Zouhar et al., 2025; Kartáč et al., 2025; Ramponi et al., 2025). Furthermore, Semin et al. (2026) recently investigated various strategies for automatic span annotation with LLMs. Our work is the first that systematically compares the performance of LLMs to human annotators.

## 3 Automating Span Annotation with LLMs

We first formally introduce the span annotation process in Section 3.1. Next, we discuss how to automate the process with LLMs in Section 3.2 and how to evaluate the quality of span annotations in Section 3.3.

### 3.1 Span Annotation: Task Definition

The aim of span annotation is to annotate a **text sequence**  $Y = \langle y_1, \dots, y_n \rangle$  given:

- the set of **categories**  $C = \{c_1, \dots, c_k\}$ ,
- the annotation **guidelines**  $\mathcal{G}$ ,
- the **source**  $X$  (such as the translation source; empty if we are annotating only intrinsic text aspects).

The output is a set of annotations  $A = \{a_1, \dots, a_m\}$ , where each annotation  $a_i$  is a tuple  $\langle s_i, e_i, c_i, r_i \rangle$ :

- $s_i, e_i \in \{1, \dots, n\}, s_i < e_i$  are the start and end indices of the annotated span,
- $c_i \in C$  is the assigned annotation category,
- $r_i$  is a reason for the annotation (optional).

### 3.2 Span Annotation with LLMs

In our setup, annotations  $A$  for the given input  $\langle Y, C, \mathcal{G}, X \rangle$  are collected from an LLM:

$$A = \text{LLM}(\text{prompt}(Y, C, \mathcal{G}, X)).$$

To obtain the annotations, we follow the setup of [Kasner and Dušek \(2024\)](#): we request the list of annotations in JSON format, using constrained decoding with a fixed JSON scheme to ensure that the output is syntactically valid. We require each annotation to contain the fields `reason` (the explanation  $r_i$ ), `text` (the textual content of the span), and `type` (the integer index of the error category  $c_i$ ).<sup>2</sup>

For reasoning models not supporting structured output, we retrieve the raw answer from the model, strip any parts within the `<think></think>` tags (if present), and consider the latest valid top-level JSON object as the model’s response.

### 3.3 Evaluating Span Annotations

To compare annotations automatically, we need a notion of similarity between two sets of annotations  $\mathcal{A} = \{A_1, A_2, \dots, A_{|Y|}\}$  and  $\hat{\mathcal{A}} = \{\hat{A}_1, \hat{A}_2, \dots, \hat{A}_{|Y|}\}$  over a set of texts  $\mathcal{Y} = \{Y_1, Y_2, \dots, Y_{|Y|}\}$ . Note that basic inter-annotator agreement metrics such as Cohen’s  $\kappa$  ([Cohen, 1960](#)) are not applicable in our case, as they require

<sup>2</sup>Following [Castillo \(2024\)](#), we ensure that the reason field is generated first.

a fixed set of annotation units, while the number and position of spans in the span annotation task may differ ([Mathet et al., 2015](#)). Therefore, we consider the following similarity metrics:

**Pearson correlation  $\rho$  over counts.** This metric compares how many spans were annotated for each example:

$$\text{Pearson}(\mathcal{A}, \hat{\mathcal{A}}) = \rho(|A_Y|, |\hat{A}_Y|_{Y \in \mathcal{Y}}) \quad (1)$$

The correlation serves as a sanity check: a low value would suggest that an annotator either skips examples or over-annotates, indicating unclear annotation guidelines.

**Precision, Recall, and  $F_1$ .** To quantify the degree of alignment between individual annotations, we compute precision, recall, and  $F_1$  as defined by [Da San Martino et al. \(2019\)](#). These measures are on matching annotations, adjusted to give partial credit to imperfect matches:

$$\text{Precision}(A_Y, \hat{A}_Y) = \frac{1}{|A_Y|} \sum_{a \in A_Y} \frac{|a \cap \hat{a}|}{|a|}, \quad (2)$$

$$\text{Recall}(A_Y, \hat{A}_Y) = \frac{1}{|\hat{A}_Y|} \sum_{\hat{a} \in \hat{A}_Y} \frac{|a \cap \hat{a}|}{|\hat{a}|}, \quad (3)$$

where  $a \cap \hat{a}$  is the character overlap between two annotation spans and  $|a| = e - s + 1$  is the length of the annotation span in characters (see Section 3.1). Subsequently, we compute the  $F_1$ -score as the harmonic mean of precision and recall.

For each of the metrics, we consider *soft* and *hard* variants. The *hard* variant only considers overlaps where the span category matches, while the *soft* variant disregards the categories. We consider the hard variant to be the default. In addition, we report the difference  $F_1 \Delta = F_1(\text{soft}) - F_1(\text{hard})$ .

**Gamma  $\gamma$ .** The  $F_1$  score is sensitive to varying span granularities and does not consider near-matches with no overlap or agreement by chance. To this end, we follow [Da San Martino et al. \(2019\)](#) and [Hasanain et al. \(2024\)](#) in using the  $\gamma$  score ([Mathet et al., 2015](#)) as a complementary metric. The metric builds the best possible alignment between the sets of annotations  $A_Y$  and  $\hat{A}_Y$  and computes the “disorder” of this alignment based on the *positional* and *categorical* dissimilarities of aligned annotations. The score ranges from  $-\infty$  to 1, where 1 is achieved when the annotations

Task	# Cat.	# Texts	Avg. Len	Novel Data
D2T-EVAL	6	1,296	118/715	✓
MT-EVAL	2	2,854	26/185	✗
PROPAGANDA	18	100	914/4,659	✗

Table 1: Overview of span annotation tasks used in our experiments. # *Cat.* denotes the number of categories used in the task (see Appendix C for their listings), # *Texts* the number of texts annotated, *Avg. Len* the average number of words/characters in the output, and *Novel Data* indicates newly collected data.

are perfectly aligned. The  $\gamma$  score extends Krippendorff’s  $\alpha$  (Krippendorff, 1980), another popular metric, by computing the category-aware span alignments. We use the implementation of Titeux and Riad (2021).

**$S_\emptyset$  score.** For an output  $y$ , one or both annotation sets  $A_Y, \hat{A}_Y$  may be empty. This is in fact desirable, e.g., if the goal is to annotate errors in an output that is entirely correct. However, these cases are not properly reflected by the other scores we are using: the F1 score only focuses on counting error spans and is not affected by true negatives, and the  $\gamma$  score is undefined if any of the two annotation sets is empty (these examples therefore need to be skipped during the  $\gamma$  computation). To compensate for this, we introduce a score  $S_\emptyset$  that is computed for examples where any of  $A_Y, \hat{A}_Y$  is empty:

$$S_\emptyset = 1/(1 + |A|), \quad (4)$$

where:

$$|A| = \begin{cases} |A_Y| & \text{if } |\hat{A}_Y| = 0, \\ |\hat{A}_Y| & \text{otherwise.} \end{cases} \quad (5)$$

The score is equal to 1 for the cases where no annotator produced any annotation (i.e., a perfect match) and decreases proportionally to the number of annotations from the annotator that provided a non-zero number of annotations.

## 4 Experiments

### 4.1 Tasks

We cover three span annotation tasks of different qualitative aspects. We focus on tasks that do not have extensive training data resources and cannot be readily solved by encoder models (such as, e.g., named entity tagging): evaluating data-to-text generation (D2T-EVAL; Section 4.1.1), identifying errors in machine translation (MT-EVAL; Section 4.1.2), and detecting propaganda techniques

(PROPAGANDA; Section 4.1.3). See Table 1 for an overview of our datasets.

#### 4.1.1 D2T-EVAL: Evaluation of Data-to-text Generation

In D2T-EVAL, we use span annotation to evaluate semantic accuracy and stylistic aspects of data-to-text generation outputs (Sharma et al., 2022; Celikyilmaz et al., 2020). The inputs  $X$  are the structured data used to generate the output text  $Y$ .

We use D2T-EVAL as a control task to mitigate the effects of *data contamination*: the fact that the performance of the model might be inflated by previous exposure to publicly available benchmarks (Balloccu et al., 2024; Dong et al., 2024; Jiang et al., 2024). Instead of using an existing dataset, we use the QUINTD tool (Kasner and Dušek, 2024) to download structured inputs from multiple public APIs.<sup>3</sup> To obtain output texts for the structured data, we prompt LLMs in a zero-shot setting, asking them to generate a summary of the given data using approximately five sentences. Note that we do not need to deal with the factuality of outputs here, as the sole purpose of the texts is being the input to the annotation process (in fact, having some number of errors is desirable). See Appendix B.1 for more details.

To gather annotations for our dataset, we use crowdworkers from Prolific.com. We apply best practices for gathering human annotations, including an iterative process to refine annotation guidelines and preselecting the best-performing annotators using a qualification task (Tseng et al., 2020; Iskender et al., 2020; Huang et al., 2023; Zhang et al., 2023). Our process of collecting annotations proceeded in two stages, following the setup of Zhang et al. (2023): (1) a *qualification task* for preselecting skilled annotators, and (2) the *main task* for collecting the annotations. See Appendix B.2 for more details on collecting annotations.

For quality checks, we collect additional internal gold annotation (by the authors) for subsets of the data:  $\mathcal{D}_{\text{dev}}$  for selecting the best prompt and  $\mathcal{D}_{\text{iaa}}$  for validating the performance of human annotators (cf. Appendix B.2). Here is a complete overview of our data splits for D2T-EVAL:

- $\mathcal{D}_{\text{test}}$  (1200 outputs) – for LLM evaluation, annotated with crowdworkers,

<sup>3</sup>We selected two of the existing domains: openweather for generating weather forecasts and gsmarena for generating phone descriptions. We also add the football domain (using RapidAPI - API-Football) for generating soccer game reports.

- $\mathcal{D}_{\text{dev}}$  (84 outputs) – for the study of prompt variants, annotated internally,
- $\mathcal{D}_{\text{iaa}}$  (12 outputs) – control for human crowdworkers, annotated internally.

#### 4.1.2 MT-EVAL: Identifying Errors in Machine Translation

For MT-EVAL, we use the dataset of system outputs from the WMT 2024 general shared task (Kocmi et al., 2024b). The system outputs were annotated with the Error Span Annotation (ESA) protocol (Kocmi et al., 2024c) by professional translators.

The inputs  $X$  for MT-EVAL are the texts in the source language used to produce the translation  $Y$ . We follow the WMT 2024 annotation guidelines, focusing on character-level span annotations of *Major* and *Minor* translation errors (see Table 6 for their definitions). Note that unlike for the other tasks, the annotations in MT-EVAL cannot overlap and need not be aligned with word boundaries.

We select the three textual domains present in the WMT 2024 shared task: news, literary, and social; using the data translated from English into other languages: Chinese, Czech, German, Hindi, Icelandic, Japanese, Russian, Spanish, and Ukrainian.

The original dataset has nearly 50k model outputs, making it too extensive for our evaluation campaign. Therefore, we used a balanced subsample: For each of the nine language pairs, we randomly sample ten input translation segments. We then take all available system outputs for these 90 input segments, making up 2,854 examples in total.

#### 4.1.3 PROPAGANDA: Propaganda Technique Detection

For the PROPAGANDA task, we use the dataset of Da San Martino et al. (2019) containing news collected mostly from on-line propagandistic sources. The token-level annotations in the dataset created by expert annotators cover 18 categories of logical fallacies and persuasion techniques. We use the test split for our experiments. Inputs  $X$  are empty for this task, as all annotated categories are intrinsic to the evaluated text  $Y$ .

## 4.2 Collecting LLM annotations

**Models** For collecting span annotations with LLMs, we use a mixture of open and proprietary state-of-the-art models:

Prompt	Llama 3.3			DeepSeek-R1		
	$F_1$	$\gamma$	#a/o	$F_1$	$\gamma$	#a/o
$\mathcal{P}_{\text{base}}$	0.20	0.13	2.4	<b>0.25</b>	<b>0.20</b>	1.0
$\mathcal{P}_{\text{cot}}$	0.09	0.10	0.8	0.24	0.19	1.1
$\mathcal{P}_{5\text{shot}}$	<b>0.25</b>	<b>0.18</b>	2.5	0.21	0.16	1.4
$\mathcal{P}_{\text{noguide}}$	0.11	0.08	3.4	0.20	0.16	1.6
$\mathcal{P}_{\text{noreason}}$	0.22	0.13	2.2	0.24	0.18	1.1

Table 2: Comparison of prompting techniques on the  $\mathcal{D}_{\text{dev}}$  (#a/o is the average number of annotations per output).

- **instruction-tuned models:** Llama 3.3 70B (Grattafiori et al., 2024), GPT-4o (Hurst et al., 2024), and Claude 3.7 Sonnet (Anthropic, 2025),
- **reasoning models:**<sup>4</sup> DeepSeek-R1 70B (DeepSeek-AI, 2025), o3-mini (OpenAI, 2025), and Gemini 2.0 Flash Thinking (Deepmind, 2025).

See Appendix A for details on our experimental setup.

**Prompts** We define several prompt variants for our experiments.  $\mathcal{P}_{\text{base}}$  is the base prompt that includes the guidelines  $\mathcal{G}$  as given to human annotators and asks the model to explain its annotation. By extending  $\mathcal{P}_{\text{base}}$ , we implement a few-shot prompt adding 5 examples ( $\mathcal{P}_{5\text{shot}}$ ) and a chain-of-thought prompt simply asking the model to produce intermediate reasoning ( $\mathcal{P}_{\text{cot}}$ ). We also ablate  $\mathcal{P}_{\text{base}}$  by removing extended guidelines ( $\mathcal{P}_{\text{noguide}}$ ) and not asking for explanations ( $\mathcal{P}_{\text{noreason}}$ ). The full prompts can be found in Appendix D.

## 5 Results

We first investigate the effect of prompting techniques in Section 5.1. Next, we evaluate the LLM annotations using automatic metrics (Section 5.2) and manually analyze the errors in the model outputs (Section 5.3).

### 5.1 Prompting Techniques

We perform preliminary experiments on the D2T-EVAL task  $\mathcal{D}_{\text{dev}}$  set using open models (Llama 3.3 and DeepSeek-R1) to study the differences between prompting techniques. The results are shown in Table 2.

<sup>4</sup>By *reasoning* models we understand the models that use extra inference time to generate a thinking trace before providing the answer (Marjanović et al., 2025).

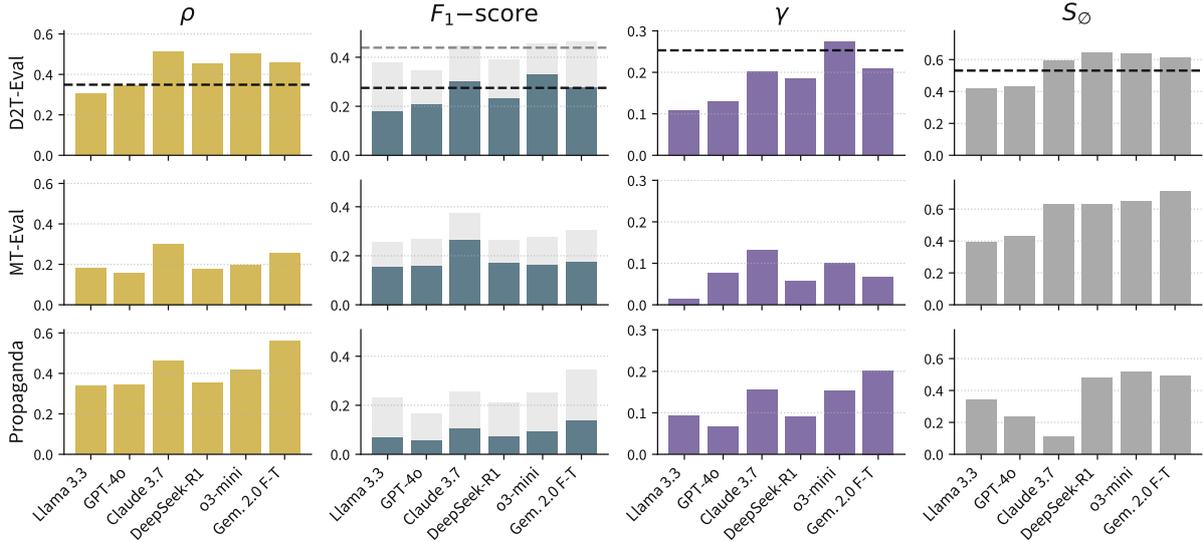


Figure 2: Comparison between LLMs using  $\mathcal{P}_{\text{base}}$  and human annotators. Rows represent different tasks (see Section 4.1), columns show different annotation similarity metrics (see Section 3.3). For the  $F_1$  score, the shadow bar denotes its *soft* variant. The dashed horizontal lines denote agreement between our human annotators for D2T-EVAL (the agreement is not available for the external datasets). More detailed results are included in Tables 11 to 14.

Including detailed guidelines seems beneficial: omitting the guidelines ( $\mathcal{P}_{\text{noguide}}$ ) lowers the performance of both models. In contrast, not letting the model explain the annotation ( $\mathcal{P}_{\text{noreason}}$ ) does not have a substantial effect. For Llama 3.3, the chain-of-thought (CoT) prompting ( $\mathcal{P}_{\text{cot}}$ ) makes it produce fewer annotations per example than the base variant (0.8 vs. 2.4), leading to lower F1 and  $\gamma$  scores. Llama 3.3 with  $\mathcal{P}_{\text{cot}}$  tends to “overthink” the annotations, deciding not to annotate cases of errors against which it can find some arguments.

Few-shot prompting ( $\mathcal{P}_{5\text{shot}}$ ) brings ambivalent results, increasing Llama 3.3 scores but doing the opposite for DeepSeek-R1. This observation is aligned with DeepSeek-AI (2025), who note that few-shot prompting degrades the performance of DeepSeek-R1. Given these considerations, we decided to use  $\mathcal{P}_{\text{base}}$  for further experiments.

## 5.2 LLM vs. Human Annotations

Next, we compare LLM and human annotations using the metrics described in Section 3.3. The overall results for all tasks are given in Figure 2. We provide detailed results for individual tasks in the Appendix F.

**Reasoning models outperform instruction-tuned models.** DeepSeek-R1 generally outperforms the

Human annotators	Model predictions					
	Contradictory	Not checkable	Misleading	Incoherent	Repetitive	Other
Contradictory	577	29	115	18	1	2
Not checkable	52	32	29	9	1	2
Misleading	91	11	51	6	0	1
Incoherent	39	3	12	9	1	0
Repetitive	7	1	6	1	4	0
Other	4	1	5	0	0	0

Figure 3: Confusion matrix for D2T-EVAL (*Contradictory*, *Not checkable*, *Misleading*, *Incoherent*, *Repetitive*, *Other*), averaged across models (see Table 5 for category descriptions).

same-sized Llama 3.3.<sup>5</sup> Its superiority is most pronounced on D2T-EVAL ( $F_1$ -score of 0.23 vs. 0.18,  $\gamma$  score of 0.19 vs. 0.11). The same observation applies to OpenAI models, where o3-mini outperforms GPT-4o. A notable exception to this trend is the non-reasoning Claude 3.7 Sonnet, which scores mostly on par with o3-mini and excels at MT-EVAL.

<sup>5</sup>The models are comparable as the 70B distilled variant of DeepSeek-R1 is based on Llama 3.3 70B. See Section A.1 for details.

Model	Cost (\$/1k out)	Time (s/out)
crowdworkers	500	129.1
Llama 3.3	-	21.6
DeepSeek-R1	-	227.5
Claude 3.7 Sonnet	10.5	9.0
o3-mini	3.6	21.8

Table 3: Estimate of costs and time requirements on D2T-EVAL: crowdworkers on Prolific, open models (Llama 3.3, DeepSeek-R1), and proprietary models (Claude 3.7 Sonnet, o3-mini).

**LLMs reach human IAA on D2T-EVAL, PROPAGANDA is harder.** For D2T-EVAL, we compare model results with an average IAA on a subset of examples annotated by two human annotators. Here, o3-mini, Claude 3.7 and Gemini 2.0 mostly reach or surpass human agreement. For PROPAGANDA, the upper bound of IAA is the result of [Da San Martino et al. \(2019\)](#), who report  $\gamma = 0.31$  for annotators before consolidation. This score is substantially higher than LLMs (the best LLM score being  $\gamma = 0.16$  for Claude 3.7 Sonnet). However, this task used expert annotators and has the largest number of categories. The latter property is reflected in the large difference between the soft and hard  $F_1$  scores.<sup>6</sup>

**Models confuse related categories.** Confusion matrices (see Figure 3 and Appendix F) suggest that the models tend mainly to confuse related categories, which may be related to ambiguity or subjective understanding of category definitions. The models also use a less diverse distribution of categories than human annotators.

**LLMs are more cost- and time-efficient than human annotators.** An important factor when comparing LLMs and human annotators is efficiency with respect to cost and time per output. For D2T-EVAL, crowdsourced annotation for 1k outputs costs approximately \$500, while annotating the same amount of outputs with the high-performance o3-mini LLM costs \$3.60 (see Table 3). In terms of time, the crowdworkers take 129.1 seconds per output on average, which is better than DeepSeek-R1 70B running on our local infrastructure, but an order of magnitude slower than the API-based

<sup>6</sup>We omit the comparison with human IAA in MT-EVAL. While the WMT24 dataset for MT-EVAL contains examples annotated with a pair of annotators, these examples take up only a small fraction and exhibit high variance between language pairs.

models.<sup>7</sup> Therefore, LLMs are a more efficient alternative in terms of costs and time.

### 5.3 Manual Analysis of LLM Annotations

To gain more insights into the qualitative aspects of LLM annotations, we manually analyzed the quality of LLM annotations on 216 samples from D2T-EVAL and PROPAGANDA.<sup>8</sup> For each model, we sampled three annotations per category in D2T-EVAL and one annotation per category in PROPAGANDA. Without access to the annotation source, we classified the annotations and their explanations as *Correct*, *Partially correct*, *Wrong category*, *Incorrect*, and *Undecidable*.

We show the results in Figure 4 and Tables 19 and 20. In total, we marked 49.5% of LLM-generated annotations and 50.5% of reasons as correct (with 9.2% of annotations and 12.5% of reasons additionally marked as partially correct). Reasoning models perform better, with 56.4% of their annotations and 58.3% reasons marked as completely correct. The most accurate annotations on D2T-EVAL were those made by Gemini 2.0 and DeepSeek-R1. o3-mini performed well on both tasks, although PROPAGANDA proved challenging for all models.

**What are the sources of model errors?** We find that the models often select wrong error categories despite identifying real issues (e.g., labeling *Contradictory* statements as *Incoherent*). Models also tend to be overly attentive, flagging noise in the data (e.g., markup or off-topic content in PROPAGANDA) as errors, or marking slight numerical variations (such as rounded values) as misleading. All of these cases could be tackled by more descriptive guidelines or additional examples. However, in some cases, the models also misread or misinterpreted the data (e.g., claiming wind speed measurements do not exist when they do), which hints at deeper issues with understanding the data. Incorrect explanations vary from incomplete explanations (addressing only part of a multi-issue span), irrelevant explanations (e.g., appealing to facts that are “missing” from the text) to incorrectly flagging subjective statements (e.g., “*long-lasting usage*”)

<sup>7</sup>Note that we do not ask the crowdworkers to give us a reason  $r$  for the annotation, which would arguably make the responses of the crowdworkers slower.

<sup>8</sup>The analysis was split among 7 authors of this paper. While we did not do double annotation due to lack of time, we discussed any unclear cases throughout the process. We do not include MT-EVAL in the manual analysis due to our insufficient expertise in most target languages.



Figure 4: Results of our manual analysis. We analyzed 18 annotations and their explanations for each model and task (216 annotations in total). The color bars show annotations that we classified as *Correct*, *Partially correct*, *Wrong category*, *Incorrect*, and *Undecidable*. Detailed results are provided in Tables 19 and 20.

as factual errors. Occasionally, the model admits that it marked a correct span as an error, such as in “*The description of the game’s duration aligns with the data, providing coherent information*”.<sup>9</sup>

**How good are human annotations?** Concerningly, the LLM annotations that were marked as correct have only 24% hard character-level overlap (51% soft) with human annotations. This fact led us to analyze the quality of human annotations in D2T-EVAL (the task in which we had the necessary domain expertise). Using the same methodology as we used for the LLM annotations, we annotated a limited sample of 108 human annotations. We marked 45.3% of the annotations as *Correct*, which is comparable to the LLM annotations (see Table 4 for the results). These findings suggest that the task is hard even for human annotators, and the quality of annotations from crowdworkers varies, even if they are preselected using a qualification task.

## 6 Discussion

Here, we summarize our findings and discuss the implications of our results.

**Can LLMs substitute human annotators?** The IAA between LLMs and human annotators is only moderate, suggesting LLMs cannot straightforwardly replace human annotators. However, using LLMs may be a reasonable option in scenarios based on crowdworkers, where the strongest LLMs reach the average IAA between human annotators themselves. In other cases, when deciding whether to employ LLMs as span annotators, one needs to balance desired output quality with other practical aspects. Here, LLMs provide better flexibility, shorter response times, and lower costs. One

<sup>9</sup>This typically happened to GPT-4o, even though OpenAI API ensures JSON key ordering so the explanation *should* have been generated before the annotation (cf. Section 3.2).

should also consider the quality of available human annotators, as even qualified crowdworkers (i.e., those who passed a qualification task) make similar amounts of errors as LLMs. It can be also assumed that LLM-based span annotation will benefit from future increases in LLM capabilities, while crowdworkers may increasingly rely on LLM to complete tasks (Veselovsky et al., 2023). A promising solution seems to be a hybrid approach in which LLMs pre-annotate the text and humans post-edit the annotations (Zouhar et al., 2025).

**How to deploy LLMs as span annotators?** We recommend providing LLMs with detailed guidelines that describe conventions and how to handle ambiguous cases (cf. Figure 5). In contrast, we do not recommend providing specific examples (cf. Figure 8), as this approach did not bring consistent improvements. Arguably, this is due to the length and complexity of the examples, making them distracting to the model. When using LLMs with custom categories or guidelines, we recommend validating the model’s annotations against examples hand-annotated by experts on a sample of the data. In general, reasoning models tend to provide more reliable annotations at the cost of higher response times and token count.

**Is the task meaningful despite the low scores?** As pointed out by an anonymous reviewer of this paper, the low annotation accuracy – as found by our manual analysis – may indicate a fundamental limitation of the proposed evaluation setup. Span annotation is indeed complex and leaves more room for subjectivity than more straightforward annotation such as simple labels or scores. However, we argue that the detailed actionable feedback gained through span annotation outweighs the increased noise, both in terms of explainability of the outputs and their potential for further processing.

Source	Annotations				
	C	P	W	I	U
Human annotators	49	4	17	31	7

Table 4: Manual evaluation results for human annotators on D2T-EVAL. Categories: C=Correct, P=Partially correct, W=Wrong category, I=Incorrect, U=Undecidable.

## 7 Conclusion

We showed that LLMs can serve as span annotators for three span annotation tasks: evaluating data-to-text generation, identifying errors in machine translation, and detecting propaganda in human-written texts. Our experiments show that LLMs achieve moderate agreement with skilled human annotators. The models perform best in D2T-EVAL, where they are comparable to verified crowdworkers who passed a qualification task. Reasoning models consistently outperform their instruction-tuned counterparts, delivering more accurate annotations and providing more valid explanations for their decisions. Automating span annotation with LLMs seems to be a promising alternative to fine-grained human evaluation sourced from crowdsourcing platforms, opening the way towards scalable and actionable automatic NLG evaluation methods.

## Limitations

Although we aimed to select a representative sample of models, prompts and tasks, our choice is constrained by our limited time frame and budget. Our estimates of the upper-bound IAA for each task are difficult to establish and depend on many factors, such as the chosen annotation categories, their ambiguity, the annotation guidelines, or the qualification level of human annotators. The estimates are also not readily available for existing datasets and require additional data collection. Due to our insufficient expertise in the target languages, we also do not provide language-specific manual error analysis of results.

As an evaluation method, span annotation is not well-suited for certain NLG evaluation tasks such as annotating omissions or rating the overall text style. In these cases, it is best to combine span annotation with other evaluation methods.

## Author Contributions

DG and SB first came up with the idea for the project, with SB further coordinating and oversee-

ing the research process. ZK led the experimental design and execution part, including conducting both preliminary and main experiments, organizing the crowdsourcing campaigns, and processing the collected data. Multiple authors (DG, IK, KO, SB, SM, VZ, ZK) participated in the collection of gold data for D2T-EVAL. Similarly, multiple authors (IK, KO, OD, OP, PS, SB, ZK) were involved in manual evaluation of the model outputs. DG provided financial resources for the Prolific campaigns. SM and SB provided expertise in preparing annotation guidelines and structuring the Prolific campaigns. Data processing and analysis were handled mainly by ZK, VZ, and PS, with VZ providing extra support with the WMT data. The paper was written by ZK, VZ, IK, PS, OD, and SB.

## Acknowledgments

This work was funded by the European Union (ERC, NG-NLG, 101039303). It was additionally supported by the National Recovery Plan funded project MPO 60273/24/21300/21000 CEDMO 2.0 NPO and the Charles University Research Centre program No. 24/SSH/009. It used resources of the LINDAT/CLARIAH-CZ Research Infrastructure (Czech Ministry of Education, Youth, and Sports project No. LM2023062). We thank David M. Howcroft for his early input and contributions to the research methodologies adopted in this study.

## Ethics Statement

The human evaluation study was approved by the internal ethics committee of our institution. Our human annotators were hired over Prolific and paid the platform-recommended wage of 9 GBP/hour (adjusted to slightly higher rates to account for real annotation times). Annotators were pre-selected on the basis of their primary language (English). All annotators were shown detailed instructions and explanation of the data types, data sources, and the purpose of the research. The domains were selected so that they do not contain sensitive or potentially offensive content. We do not collect demographic data about participants.

## References

Marah I Abidin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat S. Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Caio

- César Teodoro Mendes, Weizhu Chen, Vishrav Chaudhary, Parul Chopra, Allie Del Giorno, Gustavo de Rosa, Matthew Dixon, Ronen Eldan, Dan Iter, Amit Garg, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Jamie Huynh, Mojan Javaheripi, Xin Jin, Piero Kauffmann, Nikos Karampatziakis, Dongwoo Kim, Mahoud Khademi, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Chen Liang, Weishung Liu, Eric Lin, Zeqi Lin, Piyush Madan, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norrick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacrose, Shital Shah, Ning Shang, Hiteshi Sharma, Xia Song, Masahiro Tanaka, Xin Wang, Rachel Ward, Guanhua Wang, Philipp A. Witte, Michael Wyatt, Can Xu, Jiahang Xu, Sonali Yadav, Fan Yang, Ziyi Yang, Donghan Yu, Chengruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024. [Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone](#). *CoRR*, abs/2404.14219.
- Anthropic. 2025. [Claude 3.7 Sonnet](#).
- Simone Balloccu, Patrícia Schmidová, Mateusz Lango, and Ondrej Dusek. 2024. [Leak, Cheat, Repeat: Data Contamination and Evaluation Malpractices in Closed-source LLMs](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2024 - Volume 1: Long Papers*, pages 67–93, St. Julian’s, Malta.
- Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, André F. T. Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suggia, Aditya K. Surikuchi, Ece Takmaz, and Alberto Testoni. 2025. [LLMs instead of Human Judges? A Large Scale Empirical Study across 20 NLP Evaluation Tasks](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL 2025*, pages 238–255, Vienna, Austria.
- Dylan Castillo. 2024. [Structured Outputs: Don’t Put the Cart Before the Horse](#).
- Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. [Evaluation of Text Generation: A Survey](#). *CoRR*, abs/2006.14799.
- Yapei Chang, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024. [BoookScore: A systematic exploration of book-length summarization in the era of LLMs](#). In *The Twelfth International Conference on Learning Representations, ICLR 2024*, Vienna, Austria.
- David Cheng-Han Chiang and Hung-yi Lee. 2023. [Can Large Language Models Be an Alternative to Human Evaluations?](#) In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023*, pages 15607–15631, Toronto, Canada.
- Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.
- Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019. [Fine-grained analysis of propaganda in news article](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5636–5646, Hong Kong, China. Association for Computational Linguistics.
- Google Deepmind. 2025. [Gemini 2.0 Flash Thinking](#).
- DeepSeek-AI. 2025. [DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning](#). *CoRR*, abs/2501.12948.
- Yihong Dong, Xue Jiang, Huanyu Liu, Zhi Jin, Bin Gu, Mengfei Yang, and Ge Li. 2024. [Generalization or Memorization: Data Contamination and Trustworthy Evaluation for Large Language Models](#). In *Findings of the Association for Computational Linguistics, ACL 2024*, pages 12039–12050, Bangkok, Thailand.
- Patrick Fernandes, Daniel Deutsch, Mara Finkelstein, Parker Riley, André F. T. Martins, Graham Neubig, Ankush Garg, Jonathan H. Clark, Markus Freitag, and Orhan Firat. 2023. [The Devil Is in the Errors: Leveraging Large Language Models for Fine-grained Machine Translation Evaluation](#). In *Proceedings of the Eighth Conference on Machine Translation, WMT 2023*, pages 1066–1083, Singapore.
- Markus Freitag, George F. Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. [Experts, Errors, and Context: A Large-scale Study of Human Evaluation for Machine Translation](#). *Trans. Assoc. Comput. Linguistics*, 9:1460–1474.
- Mingqi Gao, Xinyu Hu, Jie Ruan, Xiao Pu, and Xiaojun Wan. 2024. [LLM-based NLG Evaluation: Current Status and Challenges](#). *CoRR*, abs/2402.01383.
- Sebastian Gehrmann, Elizabeth Clark, and Thibault Selam. 2023. [Repairing the Cracked Foundation: A Survey of Obstacles in Evaluation Practices for Generated Text](#). *J. Artif. Intell. Res.*, 77:103–166.
- Dimitra Gkatzia and Saad Mahamood. 2015. [A Snapshot of NLG Evaluation Practices 2005 - 2014](#). In *ENLG 2015 - Proceedings of the 15th European Workshop on Natural Language Generation, 10-11 September 2015, University of Brighton*, pages 57–60, Brighton, UK.
- Pierpaolo Goffredo, Mariana Espinoza, Serena Villata, and Elena Cabrio. 2023. [Argument-based Detection and Classification of Fallacies in Political Debates](#).

- In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023*, pages 11101–11112, Singapore.
- Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. [SNaC: Coherence Error Detection for Narrative Summarization](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 444–463, Abu Dhabi, United Arab Emirates.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Yuanzhuo Wang, and Jian Guo. 2024. [A Survey on LLM-as-a-judge](#). *CoRR*, abs/2411.15594.
- Nuno Miguel Guerreiro, Ricardo Rei, Daan van Stigt, Luísa Coheur, Pierre Colombo, and André F. T. Martins. 2024. [xcomet : Transparent Machine Translation Evaluation through Fine-grained Error Detection](#). *Trans. Assoc. Comput. Linguistics*, 12:979–995.
- Maram Hasanain, Fatema Ahmad, and Firoj Alam. 2024. [Large Language Models for Propaganda Span Annotation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14522–14532, Miami, Florida, USA.
- Olivia Huang, Eve Fleisig, and Dan Klein. 2023. [Incorporating Worker Perspectives into MTurk Annotation Practices for NLP](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023*, pages 1010–1028, Singapore.
- Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Madry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrew Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Gierlter, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll L. Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, and Dane Sherburn. 2024. [GPT-4o System Card](#). *CoRR*, abs/2410.21276.
- Neslihan Iskender, Tim Polzehl, and Sebastian Möller. 2020. [Best Practices for Crowd-based Evaluation of German Summarization: Comparing Crowd, Expert and Automatic Evaluation](#). In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems, Eval4NLP 2020*, pages 164–175, Online.
- Minhao Jiang, Ken Ziyu Liu, Ming Zhong, Rylan Schaeffer, Siru Ouyang, Jiawei Han, and Sanmi Koyejo. 2024. [Investigating Data Contamination for Pre-training Language Models](#). *CoRR*, abs/2401.06059.
- Ivan Kartáč, Mateusz Lango, and Ondrej Dusek. 2025. [OpeNLGauge: An Explainable Metric for NLG Evaluation with Open-weights LLMs](#). *CoRR*, abs/2503.11858.
- Zdeněk Kasner and Ondrej Dušek. 2024. [Beyond Traditional Benchmarks: Analyzing Behaviors of Open LLMs on Data-to-text Generation](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12045–12072, Bangkok, Thailand.
- Zdenek Kasner, Simon Mille, and Ondrej Dusek. 2021. [Text-in-Context: Token-level Error Detection for Table-to-text Generation](#). In *Proceedings of the 14th International Conference on Natural Language Generation, INLG 2021*, pages 259–265, Aberdeen, Scotland, UK.
- Zdenek Kasner, Ondrej Plátek, Patrícia Schmidtová, Simone Balloccu, and Ondrej Dusek. 2024. [factgenie: A Framework for Span-based Evaluation of Generated Texts](#). In *Proceedings of the 17th International Natural Language Generation Conference, INLG 2024 - System Demonstrations*, pages 13–15, Tokyo, Japan.
- Seungone Kim, Jamin Shin, Yejin Choi, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoon Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. 2024a. [Prometheus: Inducing Fine-grained Evaluation Capability in Language Models](#). In *The Twelfth International Conference on Learning Representations, ICLR 2024*, Vienna, Austria.
- Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024b. [Prometheus 2: An Open Source Language Model Specialized in Evaluating Other Language Models](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language*

- Processing, EMNLP 2024, Miami, FL*, pages 4334–4353, USA.
- Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, Mariya Shmatova, Steinhórf Steingrímsson, and Vilém Zouhar. 2024a. [Findings of the WMT24 General Machine Translation Shared Task: The LLM Era Is Here but MT Is Not Solved Yet](#). In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA.
- Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, Mariya Shmatova, Steinhórf Steingrímsson, and Vilém Zouhar. 2024b. [Findings of the WMT24 General Machine Translation Shared Task: The LLM Era Is Here but MT Is Not Solved Yet](#). In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA.
- Tom Kocmi and Christian Federmann. 2023. [GEMBA-MQM: Detecting Translation Quality Error Spans with GPT-4](#). In *Proceedings of the Eighth Conference on Machine Translation, WMT 2023*, pages 768–775, Singapore.
- Tom Kocmi, Vilém Zouhar, Eleftherios Avramidis, Roman Grundkiewicz, Marzena Karpinska, Maja Popovic, Mrinmaya Sachan, and Mariya Shmatova. 2024c. [Error Span Annotation: A Balanced Approach for Human Evaluation of Machine Translation](#). In *Proceedings of the Ninth Conference on Machine Translation, WMT 2024, Miami, FL*, pages 1440–1453, USA.
- Ryan Koo, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, and Dongyeop Kang. 2024. [Benchmarking Cognitive Biases in Large Language Models as Evaluators](#). In *Findings of the Association for Computational Linguistics, ACL 2024*, pages 517–545, Bangkok, Thailand.
- Klaus Krippendorff. 1980. *Content Analysis: An Introduction to Its Methodology*. SAGE Publications, Beverly Hills, CA.
- Christoph Leiter, Juri Opitz, Daniel Deutsch, Yang Gao, Rotem Dror, and Steffen Eger. 2023. [The Eval4NLP 2023 Shared Task on Prompting Large Language Models as Explainable Metrics](#). In *Proceedings of the 4th Workshop on Evaluation and Comparison of NLP Systems, Eval4NLP 2023*, pages 117–138, Bali, Indonesia.
- Junlong Li, Shichao Sun, Weizhe Yuan, Run-Ze Fan, Hai Zhao, and Pengfei Liu. 2024. [Generative Judge for Evaluating Alignment](#). In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria*.
- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. [G-Eval: NLG Evaluation using Gpt-4 with Better Human Alignment](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023*, pages 2511–2522, Singapore.
- Yuxuan Liu, Tianchi Yang, Shaohan Huang, Zihan Zhang, Haizhen Huang, Furu Wei, Weiwei Deng, Feng Sun, and Qi Zhang. 2024. [Calibrating LLM-based Evaluator](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024*, pages 2638–2656, Torino, Italy.
- Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. [Multidimensional quality metrics \(MQM\): A framework for declaring and describing translation quality metrics](#). *Tradumática*, 0(12):0455–463.
- Valerie R Mariana. 2014. *The Multidimensional Quality Metric (MQM) framework: A new framework for translation quality assessment*. Brigham Young University.
- Sara Vera Marjanović, Arkil Patel, Vaibhav Adlakha, Milad Aghajohari, Parishad BehnamGhader, Mehar Bhatia, Aditi Khandelwal, Austin Kraft, Benno Krojer, Xing Han Lü, et al. 2025. [DeepSeek-R1 Thoughtology: Let’s think about LLM Reasoning](#). *arXiv preprint arXiv:2504.07128*.
- G. Da San Martino, A. Barrón-Cedeño, H. Wachsmuth, R. Petrov, and P. Nakov. 2020. [SemEval-2020 Task 11: Detection of Propaganda Techniques in News Articles](#).
- Yann Mathet, Antoine Widlöcher, and Jean-Philippe Métévier. 2015. [The Unified and Holistic Method Gamma \( \$\gamma\$ \) for Inter-annotator Agreement Measure and Alignment](#). *Comput. Linguistics*, 41(3):437–479.
- Jekaterina Novikova, Ondrej Dusek, Amanda Cercas Curry, and Verena Rieser. 2017. [Why We Need New Evaluation Metrics for NLG](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017*, pages 2241–2252, Copenhagen, Denmark.
- OpenAI. 2025. [OpenAI o3-mini System Card](#).
- Jakub Piskorski, Nicolas Stefanovitch, Nikolaos Nikolaidis, Giovanni Da San Martino, and Preslav Nakov. 2023. [Multilingual Multifaceted Understanding of Online News in Terms of Genre, Framing, and Persuasion Techniques](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023*, pages 3001–3022, Toronto, Canada.

- Maja Popovic. 2020. [Informative Manual Evaluation of Machine Translation Output](#). In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020*, pages 5059–5069, Barcelona, Spain (Online).
- Alan Ramponi, Agnese Daffara, and Sara Tonelli. 2025. [Fine-grained Fallacy Detection with Human Label Variation](#). In *Proceedings of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 1: Long Papers, Albuquerque*, pages 762–784, New Mexico, USA.
- Ehud Reiter. 2018. [A Structured Review of the Validity of BLEU](#). *Comput. Linguistics*, 44(3).
- Ananya B. Sai, Akash Kumar Mohankumar, and Mitesh M. Khapra. 2023. [A Survey of Evaluation Metrics Used for NLG Systems](#). *ACM Comput. Surv.*, 55(2):26:1–26:39.
- Patrícia Schmidtová, Saad Mahamood, Simone Baloccu, Ondrej Dusek, Albert Gatt, Dimitra Gkatzia, David M. Howcroft, Ondrej Plátek, and Adarsa Sivaprasad. 2024. [Automatic Metrics in Natural Language Generation: A survey of Current Evaluation Practices](#). In *Proceedings of the 17th International Natural Language Generation Conference, INLG 2024*, pages 557–583, Tokyo, Japan.
- Danil Semin, Ondřej Dušek, and Zdeněk Kasner. 2026. [Strategies for span labeling with large language models](#).
- Mandar Sharma, Ajay Kumar Gogineni, and Naren Ramakrishnan. 2022. [Innovations in Neural Data-to-text Generation](#). *CoRR*, abs/2207.12571.
- Andrea Sottana, Bin Liang, Kai Zou, and Zheng Yuan. 2023. [Evaluation Metrics in the Era of GPT-4: Reliably Evaluating Large Language Models on Sequence to Sequence Tasks](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023*, pages 8776–8788, Singapore.
- Rickard Stureborg, Dimitris Alikaniotis, and Yoshi Suhara. 2024. [Large Language Models are Inconsistent and Biased Evaluators](#). *CoRR*, abs/2405.01724.
- Melanie Subbiah, Sean Zhang, Lydia B. Chilton, and Kathleen R. McKeown. 2024. [Reading Subtext: Evaluating Large Language Models on Short Story Summarization with Writers](#). *Trans. Assoc. Comput. Linguistics*, 12:1290–1310.
- Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. [Gemma 2: Improving open language models at a practical size](#). *arXiv preprint arXiv:2408.00118*.
- Craig Thomson and Ehud Reiter. 2020. [A Gold Standard Methodology for Evaluating Accuracy in Data-To-text Systems](#). In *Proceedings of the 13th International Conference on Natural Language Generation, INLG 2020*, pages 158–168, Dublin, Ireland.
- Hadrien Titeux and Rachid Riad. 2021. [pygamma-agreement: Gamma  \$\gamma\$  measure for inter/intra-annotator agreement in Python](#). *Journal of Open Source Software*, 6(62):2989.
- Tina Tseng, Amanda Stent, and Domenic Maida. 2020. [Best Practices for Managing Data Annotation Projects](#). *CoRR*, abs/2009.11654.
- Veniamin Veselovsky, Manoel Horta Ribeiro, and Robert West. 2023. [Artificial Artificial Artificial Intelligence: Crowd Workers Widely Use Large Language Models for Text Production Tasks](#). *CoRR*, abs/2306.07899.
- Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024. [Large Language Models are not Fair Evaluators](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024*, pages 9440–9450, Bangkok, Thailand.
- Lining Zhang, Simon Mille, Yufang Hou, Daniel Deutsch, Elizabeth Clark, Yixin Liu, Saad Mahamood, Sebastian Gehrmann, Miruna Clinciu, Khyathi Raghavi Chandu, and João Sedoc. 2023. [A Needle in a Haystack: An Analysis of High-agreement Workers on MTurk for Summarization](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14944–14982, Toronto, Canada.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. [Judging LLM-as-a-judge with MT-bench and Chatbot Arena](#). In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023*, New Orleans, LA, USA.
- Vilém Zouhar, Tom Kocmi, and Mrinmaya Sachan. 2025. [AI-assisted Human Evaluation of Machine Translation](#). In *Proceedings of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 1: Long Papers, Albuquerque*, pages 4936–4950, New Mexico, USA.

## A Implementation Details

### A.1 Open Models

We run the local models using the `ollama` framework in 4-bit quantization. Specifically, we use `llama3.3:70b` and `deepseek-r1:70b` (which is based on Llama 3.3 70B) for span annotations. We also use `gemma2:2b` and `phi3.5:3.8b` for generating texts in D2T-EVAL.

For better reproducibility, we set the seed to 42 and the temperature to 0 for the local models. We do not use these parameters for proprietary models as these parameters are generally not supported.

We run the models using several GPU variants, including NVIDIA H100 NVL (95G), AMD MI210 (64G), and NVIDIA RTX 3090 (24G).

### A.2 Proprietary Models

We use the following proprietary model versions:

- GPT-4o: `gpt-4o-2024-11-20`
- Claude 3.7 Sonnet:  
`claude-3-7-sonnet-20250219`
- o3-mini: `o3-mini-2025-01-31`
- Gemini 2.0 Flash Thinking:  
`gemini-2.0-flash-thinking-exp-01-21`

### A.3 Web Interface

We implement our span annotation process using `factgenie` (Kasner et al., 2024): a tool that supports both collecting span annotations from humans via a web interface and from LLMs via API calls.

Figure 12 shows samples of our annotation interface implemented in `factgenie` for human annotators, including data visualizations from the football and openweather domains.

## B Annotating D2T-EVAL

### B.1 Generating Outputs

For generating the outputs for the structured inputs we collected, we use two larger models – Llama 3.3 70B (Grattafiori et al., 2024) and GPT-4o (Hurst et al., 2024) – and two smaller models – Gemma 2 2B (Team et al., 2024) and Phi-3.5 3.8B (Abdin et al., 2024). See more details on the models in Appendix A and prompts in Appendix D.

### B.2 Collecting annotations

**Annotation guidelines** For the annotation guidelines, we went through an iterative process to establish the annotation guidelines  $\mathcal{G}$  and the annotation categories  $C$ . We started with a preliminary version of the guidelines and annotation categories, drawing inspiration from the guidelines in previous works (Kasner and Dušek, 2024; Thomson and Reiter, 2020). We settled on the following annotation categories (see Table 5 for details): semantic accuracy errors due to information *Contradictory* to the input, *Not Checkable*, or *Misleading*; any *Incoherent* and *Repetitive* content, and any *Other* errors.

**Gold annotations** With the annotation guidelines established, we proceeded to collect our own internally annotated gold data:  $\mathcal{D}_{\text{dev}}$ , which contains 84 examples annotated individually by one of 7 annotators (12 examples per annotator) and  $\mathcal{D}_{\text{iaa}}$ , which contains 12 examples annotated commonly by all annotators.<sup>10</sup> The purpose of  $\mathcal{D}_{\text{dev}}$  is to create a high-quality development set for the model prompting study, while the purpose of  $\mathcal{D}_{\text{iaa}}$  is to pre-select skilled crowdworkers and quantify the performance of crowdworkers during data collection. Our average IAA on  $\mathcal{D}_{\text{iaa}}$  was  $F_1 = 0.444$  and  $\gamma = 0.399$ .

**Crowdsourcing annotations** We gather span annotations for  $\mathcal{D}_{\text{test}}$  with crowdworkers from `ProLific.com`. Our process of collecting annotations proceeded in two stages, following the setup of Zhang et al. (2023): (1) a *qualification task* for pre-selecting skilled annotators, and (2) the *main task* for collecting the annotations.

- **Qualification task:** For the qualification task, we pre-selected workers whose first language is English, with >99% approval rate and more than 100 submissions. The workers were presented with a detailed tutorial with annotation guidelines and examples of individual errors. After the tutorial, we tested the worker performance on five manually pre-selected examples from  $\mathcal{D}_{\text{iaa}}$ . We invited annotators with the  $F_1$  score higher than 0.5 w.r.t. our internal annotations for the main task.
- **Main task:** Of the 230 annotators who participated in the qualification task, 50 annotators

<sup>10</sup>We selected an example for each of the 4 domains and 3 models.

(21.7%) qualified. Of these, 45 annotators accepted (=90% turnover rate). For annotating the data in  $\mathcal{D}_{\text{test}}$ , we presented each annotator with a batch of 32 examples: 25 examples from  $\mathcal{D}_{\text{test}}$  and 7 remaining examples from  $\mathcal{D}_{\text{iaa}}$  (that is, the examples that we did not use for the qualification task). All the 1200 outputs in  $\mathcal{D}_{\text{test}}$  were annotated by at least one annotator. Furthermore, 475 outputs (39.6%) were annotated by an additional annotator.<sup>11</sup>

For the qualification task, we paid all the annotators an average reward of 9.58 GBP / hour regardless of the qualification outcome. For the main task, we pay all the annotators an average reward of 10.70 GBP / hour.

## C Annotation Categories

Tables 5 to 7 show an overview of the annotation span categories that we used for our tasks along with their descriptions.

Category Name	Description
<i>Contradictory</i>	The fact contradicts the data.
<i>Not checkable</i>	The fact cannot be verified from the data.
<i>Misleading</i>	The fact is technically true, but leaves out important information or otherwise distorts the context.
<i>Incoherent</i>	The text uses unnatural phrasing or does not fit the discourse.
<i>Repetitive</i>	The fact has been already mentioned earlier in the text.
<i>Other</i>	The text is problematic for another reason.

Table 5: Annotation categories for the D2T-EVAL task.

Category Name	Description
<i>Major</i>	An error that disrupts the flow and makes the understandability of text difficult or impossible.
<i>Minor</i>	An error that does not disrupt the flow significantly and what the text is trying to say is still understandable.

Table 6: Annotation categories for the MT-EVAL task.

## D Prompts

Here, we provide the model prompts:

<sup>11</sup>We use examples with two annotators to compute the average IAA for D2T-EVAL in Section 5.2. For other experiments, we use only the outputs from the first annotator as reference data.

- Figures 5 to 8 show the prompts for the D2T-EVAL that we use for the experiments in Section 5.1.
- Figure 9 shows the base prompt we used for MT-EVAL.
- Figure 10 shows the base prompt we used for PROPAGANDA.
- Figure 11 shows the prompt we used for *generating* the outputs for D2T-EVAL.

## E Examples

In Tables 8 to 10, we show examples of the annotated outputs for our tasks. Figure 12 shows our annotation interface.

## F Results

Here, we provide detailed results of our experiments:

- **Main results:** Table 11 (D2T-EVAL), Tables 12 and 13 (MT-EVAL), Table 14 (PROPAGANDA)
- **Extra statistics:** (D2T-EVAL) Tables 13 and 17 (MT-EVAL), Table 18 (PROPAGANDA).
- **Confusion matrices:** Figure 13 (MT-EVAL) and Figure 14 (PROPAGANDA).
- **Manual evaluation:** Table 19 (D2T-EVAL), Table 20 (PROPAGANDA), Table 4 (human annotators).

Category Name	Description
<i>Appeal to Authority</i>	Stating that a claim is true simply because a valid authority or expert on the issue said it was true, without any other supporting evidence offered. We consider the special case in which the reference is not an authority or an expert in this technique, although it is referred to as Testimonial in literature
<i>Appeal to fear-prejudice</i>	Seeking to build support for an idea by instilling anxiety and/or panic in the population towards an alternative. In some cases the support is built based on preconceived judgements
<i>Bandwagon</i>	Attempting to persuade the target audience to join in and take the course of action because "everyone else is taking the same action"
<i>Black-and-White Fallacy</i>	Presenting two alternative options as the only possibilities, when in fact more possibilities exist. As an extreme case, tell the audience exactly what actions to take, eliminating any other possible choices (Dictatorship)
<i>Causal Oversimplification</i>	Assuming a single cause or reason when there are actually multiple causes for an issue. It includes transferring blame to one person or group of people without investigating the complexities of the issue
<i>Doubt</i>	Questioning the credibility of someone or something
<i>Exaggeration, Minimisation</i>	Either representing something in an excessive manner: making things larger, better, worse (e.g., "the best of the best", "quality guaranteed") or making something seem less important or smaller than it really is (e.g., saying that an insult was just a joke)
<i>Flag-Waving</i>	Playing on strong national feeling (or to any group; e.g., race, gender, political preference) to justify or promote an action or idea
<i>Loaded Language</i>	Using specific words and phrases with strong emotional implications (either positive or negative) to influence an audience
<i>Name Calling, Labeling</i>	Labeling the object of the propaganda campaign as either something the target audience fears, hates, finds undesirable or loves, praises
<i>Obfuscation, Intentional Vagueness, Confusion</i>	Using words which are deliberately not clear so that the audience may have its own interpretations. For example when an unclear phrase with multiple definitions is used within the argument and, therefore, it does not support the conclusion
<i>Red Herring</i>	Introducing irrelevant material to the issue being discussed, so that everyone's attention is diverted away from the points made
<i>Reductio ad hitlerum</i>	Persuading an audience to disapprove an action or idea by suggesting that the idea is popular with groups hated in contempt by the target audience. It can refer to any person or concept with a negative connotation
<i>Repetition</i>	repeating the same message over and over again so that the audience will eventually accept it
<i>Slogans</i>	A brief and striking phrase that may include labeling and stereotyping. Slogans tend to act as emotional appeals
<i>Straw Men</i>	When an opponent's proposition is substituted with a similar one which is then refuted in place of the original proposition
<i>Thought-terminating Cliches</i>	Words or phrases that discourage critical thought and meaningful discussion about a given topic. They are typically short, generic sentences that offer seemingly simple answers to complex questions or that distract attention away from other lines of thought
<i>Whataboutism</i>	A technique that attempts to discredit an opponent's position by charging them with hypocrisy without directly disproving their argument

Table 7: Annotation categories for the PROPAGANDA task. The categories are adopted from [Da San Martino et al. \(2019\)](#).

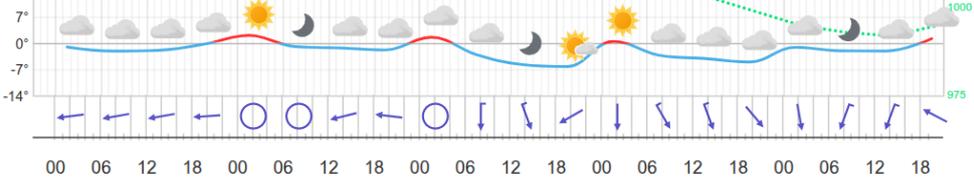
Source (part of the visualization)	
Annotated by Llama 3	Bear Creek will experience mostly overcast skies for the next five days, with <b>a chance of occasional showers on Wednesday and Thursday</b> <sup>not check.</sup> . Temperatures will remain relatively cool, hovering around freezing point or slightly below. Winds will be light to moderate throughout the period, with <b>some gusts expected on Friday</b> <sup>not check.</sup> . The overall weather pattern suggests a generally cloudy and potentially wet week.
Annotated by o3-mini	Bear Creek will experience mostly overcast skies for the next five days, with a <b>chance of occasional showers on Wednesday and Thursday</b> <sup>cont.</sup> . Temperatures will remain relatively cool, hovering around freezing point or slightly below. Winds will be light to moderate throughout the period, <b>with some gusts expected on Friday</b> <sup>cont.</sup> . The overall weather pattern suggests a generally cloudy and <b>potentially wet week</b> <sup>cont.</sup> .
Annotated by human	Bear Creek will experience mostly overcast skies for the next five days, with a chance <b>of occasional showers on Wednesday and Thursday</b> <sup>cont.</sup> . Temperatures will remain relatively cool, hovering around freezing point or slightly below. Winds will be light to moderate throughout the period, <b>with some gusts expected on Friday</b> <sup>mislead.</sup> . The overall weather pattern suggests a generally cloudy and <b>potentially wet</b> <sup>cont.</sup> <b>week</b> <sup>cont.</sup> .

Table 8: Example for span annotation of D2T-EVAL in the weather domain with **contradictory**<sup>cont.</sup>, **misleading**<sup>mislead.</sup>, and **not-checkeable**<sup>not check.</sup> as error span categories. In the annotation interface, the visualization is interactive. The top part (not visible here) shows the place name and date timeline.

Source	“It’s your birthday soon, isn’t it?” Ivory asked, remembering that the princess’ birthday was in a few days. Kari nodded, a sad glint in her light blue eyes.
Annotated by Llama 3	“¿Es su cumpleaños pronto, <b>no lo es?</b> ” <sup>major</sup> Ivory le preguntó, <sup>minor</sup> recordando que el cumpleaños de la princesa era en unos días. Kari nodó, <b>un deslumbramiento triste</b> <sup>major</sup> en sus ojos azules claros.
Annotated by o3-mini	“¿Es su cumpleaños pronto, <b>no lo es?</b> ” <sup>minor</sup> Ivory le preguntó, recordando que el cumpleaños de la princesa era en unos días. Kari nodó, un <b>deslumbramiento triste</b> <sup>major</sup> en sus ojos azules claros.
Annotated by human	“¿Es su cumpleaños pronto, <b>no lo es?</b> ” <sup>minor</sup> Ivory le preguntó, recordando que el cumpleaños de la princesa era en unos días. Kari <b>nodó</b> <sup>major</sup> , un deslumbramiento triste en sus ojos azules claros.

Table 9: An example for span annotation of MT-EVAL outputs (English→Spanish) with **minor** and **major** as error span categories.

Annotated by Llama 3	<b>When the left made Linda Sarsour into its role model, it climbed into bed with</b> <sup>fear</sup> <b>the worst of the worst</b> <sup>labelling</sup> . The father of a missing 4-year-old Georgia boy was training children at <b>a filthy New Mexico compound</b> <sup>loaded</sup> to commit school shootings, prosecutors alleged in court documents Wednesday.
Annotated by o3-mini	<b>When the left made Linda Sarsour into its role model, it climbed into bed with the worst of the worst.</b> <sup>loaded</sup> The father of a missing 4-year-old Georgia boy was training children at <b>a filthy New Mexico compound</b> <sup>loaded</sup> to commit school shootings, prosecutors alleged in court documents Wednesday.
Annotated by human	When the left made Linda Sarsour into <b>its role model</b> <sup>labelling</sup> , <b>it climbed into bed</b> <sup>loaded</sup> <b>with the worst of the worst.</b> <sup>exag.</sup> The father of a missing 4-year-old Georgia boy was training children at <b>a filthy New Mexico compound</b> <sup>labelling</sup> to commit school shootings, prosecutors alleged in court documents Wednesday.

Table 10: Two examples for span annotation of PROPAGANDA outputs with **appeal-to-fear**<sup>fear</sup>, **name-calling-labelling**<sup>labelling</sup>, **loaded-language**<sup>loaded</sup>, and **exaggeration**<sup>exag.</sup> as span categories.

Model	$\rho$	Precision		Recall		F1		$\Delta$	$\gamma$	$S_\emptyset$
		Hard	Soft	Hard	Soft	Hard	Soft			
Llama 3.3	0.307	0.176	0.365	0.187	0.388	0.181	0.377	0.196	0.109	0.418
GPT-4o	0.346	0.233	0.391	0.184	0.308	0.206	0.345	0.139	0.130	0.429
Claude 3.7	<b>0.512</b>	0.294	0.442	<b>0.304</b>	<b>0.457</b>	0.299	0.449	0.150	0.203	0.592
DeepS. R1	0.453	0.317	0.532	0.185	0.310	0.233	0.392	0.159	0.185	<b>0.645</b>
o3-mini	0.505	<b>0.392</b>	<b>0.542</b>	0.285	0.395	<b>0.330</b>	0.457	0.127	<b>0.273</b>	0.637
Gem. 2-FT	0.458	0.293	0.488	0.263	0.438	0.277	<b>0.462</b>	0.185	0.209	0.612

Table 11: Evaluation of human and LLM annotations using  $\mathcal{P}_{\text{base}}$  on D2T-EVAL. See Figure 2 for visualizaton of this table.

Model	$\rho$	Precision		Recall		F1		$\Delta$	$\gamma$	$S_\emptyset$
		Hard	Soft	Hard	Soft	Hard	Soft			
Llama 3.3	0.182	0.121	0.200	0.229	0.378	0.155	0.257	0.102	0.014	0.392
GPT-4o	0.158	0.141	0.240	0.195	0.327	0.156	0.266	0.110	0.076	0.428
Claude 3.7	<b>0.301</b>	<b>0.226</b>	<b>0.325</b>	<b>0.335</b>	<b>0.469</b>	<b>0.262</b>	<b>0.373</b>	0.111	<b>0.131</b>	0.628
DeepS. R1	0.177	0.169	0.268	0.183	0.280	0.168	0.262	0.094	0.058	0.631
o3-mini	0.197	0.169	0.291	0.161	0.275	0.160	0.275	0.115	0.100	0.646
Gem. 2-FT	0.257	0.184	0.312	0.180	0.339	0.173	0.304	0.130	0.066	<b>0.710</b>

Table 12: Evaluation of human and LLM annotations using  $\mathcal{P}_{\text{base}}$  on the MT-EVAL – average across languages.

Model	$\rho$	Precision		Recall		F1		$\Delta$	$\gamma$	$S_\emptyset$
		Hard	Soft	Hard	Soft	Hard	Soft			
en-cs	0.303	0.144	0.268	0.180	0.326	0.156	0.286	0.130	0.084	0.582
en-es	0.171	0.161	0.243	0.236	0.362	0.190	0.288	0.098	0.080	<b>0.631</b>
en-hi	0.170	0.173	0.265	0.208	0.327	0.173	0.269	0.096	-0.0	0.552
en-is	<b>0.347</b>	0.136	0.246	0.187	0.361	0.145	0.269	0.124	0.108	0.493
en-ja	0.127	0.193	0.302	0.249	0.363	<b>0.209</b>	<b>0.318</b>	0.109	0.063	0.569
en-ru	0.225	0.178	0.256	<b>0.273</b>	<b>0.386</b>	0.208	0.298	0.090	<b>0.162</b>	0.588
en-uk	0.192	0.166	0.254	0.214	0.339	0.184	0.286	0.102	0.031	0.542
en-zh	0.163	<b>0.196</b>	<b>0.346</b>	0.163	0.294	0.169	0.302	0.133	0.075	0.623

Table 13: Evaluation of human and LLM annotations using  $\mathcal{P}_{\text{base}}$  on the MT-EVAL separately for each language (average across models).

Model	$\rho$	Precision		Recall		F1		$\Delta$	$\gamma$	$S_\emptyset$
		Hard	Soft	Hard	Soft	Hard	Soft			
Llama 3.3	0.336	0.070	0.243	0.063	0.219	0.066	0.230	0.164	0.092	0.343
GPT-4o	0.344	0.095	0.293	0.038	0.115	0.054	0.166	0.112	0.066	0.234
Claude 3.7	0.460	0.110	0.274	0.096	0.239	0.103	0.255	0.152	0.155	0.113
DeepS. R1	0.354	0.083	0.246	0.062	0.182	0.071	0.209	0.138	0.091	0.476
o3-mini	0.418	<b>0.152</b>	<b>0.411</b>	0.066	0.179	0.092	0.249	0.157	0.154	<b>0.517</b>
Gem. 2-FT	<b>0.560</b>	0.106	0.268	<b>0.190</b>	<b>0.477</b>	<b>0.136</b>	<b>0.343</b>	0.207	<b>0.202</b>	0.493

Table 14: Evaluation of human and LLM annotations using  $\mathcal{P}_{\text{base}}$  on the PROPAGANDA.

Annotator	Ann	Ann/Ex	w/o%	Char/Ann
Human	2981	2.5	28.8	50.3
Llama 3.3	3214	2.7	7.4	65.5
GPT-4o	2284	1.9	4.8	66.3
Claude 3.7	2865	2.4	22.5	57.2
DeepS. R1	1387	1.2	44.2	56.8
o3-mini	1836	1.5	35.6	58.0
Gem. 2-FT	2517	2.1	28.9	54.3

Table 15: Statistics of models and human annotators using  $\mathcal{P}_{\text{base}}$  on D2T-EVAL. Ann=# of annotations, Ann/Ex=ann. per example. w/o=% ex. without annotations, Char/Ann=# chars per ann.

Annotator	Ann.	Ann/Ex	w/o%	Char/Ann
Human	2090	0.7	66.0	14.5
Llama 3.3	6361	2.3	6.2	17.4
GPT-4o	4866	1.7	7.0	15.9
Claude 3.7	3782	1.4	30.6	15.9
DeepS. R1	2586	0.9	36.3	15.1
o3-mini	3039	1.1	35.8	13.8
Gem. 2-FT	2181	0.8	50.0	15.2

Table 16: Statistics of models and human annotators using  $\mathcal{P}_{\text{base}}$  on MT-EVAL. See Table 15 for the legend.

Lang.	Annot.	Ann.	Ann/Ex	w/o%	Char/Ann
en-cs	Model	600	1.4	27.0	16.6
	Human	399	0.7	66.1	13.0
en-es	Model	417	1.1	38.9	18.8
	Human	248	0.6	70.3	10.3
en-hi	Model	396	1.3	26.2	19.0
	Human	222	0.5	71.2	10.7
en-is	Model	563	1.9	14.3	15.7
	Human	752	2.5	18.3	16.6
en-ja	Model	471	1.3	28.7	11.1
	Human	118	0.2	87.5	14.8
en-ru	Model	500	1.3	25.9	18.2
	Human	287	0.7	58.7	19.4
en-uk	Model	436	1.5	25.4	17.8
	Human	208	0.7	64.3	12.3
en-zh	Model	420	1.2	34.6	7.2
	Human	171	0.2	85.1	6.6

Table 17: Statistics of models (averaged) and human annotators using  $\mathcal{P}_{\text{base}}$  on the MT-EVAL separately for each language. See Table 15 for the legend.

Annotator	Ann.	Ann/Ex	w/o%	Char/Ann
Human	1439	14.2	4.0	40.2
Llama 3.3	574	5.7	3.0	92.0
GPT-4o	246	2.4	8.9	91.1
Claude 3.7	803	8.0	7.9	58.5
DeepS. R1	459	4.5	9.9	89.3
o3-mini	376	3.7	10.9	65.3
Gem. 2-FT	1864	18.5	3.0	54.1

Table 18: Statistics of models and human annotators using  $\mathcal{P}_{\text{base}}$  on the PROPAGANDA. See Table 15 for the legend.

Model	Annotations					Explanations				
	C	P	W	I	U	C	P	W	I	U
Llama 3.3	7	1	2	8	0	6	0	1	10	1
GPT-4o	5	2	1	10	0	6	1	0	11	0
Claude 3.7	7	2	3	6	0	9	2	0	7	0
DeepSeek	11	2	4	1	0	12	3	0	3	0
o3-mini	10	3	2	0	3	8	7	0	0	3
Gemini 2 F-T	12	4	1	1	0	12	5	1	0	0
Total	52	14	13	26	3	53	18	2	31	4

Table 19: Manual evaluation results for D2T-Eval domain. Categories for annotation and reason: C=Correct, P=Partially correct, W=Wrong category, I=Incorrect, U=Undecidable.

Model	Annotations					Explanations				
	C	P	W	I	U	C	P	W	I	U
Llama 3.3	9	1	2	5	1	8	2	1	6	1
GPT-4o	9	0	3	6	0	8	2	1	7	0
Claude 3.7	9	2	3	4	0	9	3	3	3	0
DeepSeek	6	0	4	8	0	7	0	3	8	0
o3-mini	15	0	0	2	1	15	0	0	2	1
Gemini 2 F-T	7	3	2	6	0	9	2	1	6	0
Total	55	6	14	31	2	56	9	9	32	2

Table 20: Manual evaluation results for Propaganda domain. See Table 19 for the legend.

Your task is to identify errors in the text and classify them.

Output the errors as a JSON object with a single key "annotations". The value of "annotations" is a list in which each object contains fields "reason", "text", and "annotation\_type". The value of "reason" is the short sentence justifying the annotation. The value of "text" is the literal value of the identified span (we will later identify the span using string matching). The value of "annotation\_type" is an integer index of the error based on the following list:

```
{categories}
```

Examples:

- Contradictory: The lowest temperature does not drop below 4°C, but the text mentions 2°C.
- Not checkable: The text mentions that "both teams display aggressive play", which cannot be checked from the data.
- Misleading: The tone of the text suggests there are many sensors out of which just a few are listed here. However, according to the data, the device has exactly these four sensors.
- Incoherent: The text states that both teams had "equal chances until the first half ended scoreless." While this is technically true, the expression sounds unnatural for a sport summary.
- Repetitive: The output text unnecessarily re-states information about a smartphone battery that was mentioned earlier.
- Other: Use this as a last resort when you notice something else not covered by the above categories.

Hints:

- Always try to annotate the longest continuous span (i.e., the whole fact instead of a single word).
- Annotate only the spans that you are sure about. If you are not sure about an annotation, skip it.
- Ignore subjective statements: for example "a lightweight smartphone" highly depends on the context: you should not annotate these statements.
- Numerical conventions: For weather forecasts, we accept both precise numbers (e.g. 10.71°C) and the rounded ones (e.g. 11°C) as long as they agree with the data.
- Annotate only according to your own knowledge. If you are considering using an external tool (Google, ChatGPT etc.), just skip that specific fact.
- If there is nothing to annotate in the text, "annotations" will be an empty list.

Given the data:

```
{data}
```

annotate the errors in the corresponding text generated from the data:

```
{text}
```

Figure 5: The prompt  $\mathcal{P}_{\text{base}}$  for D2T-EVAL.

Your task is to identify errors in the text and classify them.

Output the errors as a JSON object with a single key "annotations". The value of "annotations" is a list in which each object contains fields "reason", "text", and "annotation\_type". The value of "reason" is the short sentence justifying the annotation. The value of "text" is the literal value of the identified span (we will later identify the span using string matching). The value of "annotation\_type" is an integer index of the error based on the following list:

```
{categories}
```

Given the data:

```
```\n
```

```
{data}
```

```
```\n
```

annotate the errors in the corresponding text generated from the data:

```
```\n
```

```
{text}
```

```
```\n
```

Figure 6: The prompt  $\mathcal{P}_{\text{noguide}}$  for D2T-EVAL.

Think about it step-by-step. You should enclose your chain of thoughts between the <think> and </think> tags. Once you are ready, output the JSON object in the required format.

Example:

```
```\n
```

```
<think> ... chain of thoughts ... </think> ...  
JSON object ...
```

```
```\n
```

Figure 7: The additional text added for  $\mathcal{P}_{\text{cot}}$ .

In order to help you with the task, we provide you with five examples of inputs, outputs and annotations:

Example #1:

data:

```
```\n
```

```
{data}
```

```
```\n
```

text:

```
```\n
```

```
{text}
```

```
```\n
```

output:

```
```\n
```

```
{annotations}
```

```
```\n
```

```
(...)
```

Figure 8: The additional text added for  $\mathcal{P}_{\text{5shot}}$ .

Your task is to identify errors in the translation and classify them.

Output the errors as a JSON object with a single key "annotations". The value of "annotations" is a list in which each object contains fields "reason", "text", and "annotation\_type". The value of "reason" is the short sentence justifying the annotation. The value of "text" is the literal value of the identified span (we will later identify the span using string matching). The value of "annotation\_type" is an integer index of the error based on the following list:

{categories}

Error spans can include parts of the words, whole words, or multi-word phrases.  
Hint: errors are usually accuracy-related (addition, mistranslation, omission, untranslated text), fluency-related (character encoding, grammar, inconsistency, punctuation, register, spelling), style-related (awkward, terminology (inappropriate for context, inconsistent use).

Make sure that the annotations are not overlapping. If there is nothing to annotate in the text, "annotations" will be an empty list.

Given the source:

{source}

annotate its translation:

{text}

Figure 9: The prompt  $\mathcal{P}_{\text{base}}$  for MT-EVAL.

Your task is to identify spans of text that employ propaganda techniques.

Output the errors as a JSON object with a single key "annotations". The value of "annotations" is a list in which each object contains fields "reason", "text", and "annotation\_type". The value of "reason" is the short sentence justifying the annotation. The value of "text" is the literal value of the identified span (we will later identify the span using string matching). The value of "annotation\_type" is an integer index of the error based on the following list:

{categories}

Now annotate the following text:

{text}

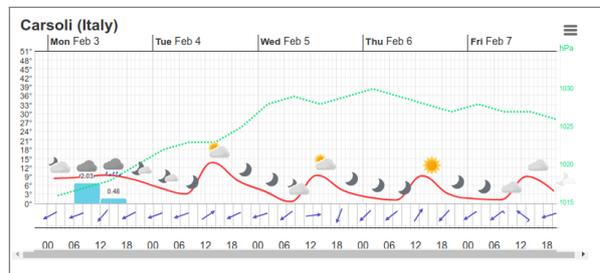
Figure 10: The prompt  $\mathcal{P}_{\text{base}}$  for PROPAGANDA.

Given the structured summary of a football game:

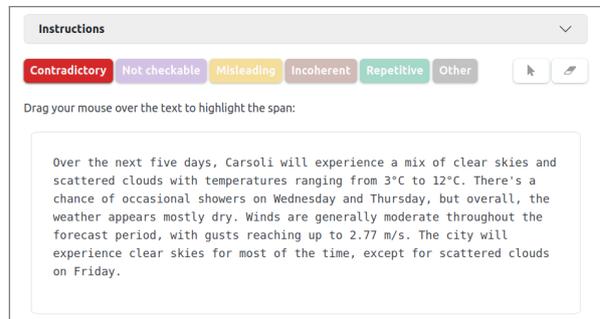
{data}

Generate a match summary using approximately five sentences. The summary should sound natural, reporting on the important moments of the game. Avoid subjective statements, keep the tone of the summary neutral. Do not fabricate any facts that are not explicitly stated in the data.

Figure 11: The prompt used for generating outputs in the football domain for D2T-EVAL. The prompts for the other domains are analogical. For more robust parsing, we initialize the model response with 'Sure, here is the summary: "' .



(a) Data visualization – openweather



(b) Interface for highlighting spans

Figure 12: Samples from the factgenie annotation interface used for collecting span annotations.

Human annotators	Major	162	134
	Minor	139	234
		Major	Minor
		Model predictions	

Figure 13: Confusion matrix for MT-EVAL, averaged across models (see Table 6 for category descriptions).

Human annotators	Thought-terminating_Cliches	Whataboutism	Straw_Men	Repetition	Reductio_ad_hitlerum	Red_Herring	Intentional_Vagueness,Confusion	Name_Calling,Labeling	Loaded_Language	Flag-Waving	Exaggeration,Minimisation	Doubt	Causal_Oversimplification	Black-and-White_Fallacy	Bandwagon	Appeal_to_fear-prejudice	Appeal_to_Authority	
Thought-terminating_Cliches	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Whataboutism	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Straw_Men	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Repetition	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Reductio_ad_hitlerum	0	3	0	2	1	1	3	0	6	3	0	0	0	0	0	0	0	0
Red_Herring	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
Intentional_Vagueness,Confusion	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Name_Calling,Labeling	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Loaded_Language	1	5	0	0	1	0	1	0	25	27	0	0	0	0	0	0	0	0
Flag-Waving	1	8	0	1	2	2	11	1	58	17	0	1	0	0	0	0	1	0
Exaggeration,Minimisation	1	2	1	0	0	0	1	5	5	2	0	0	0	0	0	0	0	0
Doubt	0	2	0	0	0	0	7	0	10	2	0	0	0	0	0	0	0	0
Causal_Oversimplification	0	1	0	0	0	4	1	0	6	2	0	0	0	0	0	0	0	0
Black-and-White_Fallacy	0	0	0	0	2	1	1	0	3	2	0	0	0	0	0	0	0	0
Bandwagon	0	1	0	2	0	0	0	0	1	0	0	0	0	0	0	0	0	0
Appeal_to_fear-prejudice	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Appeal_to_Authority	2	4	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0

Figure 14: Confusion matrix comparing human annotations (rows) with model predictions (columns) for PROPAGANDA, averaged across models. (see Table 7 for the description of categories).

# On the Credibility of Evaluating LLMs using Survey Questions

Jindřich Libovický

Charles University, Faculty of Mathematics and Physics,  
Institute of Formal and Applied Linguistics  
V Holešovičkách 747/2, 180 00 Praha, Czechia  
libovicky@ufal.mff.cuni.cz

## Abstract

Recent studies evaluate the value orientation of large language models (LLMs) using adapted social surveys, typically by prompting models with survey questions and comparing their responses to average human responses. This paper identifies limitations in this methodology that, depending on the exact setup, can lead to both underestimating and overestimating the similarity of value orientation. Using the World Value Survey in three languages across five countries, we demonstrate that prompting methods (direct vs. chain-of-thought) and decoding strategies (greedy vs. sampling) significantly affect results. To assess the interaction between answers, we introduce a novel metric, self-correlation distance. This metric measures whether LLMs maintain consistent relationships between answers across different questions, as humans do. This indicates that even a high average agreement with human data, when considering LLM responses independently, does not guarantee structural alignment in responses. Additionally, we reveal a weak correlation between two common evaluation metrics, mean-squared distance and KL divergence, which assume that survey answers are independent of each other. For future research, we recommend CoT prompting, sampling-based decoding with dozens of samples, and robust analysis using multiple metrics, including self-correlation distance.

## 1 Introduction

Evaluating the values expressed in texts generated by Large Language Models (LLMs) is crucial for shaping public perception, informing policy, and ensuring the ethical use of AI. A common evaluation method involves prompting LLMs with questions from standardized surveys and comparing their responses to human answers or calibrated scores (see Table 1 for a comprehensive list of related work).

This approach has been criticized for its inconsistency with open-ended generation (Wright et al., 2024) and sensitivity to prompt formulation (Röttger et al., 2024; Motoki et al., 2024). We address a related issue: the reliability of measuring similarity between model

and human responses, particularly in how answers to different questions correlate with one another.

We examine this framework using (1) direct vs. Chain-of-Thought (CoT) prompting, (2) greedy decoding vs. nucleus sampling, and (3) three similarity metrics: mean squared difference, KL divergence, and our novel self-correlation distance. Unlike previous metrics that estimate average alignment by treating survey questions in isolation, our metric accounts for interactions among survey answers.

Using LLaMA 3, Mistral 2, EuroLLM, and Qwen 2.5, we compare model responses in the World Value Survey (WVS; Haerper et al., 2020) with opinions from selected countries. We find that choices in prompting, decoding, and metrics yield different conclusions. For example, greedy decoding deviates from typical results obtained with nucleus sampling, while short categorical answers underestimate alignment compared to CoT prompting. The self-correlation distance indicates that, despite a high average alignment with survey data, different correlation patterns reveal potential overgeneralization. Metrics that treat answers independently can thus overestimate alignment.

In this paper, we first review studies that compare LLM-generated answers with population survey responses (Section 2). We then describe our experiments (Section 3), including the models we use, the prompts we employ, and the inference algorithm we employ. We describe the metrics we use, including the newly introduced self-correlation distance. Section 4 presents the results, and Section 5 concludes the paper.

## 2 LLMs and Surveys

Many studies examine the values encoded in language models, focusing on moral, cultural, and political biases. These often rely on surveys designed for human respondents (see Table 1). They cover topics ranging from general ethics to specific domains, such as political ideologies (Feng et al., 2023), autonomous vehicle ethics (Vida et al., 2024), and religious values (Liu et al., 2024).

Model responses are typically evaluated using two approaches: (1) evaluation keys, as in frameworks like the Political Compass or Moral Foundations Questionnaire (Graham et al., 2011; MFQ), or (2) comparisons with human population data.

Most studies prompt models to produce categorical

	Gold survey data	Open-source models	Persona	Decoding (# samples)	Categorical output	Comparison w/ humans
Santurkar et al. (2023)	Custom	No	Yes	Logits	Yes	WD
Cao et al. (2023)	Hofstede (1984)	No	No	Sampling (1)	Yes	Accuracy
Feng et al. (2023)	Political compass	Yes	No	10-best tokens	No	—
Olmedo et al. (2023)	Mather et al. (2005)	Yes	No	Logits	Yes	KL
Scherrer et al. (2023)	Custom	Yes	No	Full samp. (5/10)	No	—
Sanders et al. (2023)	Custom	No	Yes	Sampling? (100)	Yes	WD
Benkler et al. (2023)	WVS	No	Yes	Nucleus samp. (1)	No	—
Tao et al. (2024)	WVS, EVS	No	Yes	Greedy	Yes	MSD
Durmus et al. (2024)	Custom	No	Yes	Logits	Yes	JSD
Ceron et al. (2024)	Custom	Yes	Yes	Nucleus samp. (30)	Yes	—
Nunes et al. (2024)	MFQ, MFV	Yes	No	Sampling? (1)	Yes	—
Vida et al. (2024)	Awad et al. (2018)	Yes	No	Sampling? (1)	Yes	Accuracy
Xu et al. (2024)	WVS	Yes	Yes	Nucleus samp. (1)	Yes	Norm. MSD
Liu et al. (2024)	Center (2018)	Yes	No	Greedy	Yes	—
Kim and Baek (2024)	WVS	Yes	No	Logits	Yes	Pearson
Aksoy (2024)	MFQ	Yes	No	Sampling? (100)	Yes	—
Shen et al. (2024)	Schwartz (1992)	Yes	No	Sampling? (10)	Yes	L1
Sukiennik et al. (2025)	Hofstede (1984)	Yes	No	Greedy	Yes	Norm. L1
Qu and Wang (2024)	WVS	No	Yes	Sampling (100)	Yes	MSD
Kazemi et al. (2024)	WVS	No	No	Not specified	Yes	Accuracy
Rupprecht et al. (2025)	WVS	Yes	No	Logits	Yes	Accuracy
Gurgurov et al. (2025)	Political compass	Yes	No	Nucleus sampling (1)	Yes	—
Bulté and Rigouts (2025)	(Hofstede, 1984) + WVS	Yes	No	Nucleus sampling (6)	Yes	Accuracy
Costa et al. (2025)	MFQ	Yes	Yes	Sampling? (1)	Yes	—
Atari et al. (2023)	WVS	No	No	Sampling? (100)	Yes	Fixation ind.

*Gold survey data:* WVS = World Value Survey (Haerpfer et al., 2020), EVS = European Value Study (EVS, 2022), MFQ = Moral Foundation Questionnaire (Graham et al., 2011), MFV = Moral Foundation Vignettes (Clifford et al., 2015)  
*Comparison methods:* WD = Wasserstein Distance, Acc. = Accuracy, KL = Kullback-Leibler Divergence, JSD = Jensen-Shannon Distance, MSD = Mean Squared Difference

Table 1: Overview of publications using standardized surveys to evaluate values in LLMs.

answers, such as selecting an option or providing a score. These responses are compared to population data using metrics like KL Divergence (Olmedo et al., 2023), Jensen-Shannon Divergence (Durmus et al., 2024), or Wasserstein Distance (Santurkar et al., 2023; Sanders et al., 2023). However, these methods often focus on single-token generation, which limits their generalization to longer text samples. Some evaluations use greedy decoding or a single sampled response, with sampling details often unspecified. Persona probing (i.e., specifying demographic traits for models to emulate) is also common but typically focuses on English outputs, leaving biases in other languages underexplored.

Several recent studies have extended survey-based evaluation methodologies. Sukiennik et al. (2025) conducted the first large-scale evaluation of cultural alignment across 20 countries and 10 LLMs, using Hofstede’s Cultural Values Questionnaire. The results found that models generally represent a moderate cultural middle ground, with the United States showing the best alignment. Qu and Wang (2024) used WVS data to evaluate ChatGPT’s public opinion simulation capabilities, revealing significant performance disparities favoring Western, English-speaking nations and demographic biases across gender, ethnicity, and social class. Kazemi et al. (2024) demonstrated that 44% of GPT-4o’s ability to reflect societal values correlates with digital resource

availability in a society’s primary language, with error rates in low-resource languages exceeding those in high-resource languages by a factor of five. Most recently, Rupprecht et al. (2025) extended bias research in LLM survey responses using WVS data, testing perturbations in answer and question phrasing across multiple models and finding significant sensitivity to prompt variations that mirror human response biases.

Röttger et al. (2024); Wang et al. (2024) and Moore et al. (2024) highlight how prompt formulations, such as multiple-choice setups, significantly influence model outputs and score robustness. While their work examines the robustness of the prompt with respect to evaluation keys (such as the Political Compass), it does not address the methodological aspects of comparing model outputs to population survey data. Also, as far as we know, all previous work treats survey responses as independent and disregards correlations between questions, an issue we address by introducing the self-correlation distance.

Previous work also evaluated generation consistency (Kumar and Joshi, 2022; Bonagiri et al., 2024). However, it focuses on cases with a well-specified ground truth. We are interested in a slightly different type of consistency: Statements like people who tend to say *A*, also tend to say *B* to some extent. In this paper, we introduce a metric to measure the extent to which LLMs

capture these tendencies, independent of the actual content.

Recent work has also examined cultural adaptation methods using survey data. [Adilazuarda et al. \(2025\)](#) found that WVS-based training can lead to cultural homogenization and undermine factual knowledge, and introduced a cultural distinctiveness metric that complements existing evaluation approaches. Their findings that survey data alone may be insufficient for cultural adaptation align with our observations about the limitations of current evaluation methodologies.

### 3 Experiments

Following several previous studies ([Benkler et al., 2023](#); [Tao et al., 2024](#); [Kim and Baek, 2024](#)), we prompt LLMs with World Value Survey (Round 7, version 5.0) questions and compare their answers with human data using three evaluation methods. WVS is a global research project that has comprehensively explored people’s values and beliefs since 1981. The World Value Survey covers 55 countries and 80 languages, making it likely the most comprehensive standardized resource for comparing values in LLM outputs with the human population across languages and cultures. In this work, we focus on evaluation metrics and conduct experiments only on a small subset.

As in related work, we simulate the survey using an LLM and compare the results with those obtained from the human population in the respective countries. Since the answers to all questions are integers, depending on the evaluation metric, we either use the average answers or the categorical distributions of answers as the ground truth for comparison. For the correlation study, we use responses from individual respondents and compute how responses to individual questions correlate with each other.

The source code to replicate the experiments is available at <https://github.com/jlibovicky/llm-survey-eval>.

#### 3.1 Model Prompting

**Questionnaire design.** WVS is not a self-assessment questionnaire. Interlocutors interview the subjects and, based on their answers, record integer scores for each question, most often on a scale from 1 to 10, indicating the extent to which they agree with a statement.

We use general, non-personalized formulations of the questions, i.e., we exclude questions about income, health, or personal experiences, which would likely be rejected by the models. We reformulated the questions to contain more general, non-personal statements (e.g., replacing “your life” with “human life”) to further reduce rejections. After excluding questions that were not in all language versions, 143 prompts remained. Prompts were created in English, machine-translated into German and Czech using Google Translate, and then manually post-edited with reference to official WVS translations by native speakers. The question-

naire was administered in a single conversation session to allow evaluation of answer consistency (see examples in Appendix A).

**Scores vs. Chain-of-Thought.** We compare two prompt types: direct numeric answers and chain-of-thought prompts ([Wei et al., 2022](#)), where justification precedes the answer. We posit that chain-of-thought is closer to real-world LLM use, as chat-like user interactions typically involve longer generations than a single categorical output, e.g., in interactive sessions.

**Greedy vs. Sampling.** Studies often use greedy decoding for its efficiency and for producing a deterministic output. It approximates the most probable sequence but may not yield typical responses, as the probability mass is distributed across similar sequences ([Eikema and Aziz, 2020](#); [Wiher et al., 2022](#)). Because of this, and since sampling is more common in practice, we compare greedy decoding with nucleus sampling (nucleus 0.9, temperature 0.7), including an estimation of how many samples are needed to obtain a stable result for the given metrics. Following [Andreas \(2022\)](#), who argues that language models should be treated as ensembles of different multiple agents, and [Lederman and Mahowald \(2024\)](#), who argue that language models are compressed libraries, we assume that sampling one conversation session might correspond to one agent within the language model in the agent metaphor and retrieving one set of world knowledge from the library in the library metaphor. Therefore, we treat the conversation sessions as comparable to individual respondents in the survey.

#### 3.2 Evaluation

We evaluate model outputs against data from the USA, UK, Czechia, and Germany, where the prompt languages are spoken. Iran and China are included as culturally distinct cases, a sanity check for metric validation.<sup>1</sup> Based on the results of previous studies, we expect that due to the prevalence of English data, models will tend to better align with Western countries. We use three metrics: Mean Squared Difference (MSD), Kullback-Leibler Divergence (KLD), and a novel self-correlation distance.

**Models.** We use four instruction-tuned models: LLaMA 3 8B Instruct ([Dubey et al., 2024](#)) and Mistral v0.1 7B Instruct ([Jiang et al., 2023](#)), EuroLLM 9B Instruct ([Martins et al., 2024](#)), and Qwen 2.5 7B Instruct ([Yang et al., 2024](#)). This selection includes both general-purpose models (LLaMA 3 from the USA, Mistral from France) and models with specific regional focuses (EuroLLM for European languages, Qwen developed in China for multilingual applications), allowing us to examine how model design influences value alignment across cultures.

<sup>1</sup>The number of WVS participants in the respective countries was: USA: 2,596, UK: 2,609, Czechia: 1,200, Germany: 1,528, Iran: 1,499, China: 3,036.

**Mean Square Difference.** Most WVS questions have numerical answers on a 1-to-10 scale or lower. We scale the answer to the 0–1 interval, compute the squared differences between the scores sampled from the model and human population averages, and average them over all questions.

**KL Divergence.** We treat the model and human answers as categorical distributions. For the model outputs, we normalize the distribution over model runs. For the survey data, we normalize over the participants. We compute the Kullback-Leibler divergence between the sampled model answer distribution and the distribution of answers in the human population.

**Self-Correlation Distance.** The previous metrics fail to account for the fact that questionnaire responses often correlate with one another due to underlying consistency in values and opinions, yet all questions are treated as conditionally independent. This assumption is not realistic. Value opinions often come in bundles and follow patterns that may differ across various societies. Simple examples might include people who believe that religion should play a stronger role in society being more likely to say that mothers of young children should stay at home with their children, or that individuals concerned about social justice are often also concerned about the environment. These correlations between individual respondents’ answers capture second-order patterns, relationships between answers, that are not apparent when comparing individual answers alone, as with MSE or KL-Divergence.

To analyze this, we compute self-correlation matrices that measure the Pearson correlation between all pairs of questionnaire responses. In the survey data, we calculate the correlation between responses to individual questions across participants. In the LLM case, we calculate the correlation between answers across model runs. This gives a matrix with all question pairs. Using the Frobenius norm, we quantify how internally consistent or “principled” the answers are. High norm means that the absolute value of the correlations tends to be high, whereas a norm close to zero means that questions in the survey are more independent of each other.

Additionally, we compare the self-correlation matrices of model outputs and human responses using the Frobenius norm. This metric allows us to evaluate whether the underlying structure of the answers aligns between models and humans, going beyond simple agreement on individual responses taken independently.

## 4 Results

**Comparing MSD, KLD, and self-correlation distance.** The results comparing model responses to human surveys are presented in Table 2 for the USA, with additional results for other countries and languages provided in Table 6 in the Appendix. Within languages and countries, results follow similar trends across setups.

To interpret our results, we first establish baseline values from human populations. Country-level comparisons in the WVS yield MSD values ranging from 0.009 (USA-UK) to 0.069 (Germany-Iran), with Western countries showing differences of 0.009–0.024 (see Table 7 in the Appendix). KLD between countries ranges from 0.07 (USA-UK) to 0.44 (Germany-Iran), with Western countries showing 0.07–0.22. Self-correlation distances between human populations range from 0.64 (China-Iran) to 0.95 (USA-Iran), with typical values between 0.79 and 0.92.

Using MSD and KLD metrics, which treat answers independently, we observe substantial variation across prompting and decoding strategies. With CoT prompting and sampling, Mistral 2 achieves remarkably low MSD (0.022) and KLD (0.26) for USA data, comparable to differences between Western countries. However, the same model with greedy decoding and CoT prompting shows drastically worse alignment (MSD=0.188), exceeding even USA-Iran differences. LLaMA 3 shows more moderate values (MSD=0.059, KLD=1.47 for CoT+Sampling), falling between Western and cross-cultural differences. EuroLLM exhibits the poorest alignment with direct prompting and greedy decoding (MSD=0.165–0.284), though sampling decreases the distance between model and population substantially. Qwen demonstrates relatively stable performance across setups (MSD=0.041–0.199).

The self-correlation distance metric reveals a paradox: setups that achieve the best surface-level alignment often exhibit the poorest structural alignment. Mistral 2, with CoT and sampling, despite having the lowest MSD (0.022) and KLD (0.26), exhibits a self-correlation distance of 1.62, which is far greater than the distances between any human populations (0.64–0.95). Its correlation norm (2.80) is also substantially higher than human values ( $\sim 1.66$ ), which indicates overly rigid response patterns. In contrast, LLaMA 3 with the same setup shows better structural alignment (self-correlation distance=1.29, correlation norm=1.70) despite worse surface metrics (MSD=0.059, KLD=1.47). Qwen, with score-only prompts and sampling, produces the most structured responses (correlation norm=3.33, self-correlation distance=2.13) with an even greater departure from human response variability.

Across all models and prompting strategies, greedy decoding consistently underestimates alignment when measured with MSD. For instance, comparing LLaMA 3 with score-only prompts, greedy decoding yields MSD=0.098, versus 0.073 with sampling, roughly a one-third increase. The disparity is even more pronounced for KLD, where sampling values are often 2–3 $\times$  higher than greedy decoding, indicating that greedy decoding captures only a narrow slice of the probability distribution reachable by common decoding algorithms.

The effect of CoT prompting varies across models and decoding strategies. For LLaMA 3 with sampling, CoT improves alignment (MSD decreases from 0.073 to 0.059). For Mistral 2 with sampling, CoT dramat-

Model	Prompt type	Decode	MSD	KLD	Corr. norm	Self-corr. dist.
LLaMA 3	Score only	Gr.	.098	1.71	0.90	1.26
		Spl.	.073	2.99		
LLaMA 3	CoT	Gr.	.088	1.68	1.70	1.29
		Spl.	.059	1.47		
Mistral2	Score only	Gr.	.094	1.80	1.77	1.17
		Spl.	.041	0.77		
Mistral2	CoT	Gr.	.188	1.95	2.80	1.62
		Spl.	.022	0.26		
EuroLLM	Score only	Gr.	.165	1.91	2.55	1.56
		Spl.	.059	0.72		
EuroLLM	CoT	Gr.	.130	1.76	2.97	1.74
		Spl.	.125	0.97		
Qwen	Score only	Gr.	.082	1.66	3.33	2.13
		Spl.	.074	1.27		
Qwen	CoT	Gr.	.199	1.60	1.52	1.25
		Spl.	.062	1.13		

Table 2: A comparison of model outputs with different prompting strategies (Score-only, CoT: Chain of Thought), decoding method (Gr.: greedy, Spl.: sampling). For other countries and languages, see Table 6 in the Appendix.

	MSD	KLD	CorrD
MSD	—	.465	-.389
KLD	.717	—	-.083
CorrD	-.374	-.064	—

Table 3: Correlation of the metrics (Pearson above the diagonal, Spearman under the diagonal) for both models over all languages and countries.

ically improves both MSD (0.041→0.022) and KLD (0.77→0.26). However, with greedy decoding, CoT can worsen results: Mistral 2’s MSD increases from 0.094 to 0.188. This suggests that CoT prompting is beneficial primarily when combined with sampling-based decoding, likely because it provides more stable generation that better leverages the sampling strategy.

Table 3 shows moderate overall correlation between MSD and KLD (Pearson=0.465, Spearman=0.717), but Table 5 reveals this masks dramatic model-specific differences. Mistral 2 shows a very strong correlation (Pearson=0.832, Spearman=0.925), meaning MSD and KLD largely agree on what constitutes good alignment for this model. LLaMA 3, however, shows weak correlation (Pearson=0.276, Spearman=0.389), indicating that these metrics measure somewhat different aspects of value alignment for this model. Critically, self-correlation distance is negatively correlated with both MSD (Pearson=−0.389) and KLD (Pearson=−0.083), confirming that surface-level and structural alignment are different, and sometimes opposing, qualities.

Model	Prompt type	Decode	MSD		KLD		Self-corr. dist.	
LLaMA 3	Score only	Gr.	-.19	.18	-.07	.21	.02	-.25
		Spl.	-.08	.24	.08	.24		
LLaMA 3	CoT	Gr.	-.12	.26	-.29	.21	-.12	-.28
		Spl.	-.03	.34	-.17	.12		
Mistral2	Score only	Gr.	.26	.23	.25	.36	.07	.01
		Spl.	.01	.33	-.08	.17		
Mistral2	CoT	Gr.	-.16	.08	-.01	.13	.19	.21
		Spl.	.24	.26	.21	.20		
EuroLLM	Score only	Gr.	.37	.16	.30	.08	.02	.03
		Spl.	.24	-.01	.18	.07		
EuroLLM	CoT	Gr.	.28	.12	-.18	-.16	-.17	-.00
		Spl.	.07	-.26	.15	-.15		
Qwen	Score only	Gr.	.46	.45	-.20	-.07	-.17	-.15
		Spl.	-.26	-.00	-.14	.03		
Qwen	CoT	Gr.	-.12	.11	-.10	.03	.03	-.16
		Spl.	-.05	.28	-.12	.09		

Color scale from -.3 through 0 to .3

Table 4: Point-Biserial Correlation of the alignment metrics with matching country and language as a binary indicator variable. The left part shows the correlation only within Western countries; the right part also includes Iran and China.

**Cross-lingual and cross-cultural patterns.** Table 4 summarizes how language-country matching affects alignment metrics. The correlations are generally weak, but patterns emerge when culturally distant countries (Iran and China) are included. Most models exhibit positive correlations between matching conditions and alignment metrics, suggesting that prompts in English, Czech, and German align more closely with Western surveys than with those from Iran or China. However, the strength of this effect varies: EuroLLM shows the strongest language-country matching effects (up to 0.37), likely reflecting its European-centric training, while Qwen exhibits more uniform performance across language-country combinations, consistent with its multilingual design.

Notably, the correlation patterns differ substantially across prompting and decoding strategies. Mistral 2 with CoT prompting and nucleus sampling shows the most consistent positive correlations with both language matching (0.24 for MSD, 0.21 for KLD) and the inclusion of distant cultures (0.26 for MSD, 0.20 for KLD). This suggests that this setup not only achieves the best average alignment but also shows more predictable cross-cultural patterns.

**Self-correlation patterns.** Figure 1 visualizes the self-correlation matrices, revealing systematic differences in how models structure their responses compared to humans. The visualization shows three major patterns: (1) Block diagonal structures indicate question clusters with strong internal correlations. (2) The intensity of col-

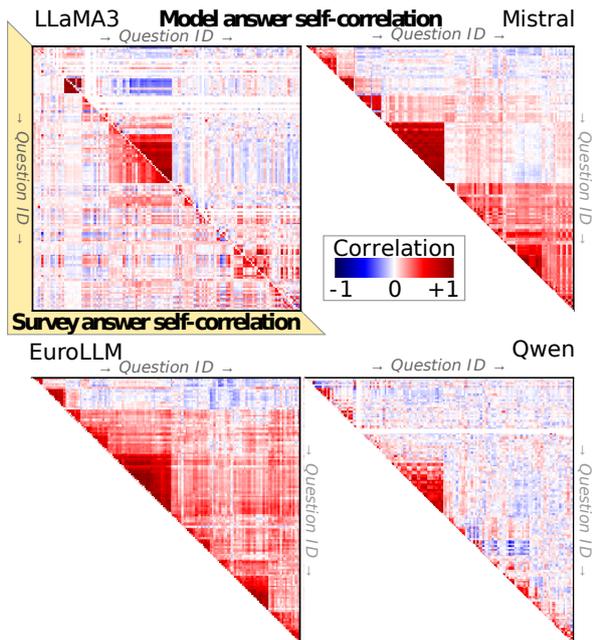


Figure 1: Correlation patterns between human answers in the USA (under the diagonal) and between answers of the LLaMA 3 and Mistral 2 models in English (above the diagonal).

ors shows correlation strength: models produce darker, more saturated colors than humans. (3) Off-diagonal patterns reveal how different value domains relate; models show simpler, more predictable patterns than the human answers.

All models show overly consistent responses in sections about cultural identity and national pride (visible as darker red blocks along the diagonal). Mistral 2 displays particularly strong self-correlation in questions about social and political attitudes, far exceeding human patterns. These visualizations confirm that, while models may match human averages on individual questions, their internal response structures systematically diverge from human response patterns.

The quantitative self-correlation distances reinforce this finding: model-to-human distances (1.1–2.1 across setups) consistently exceed human-to-human distances (0.64–0.95), indicating that all tested models impose more rigid correlation structures than exist in human populations. This suggests that models generate “principled” but overly simplistic response patterns, potentially missing the nuanced and sometimes inconsistent nature of human value systems.

**Number of sampled responses.** Figure 2 demonstrates how metric estimates stabilize with increasing sample size. For MSD, approximately 100 samples are needed for stable estimates, with values converging to within 0.005 of the final estimate. KLD requires more samples, several hundred for full stability. CoT prompting produces more stable estimates with fewer samples compared to direct prompts, likely because the

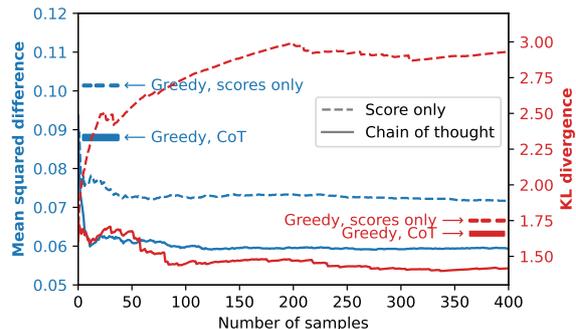


Figure 2: Mean-squared difference and KL-divergence of LLaMA 3 answers when compared to the USA data of the World Value Survey. It compares the greedy decoding and sampling from the model.

reasoning process provides additional structure that reduces sampling variance. Importantly, greedy decoding produces estimates that deviate significantly from the sampling-based norm across both metrics and all sample sizes. This confirms that greedy decoding systematically misestimates both average alignment and response distributions. Most prior studies (Table 1) used far fewer samples (typically 1–10), suggesting their conclusions may have differed substantially with more comprehensive sampling, particularly for KLD estimates.

**Model-specific behaviors.** The four tested models show distinct patterns. Mistral 2 achieves the best surface-level scores but at the cost of the highest structural rigidity, suggesting it may be overfitting to typical responses. LLaMA 3 shows the most balanced profile across metrics. EuroLLM’s strong language-matching effects (correlation of up to 0.37) reflect its European-centric training data. Qwen, despite being developed in China, shows relatively uniform cross-lingual performance but produces highly structured responses (correlation norms up to 3.33).

## 5 Conclusions

This study examined the impact of decoding strategies and evaluation metrics on comparisons of LLM responses to population survey data, using the World Value Survey as a case study across three languages and six countries.

We found that setups closely mirroring real-world LLM usage, specifically, Chain-of-Thought prompting with sampling-based decoding, achieve the best alignment with survey data. **Prior work relying on direct prompts and greedy decoding may underestimate average alignment when evaluating answers independently.**

To address gaps in current evaluations, we introduced the self-correlation distance, a novel metric that captures consistency and interaction between answers. Unlike traditional metrics such as MSD and KL Divergence, the self-correlation distance showed that high scores in

some setups indicate a lower diversity of responses than in the human population. The high average numbers are achieved at the cost of generating typical cases, resulting in overly structured responses rather than an accurate reflection of survey variability, especially in social and political topics. **This shows that metrics treating answers independently overestimate the alignment.** Discrepancies between metrics, such as LLaMA 3's low correlation between MSD and KLD (Pearson=0.276), underscore the importance of multi-metric evaluations.

For future research, we recommend using CoT prompting, nucleus sampling with at least 100 samples to achieve stable estimates, and a multi-metric approach that incorporates our proposed self-correlation distance to capture both surface-level and structural alignment.

## Limitations

The assumption behind all surveys is that the answers provided by the survey subjects reflect their actual behavior in the real world. For example, a man might claim that he believes men should do a fair share of invisible household labor and will vote for parties with a compatible election program. With LLMs, there is no such guarantee. LLMs might advocate for certain values when prompted to generate text directly related to those values, but still generate texts with underlying values that do not align with the survey answer. We do not challenge this assumption in this paper. We are not aware of any existing methodology that would approach this challenge.

Comparing answers across countries based on answer distributions is a simplifying assumption. Demographic factors other than nationality, such as age or economic status, may also play a role. This study used cross-country distributions to gain insights into LLM evaluation, rather than making claims about LLMs being more representative of certain countries than others. Such claims would require a more detailed methodology.

The primary objective of this study was to compare metrics; therefore, we included only a few variables that could influence the model's behavior. It is possible that different prompt formulations and question order in the questionnaire may also yield slightly different results.

## Acknowledgments

Many thanks to Jan Hajič Jr. and Dominik Macháček for comments on the draft of this paper.

We used GitHub Copilot when writing our code. When writing the paper, we used Grammarly and Claude to improve grammar and spelling.

This research was supported by the Charles University project PRIMUS/23/SCI/023 and project CZ.02.01.01/00/23\_020/0008518 of the Czech Ministry of Education.

## References

- Muhammad Farid Adilazuarda, Chen Cecilia Liu, Iryna Gurevych, and Alham Fikri Aji. 2025. [From surveys to narratives: Rethinking cultural value adaptation in llms](#). *CoRR*, abs/2505.16408.
- Meltem Aksoy. 2024. [Whose morality do they speak? unraveling cultural bias in multilingual language models](#). *CoRR*, abs/2412.18863.
- Jacob Andreas. 2022. [Language models as agent models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5769–5779, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Mohammad Atari, Mona J. Xue, Peter S. Park, Damián E. Blasi, and Joseph Henrich. 2023. [Which humans?](#) *PsyArXiv*.
- Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. 2018. The moral machine experiment. *Nature*, 563(7729):59–64.
- Noam Benkler, Drisana Mosaphir, Scott Friedman, Andrew Smart, and Sonja Schmer-Galunder. 2023. [Assessing llms for moral value pluralism](#). *CoRR*, abs/2312.10075.
- Vamshi Krishna Bonagiri, Sreeram Vennam, Priyanshu Govil, Ponnurangam Kumaraguru, and Manas Gaur. 2024. [SaGE: Evaluating moral consistency in large language models](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14272–14284, Torino, Italia. ELRA and ICCL.
- Bram Bulté and Terryn Ayla Rigouts. 2025. [Llms and cultural values: The impact of prompt language and explicit cultural framing](#). *Computational Linguistics*, pages 1–85.
- Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. 2023. [Assessing cross-cultural alignment between ChatGPT and human societies: An empirical study](#). In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 53–67, Dubrovnik, Croatia. Association for Computational Linguistics.
- Pew Research Center. 2018. The religious typology: A new way to categorize americans by religion.
- Tanise Ceron, Neele Falk, Ana Barić, Dmitry Nikolaev, and Sebastian Padó. 2024. [Beyond prompt brittleness: Evaluating the reliability and consistency of political worldviews in LLMs](#). *Transactions of the Association for Computational Linguistics*, 12:1378–1400.
- Scott Clifford, Vijeth Iyengar, Roberto Cabeza, and Walter Sinnott-Armstrong. 2015. Moral foundations vignettes: A standardized stimulus database of scenarios based on moral foundations theory. *Behavior research methods*, 47(4):1178–1198.

- Davi Bastos Costa, Felipe Alves, and Renato Vicente. 2025. [Moral susceptibility and robustness under persona role-play in large language models](#). *CoRR*, abs/2511.08565.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, et al. 2024. [The llama 3 herd of models](#). *CoRR*, abs/2407.21783.
- Esin Durmus, Karina Nguyen, Thomas Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. 2024. [Towards measuring the representation of subjective global opinions in language models](#). In *First Conference on Language Modeling*.
- Bryan Eikema and Wilker Aziz. 2020. [Is MAP decoding all you need? the inadequacy of the mode in neural machine translation](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4506–4520, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- EVS. 2022. [Evs trend file 1981-2017](#). GESIS, Cologne. ZA7503 Data file Version 3.0.0, <https://doi.org/10.4232/1.14021>.
- Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. [From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair NLP models](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11737–11762, Toronto, Canada. Association for Computational Linguistics.
- Jesse Graham, Brian Nosek, Jonathan Haidt, Ravi Iyer, Sena P Koleva, and Peter H Ditto. 2011. Mapping the moral domain. *Journal of Personality and Social Psychology*, 101 (2):366–385.
- Daniil Gurgurov, Katharina Trinley, Ivan Vykopal, Josef van Genabith, Simon Ostermann, and Roberto Zamparelli. 2025. [Multilingual political views of large language models: Identification and steering](#). *CoRR*, abs/2507.22623.
- Christian Haerpfer, Ronald Inglehart, Alejandro Moreno, Christian Welzel, Kseniya Kizilova, Jaime Diez-Medrano, Marta Lagos, Pippa Norris, Eduard Ponarin, and Bi Puranen. 2020. World values survey wave 7 (2017-2020) cross-national data-set. (*No Title*).
- Geert Hofstede. 1984. *Culture’s consequences: International differences in work-related values*, volume 5. sage.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Léo Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. [Mistral 7b](#). *CoRR*, abs/2310.06825.
- Sharif Kazemi, Gloria Gerhardt, Jonty Katz, Caroline Ida Kuria, Estelle Pan, and Umang Prabhakar. 2024. [Cultural fidelity in large-language models: An evaluation of online language resources as a driver of model performance in value representation](#). *CoRR*, abs/2410.10489.
- Minsang Kim and Seungjun Baek. 2024. [Exploring large language models on cross-cultural values in connection with training methodology](#). *CoRR*, abs/2412.08846.
- Ashutosh Kumar and Aditya Joshi. 2022. [Striking a balance: Alleviating inconsistency in pre-trained models for symmetric classification tasks](#). In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1887–1895, Dublin, Ireland. Association for Computational Linguistics.
- Harvey Lederman and Kyle Mahowald. 2024. [Are language models more like libraries or like librarians? bibliotechnism, the novel reference problem, and the attitudes of LLMs](#). *Transactions of the Association for Computational Linguistics*, 12:1087–1103.
- Songyuan Liu, Ziyang Zhang, Runze Yan, Wei Wu, Carl Yang, and Jiaying Lu. 2024. [Measuring spiritual values and bias of large language models](#). *CoRR*, abs/2410.11647.
- Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno Miguel Guerreiro, Ricardo Rei, Duarte M. Alves, José Pombal, M. Amin Farajian, Manuel

- Faysse, Mateusz Klimaszewski, Pierre Colombo, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. 2024. [Eurollm: Multilingual language models for europe](#). *CoRR*, abs/2409.16235.
- Mark Mather, Kerri L Rivers, and Linda A Jacobsen. 2005. The american community survey. *Population Bulletin*, 60(3).
- Jared Moore, Tanvi Deshpande, and Diyi Yang. 2024. [Are large language models consistent over value-laden questions?](#) In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15185–15221, Miami, Florida, USA. Association for Computational Linguistics.
- Fabio Motoki, Valdemar Pinho Neto, and Victor Rodrigues. 2024. More human than human: measuring chatgpt political bias. *Public Choice*, 198(1):3–23.
- José Luiz Nunes, Guilherme F. C. F. Almeida, Marcelo de Araujo, and Simone D. J. Barbosa. 2024. [Are large language models moral hypocrites? a study based on moral foundations](#). *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7(1):1074–1087.
- Ricardo Dominguez Olmedo, Moritz Hardt, and Celestine Mender-Dünner. 2023. [Questioning the survey responses of large language models](#). *CoRR*, abs/2306.07951.
- Yao Qu and Jue Wang. 2024. Performance and biases of large language models in public opinion simulation. *Humanities and Social Sciences Communications*, 11(1):1–13.
- Paul Röttger, Valentin Hofmann, Valentina Pyatkin, Musashi Hinck, Hannah Kirk, Hinrich Schuetze, and Dirk Hovy. 2024. [Political compass or spinning arrow? towards more meaningful evaluations for values and opinions in large language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15295–15311, Bangkok, Thailand. Association for Computational Linguistics.
- Jens Rupperecht, Georg Ahnert, and Markus Strohmaier. 2025. [Prompt perturbations reveal human-like biases in LLM survey responses](#). *CoRR*, abs/2507.07188.
- Nathan E. Sanders, Alex Ulinich, and Bruce Schneier. 2023. Demonstrations of the potential of AI-based political issue polling. *Harvard Data Science Review*, 5(4). <https://hdsr.mitpress.mit.edu/pub/dm2hrtx0>.
- Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. [Whose opinions do language models reflect?](#) In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 29971–30004. PMLR.
- Nino Scherrer, Claudia Shi, Amir Feder, and David M. Blei. 2023. Evaluating the moral beliefs encoded in llms. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA*.
- Shalom H Schwartz. 1992. Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. In *Advances in experimental social psychology*, volume 25, pages 1–65. Elsevier.
- Hua Shen, Tiffany Knearem, Reshmi Ghosh, Yu-Ju Yang, Tanushree Mitra, and Yun Huang. 2024. [Valuecompass: A framework of fundamental values for human-ai alignment](#). *CoRR*, abs/2409.09586.
- Nicholas Sukiennik, Chen Gao, Fengli Xu, and Yong Li. 2025. [An evaluation of cultural value alignment in LLM](#). *CoRR*, abs/2504.08863.
- Yan Tao, Olga Viberg, Ryan S Baker, and René F Kizilcec. 2024. [Cultural bias and cultural alignment of large language models](#). *PNAS Nexus*, 3(9):pgae346.
- Karina Vida, Fabian Damken, and Anne Lauscher. 2024. [Decoding multilingual moral preferences: Unveiling llm’s biases through the moral machine experiment](#). *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7(1):1490–1501.
- Xinpeng Wang, Bolei Ma, Chengzhi Hu, Leon Weber-Genzel, Paul Röttger, Frauke Kreuter, Dirk Hovy, and Barbara Plank. 2024. [“my answer is C”: First-token probabilities do not match text answers in instruction-tuned language models](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7407–7416, Bangkok, Thailand. Association for Computational Linguistics.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. [Chain-of-thought prompting elicits reasoning in large language models](#). In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.
- Gian Wiher, Clara Meister, and Ryan Cotterell. 2022. [On decoding strategies for neural text generators](#). *Transactions of the Association for Computational Linguistics*, 10:997–1012.
- Dustin Wright, Arnav Arora, Nadav Borenstein, Srishri Yadav, Serge Belongie, and Isabelle Augenstein. 2024. [LLM tropes: Revealing fine-grained values and opinions in large language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 17085–17112, Miami, Florida, USA. Association for Computational Linguistics.
- Shaoyang Xu, Yongqi Leng, Linhao Yu, and Deyi Xiong. 2024. [Self-pluralising culture alignment for large language models](#). *CoRR*, abs/2410.12971.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian

Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024. [Qwen2.5 technical report](#). *CoRR*, abs/2412.15115.

## A Example of Prompt Formulation

### Question 1: Direct Prompt

*English:* For each of the following, indicate how important it is in human life. 1 is very important, 4 is not important at all. Say a number from 1 to 4 and nothing else. Family.

*German:* Geben Sie für jeden der folgenden Punkte an, wie wichtig er im menschlichen Leben ist. 1 ist sehr wichtig, 4 ist überhaupt nicht wichtig. Nennen Sie die Zahl von 1 bis 4 und sonst nichts. Familie.

*Czech:* U každé z následujících označte, jak je důležitá v lidském životě. 1 je velmi důležitá, 4 není vůbec důležitá. Napiš číslo od 1 do 4 a nic jiného. Rodina.

### Question 1: CoT Prompt

*English:* For each of the following, indicate how important it is in human life. 1 is very important, 4 is not important at all. First, provide a step-by-step justification, and finally, say the number from 1 to 4 and nothing else. Family.

*German:* Geben Sie für jeden der folgenden Punkte an, wie wichtig er im menschlichen Leben ist. 1 ist sehr wichtig, 4 ist überhaupt nicht wichtig. Geben Sie zunächst eine schrittweise Begründung an und nennen Sie abschließend die Zahl von 1 bis 4 und sonst nichts. Familie.

*Czech:* U každé z následujících označte, jak je důležitá v lidském životě. 1 je velmi důležitá, 4 není vůbec důležitá. Nejprve napiš zdůvodnění krok za krokem a nakonec napiš číslo od 1 do 4 a nic jiného. Rodina.

### Question 41: Direct Prompt

*English:* For the following statement, tell me how much agreeable it is. 1 is strongly agree 3 is neutral, 5 is strongly disagree. Say a number from 1 to 5 and nothing else. Work should always come first, even if it means less spare time.

*German:* Sagen Sie mir, wie sehr kann man der folgenden Aussage zustimmen. 1 bedeutet völlige Zustimmung, 3 bedeutet neutral, 5 bedeutet überhaupt keine Zustimmung. Nennen Sie eine Zahl von 1 und 5 und sonst nichts. Die Arbeit sollte immer an erster Stelle stehen, auch wenn dies weniger Freizeit bedeutet.

*Czech:* U následující tvrzení uveďte, jak moc s ním lze souhlasit. 1 úplný souhlas, 3 je neutrální, 5 naprostý nesouhlas. Napiš číslo od 1 do 5 a nic jiného. Práce

by měla být vždy na prvním místě, i když to znamená méně volného času.

### Question 41: CoT Prompt

*English:* For the following statement, tell me how much agreeable it is. 1 is strongly agree 3 is neutral, 5 is strongly disagree. First provide a step-by-step justification and finally say a number from 1 to 5 and nothing else. Work should always come first, even if it means less spare time.

*German:* Sagen Sie mir, wie sehr kann man der folgenden Aussage zustimmen. 1 bedeutet völlige Zustimmung, 3 bedeutet neutral, 5 bedeutet überhaupt keine Zustimmung. Geben Sie zunächst eine schrittweise Begründung an und nennen Sie abschließend eine Zahl von 1 und 5 und sonst nichts. Die Arbeit sollte immer an erster Stelle stehen, auch wenn dies weniger Freizeit bedeutet.

*Czech:* U následující tvrzení uveďte, jak moc s ním lze souhlasit. 1 úplný souhlas, 3 je neutrální, 5 naprostý nesouhlas. Nejprve napiš zdůvodnění krok za krokem a nakonec napiš číslo od 1 do 5 a nic jiného. Práce by měla být vždy na prvním místě, i když to znamená méně volného času.

## B Detailed Multilingual Results

Here, we present the MSD, KL Divergence, Correlation norm, and Self-correlation distance for all combinations of languages (English: en, German: de, Czech: cs) and all countries (United States: USA, United Kingdom: GBR, Czechia: CZE, Germany: DEU, Iran: IRN, China: CHN) in Table 7. Table 2 is a subset of this table. Table 4 with point-biserial correlation of country-language matching and evaluation metrics and Table 3 are computed from numbers in Table 7. Table 5 shows the metric correlation separately from LLaMA 3 and Mistral 2.

Table 7 shows country comparison using the data from WVS with the metrics that we use in the paper. In the paper, we occasionally compare the model alignment to differences between countries. For this, we use data from this table.

MSD KLD CorrD				MSD KLD CorrD				MSD KLD CorrD				MSD KLD CorrD			
MSD	—	.276	-.368	MSD	—	.832	-.657	MSD	—	.691	.279	MSD	—	.617	.120
KLD	.389	—	-.064	KLD	.925	—	-.500	KLD	.724	—	.251	KLD	.656	—	.495
CorrD	-.342	-.005	—	CorrD	-.666	-.526	—	CorrD	.328	.271	—	CorrD	.009	.396	—
(a) LLaMA 3				(b) Mistral 2				(c) EuroLLM				(d) Qwen 2.5			

Table 5: Breakdown of correlation of the metrics (Pearson above the diagonal, Spearman under the diagonal) over all languages and countries for (a) LLaMA 3, (b) Mistral 2, (c) EuroLLM, and (d) Qwen 2.5

Model	Prompt type	De-code	Lang	Mean Sq. Difference						KL Divergence						Corr. norm	Self-correlation distance						
				USA	GBR	CZE	DEU	IRN	CHN	USA	GBR	CZE	DEU	IRN	CHN		USA	GBR	CZE	DEU	IRN	CHN	
LLaMA 3	Score only	Gr.	en	.098	.098	.110	.096	.115	.102	1.71	1.66	1.74	1.67	1.81	1.70								
			de	.084	.076	.103	.079	.132	.112	1.65	1.56	1.70	1.62	1.93	1.85								
			cs	.086	.071	.110	.077	.145	.119	1.56	1.41	1.64	1.41	1.83	1.69								
	CoT	Spl.	en	.073	.073	.083	.073	.090	.083	2.99	2.90	3.26	2.81	3.43	3.23	0.90	1.26	1.20	1.17	0.95	0.96	0.93	
			de	.084	.073	.099	.079	.127	.114	3.58	3.33	3.76	3.30	4.14	3.79	0.67	1.31	1.25	1.24	1.01	0.99	1.01	
			cs	.067	.047	.080	.057	.114	.105	2.90	2.49	3.05	2.46	3.52	3.04	0.78	1.26	1.20	1.20	0.96	0.94	0.95	
Mistral2	Score only	Gr.	en	.088	.083	.112	.090	.152	.133	1.68	1.57	1.73	1.63	2.01	1.89								
			de	.085	.072	.104	.080	.134	.117	1.61	1.47	1.69	1.55	1.96	1.82								
			cs	.071	.060	.089	.068	.125	.104	1.57	1.46	1.63	1.58	1.87	1.71								
	CoT	Spl.	en	.059	.046	.075	.054	.118	.111	1.47	1.30	1.70	1.29	1.94	1.93	1.70	1.29	1.26	1.22	1.14	1.19	1.03	
			de	.054	.042	.063	.050	.105	.094	1.05	0.87	1.12	0.87	1.35	1.34	1.50	1.17	1.14	1.10	0.97	1.02	0.89	
			cs	.049	.035	.059	.042	.095	.087	0.98	0.81	1.11	0.84	1.37	1.27	1.60	1.24	1.19	1.14	1.04	1.11	0.93	
EuroLLM	Score only	Gr.	en	.094	.095	.114	.110	.121	.114	1.80	1.75	1.76	1.90	2.01	1.88								
			de	.096	.107	.105	.110	.100	.105	1.90	1.97	1.82	2.08	2.08	1.94								
			cs	.114	.129	.121	.131	.111	.118	1.91	1.94	1.83	2.06	2.07	1.98								
	CoT	Spl.	en	.041	.040	.051	.051	.072	.056	0.77	0.70	0.82	0.82	1.41	0.97	1.77	1.17	1.18	1.10	1.06	1.10	0.95	
			de	.035	.039	.040	.051	.072	.068	0.56	0.57	0.51	0.73	0.94	0.79	2.26	1.37	1.35	1.32	1.32	1.37	1.19	
			cs	.034	.036	.037	.046	.060	.058	0.41	0.41	0.40	0.49	0.62	0.55	2.64	1.53	1.44	1.51	1.58	1.67	1.46	
Qwen	Score only	Gr.	en	.165	.167	.193	.186	.162	.156	1.91	1.80	1.97	1.81	1.72	1.75								
			de	.284	.290	.261	.300	.254	.274	2.40	2.35	2.33	2.40	2.08	2.38								
			cs	.223	.244	.243	.248	.176	.209	2.19	2.21	2.26	2.22	1.84	2.14								
	CoT	Spl.	en	.059	.069	.068	.081	.045	.055	0.72	0.65	0.78	0.68	0.55	0.75	2.55	1.56	1.54	1.54	1.57	1.62	1.44	
			de	.094	.111	.093	.121	.065	.080	1.00	0.95	1.04	0.93	0.69	0.92	2.76	1.66	1.59	1.70	1.69	1.77	1.57	
			cs	.129	.149	.129	.157	.082	.103	1.52	1.50	1.59	1.53	1.17	1.55	2.95	1.71	1.66	1.78	1.79	1.87	1.69	
Qwen	Score only	Gr.	en	.130	.130	.157	.142	.155	.170	1.76	1.69	1.78	1.76	1.70	1.82								
			de	.369	.402	.322	.393	.207	.279	0.91	0.97	0.80	1.08	0.79	0.94								
			cs	.358	.421	.380	.442	.286	.296	1.15	1.25	1.16	1.28	0.98	1.12								
	CoT	Spl.	en	.125	.143	.126	.153	.091	.099	0.97	0.93	1.00	0.94	0.68	0.87	2.97	1.74	1.64	1.78	1.79	1.88	1.68	
			de	.149	.171	.151	.179	.106	.121	1.29	1.27	1.38	1.21	0.89	1.12	2.83	1.71	1.59	1.74	1.73	1.85	1.64	
			cs	.120	.140	.119	.149	.086	.097	1.12	1.09	1.14	1.07	0.76	0.98	2.84	1.67	1.52	1.67	1.71	1.82	1.57	
Qwen	Score only	Gr.	en	.082	.068	.105	.079	.136	.107	1.66	1.60	1.82	1.68	2.02	1.88								
			de	.070	.102	.069	.088	.092	.100	0.82	0.88	0.73	0.92	0.92	0.98								
			cs	.081	.053	.099	.057	.178	.147	0.66	0.65	0.60	0.78	0.83	0.87								
	CoT	Spl.	en	.074	.075	.080	.081	.076	.065	1.27	1.17	1.48	1.15	1.35	1.33	3.33	2.13	2.07	2.12	2.15	2.16	2.09	
			de	.054	.043	.062	.046	.093	.073	1.13	0.95	1.24	0.95	1.40	1.31	2.12	1.61	1.56	1.51	1.45	1.50	1.36	
			cs	.041	.033	.053	.040	.083	.060	0.81	0.69	0.91	0.65	1.00	0.91	1.93	1.39	1.33	1.24	1.24	1.27	1.16	
Qwen	Score only	Gr.	en	.199	.170	.206	.193	.268	.253	1.60	1.50	1.56	1.54	1.75	1.74								
			de	.141	.133	.149	.154	.179	.177	1.70	1.66	1.73	1.78	1.84	1.87								
			cs	.116	.113	.109	.101	.142	.153	1.22	1.21	1.12	1.25	1.20	1.35								
	CoT	Spl.	en	.062	.048	.078	.058	.105	.078	1.13	0.97	1.37	1.01	1.53	1.39	1.52	1.25	1.13	1.15	1.03	1.09	0.94	
			de	.055	.050	.073	.061	.103	.093	0.69	0.59	0.75	0.65	0.92	0.84	1.84	1.24	1.16	1.12	1.12	1.21	0.98	
			cs	.048	.036	.062	.045	.088	.070	0.73	0.61	0.83	0.63	1.01	0.90	1.72	1.17	1.12	1.10	1.03	1.12	0.92	

Table 6: A comparison of model outputs with different prompting strategies (Score-only, CoT: Chain of Thought), decoding method (Gr.: greedy, Spl.: sampling) and different languages

	USA	GBR	CZE	DEU	IRN	CHN		USA	GBR	CZE	DEU	IRN	CHN		Norm.	USA	GBR	CZE	DEU	IRN	CHN
USA	—	.009	.017	.016	.046	.043	USA	—	0.07	0.11	0.16	0.34	0.32	USA	1.66	—	0.79	0.92	0.78	0.95	0.92
GBR	.009	—	.022	.011	.066	.052	GBR	0.15	—	0.13	0.09	0.39	0.35	GBR	1.54	0.79	—	0.90	0.72	0.92	0.78
CZE	.017	.022	—	.024	.044	.042	CZE	0.17	0.15	—	0.22	0.34	0.29	CZE	1.55	0.92	0.90	—	0.83	0.93	0.79
DEU	.016	.011	.024	—	.069	.049	DEU	0.21	0.08	0.17	—	0.36	0.32	DEU	1.11	0.78	0.72	0.83	—	0.66	0.64
IRN	.046	.066	.044	.069	—	.030	IRN	0.41	0.42	0.33	0.44	—	0.27	IRN	1.06	0.95	0.92	0.93	0.66	—	0.69
CHN	.043	.052	.042	.049	.030	—	CHN	0.32	0.34	0.27	0.33	0.24	—	CHN	1.13	0.92	0.78	0.79	0.64	0.69	—

Table 7: Comparison of difference between countries in the World Value Survey when measured using mean squared difference, KL Divergence, the norm of the self-correlation tables for each country and distances of the self-correlation tables across countries.

# An improved Code-Switching Detection System for some Indic Languages

**Karan Bhanushali**

University of Stuttgart  
Stuttgart, Germany

karanbhanushali314@gmail.com

**Fritz Hohl**

Sony Europe Limited  
Stuttgart, Germany

Fritz.Hohl@sony.com

## Abstract

Code-switching is a common feature of multilingual communication, and reliably identifying where the language switches is essential for downstream tasks such as generating code-switched machine translations. This paper introduces CSDI, a Code-Switching Detection (CSD) system for Indic text, which jointly learns CSD, Named Entity Recognition, and Part-of-Speech tagging through a shared encoder. Leveraging multitask learning, CSDI captures linguistic cues that signal switching boundaries and achieves a new state-of-the-art macro-F1 score with near-zero  $\Delta$ CMI across six Indic languages. The model also demonstrates strong cross-lingual transfer, effectively leveraging high-resource languages to improve low-resource performance. Despite challenges such as intra-word code-mixing and limited token-level context, CSDI establishes a new baseline for scalable, low-resource NLP research in code-mixed environments.

## 1 Introduction

Code-switching (CS) is the alteration between two or more languages within a single conversation or utterance. This is a natural characteristic of multilingual societies such as India, where speakers frequently mix English with native languages in both spoken and written communications. This phenomenon is especially common in social media, everyday conversations, and in film dialogues.

When automatically translating Indic movie dialogues, this phenomenon cannot be ignored. Analysis of our internal data shows that roughly 15% of the words in Indic movie subtitles are English elements. These mixed-language segments behave differently from pure monolingual ones during translation, making CSD an essential step in translation.

Existing systems for some Indic languages (Bengali, Hindi, Kannada, Malayalam, Marathi, and

Telugu) offered not enough accuracy for our requirements, which motivated the development of an improved system that achieves better results on standard test sets and on an internal movie dialogue dataset.

In this work, we only consider code-switching where entire words are considered monolingual; intra-word code-switching, while present, is rarer and requires a different approach from the one proposed here. Also, we assume there is maximally one code-switched language (concretely English) in the source language text. Ignoring more than one code-switched language and languages other than English does not completely reflect the reality in the datasets, but it comes relatively close statistically.

To address the limitations of existing systems, we introduce a multitask learning framework that jointly performs CSD, NER, and POS tagging through a shared encoder. This enables the model to capture richer linguistic cues around switching boundaries, leading to substantial improvements over prior work. The resulting system achieves new state-of-the-art performance across six Indic languages, exhibits near-zero  $\Delta$ CMI, and transfers effectively from high-resource to low-resource settings.

## 2 Background

Code-switching has evolved from being viewed as a linguistic imperfection (Bolonyai, 2009) to a natural feature of multilingual discourse (Grosjean, 2010). Early studies classified CS into intersentential, intra-sentential, intra-word, and tag-switching (Poplack, 1981), governed by grammatical and social factors (Muysken, 1995). In India, with 121 major languages and a high bilingualism, CS is widely observed in social media and films, often used for emphasis or identity expression (Dey and Fung, 2014; Pratapa and Choudhury, 2017).

To explain how CS is structured, the Matrix Language Frame (MLF) model (Jake and Myers-Scotton, 1997) distinguishes the dominant (matrix) language from embedded ones. In Indic CS, Hindi–English switches frequently occur at lexical boundaries, typically English nouns and verbs (Dey and Fung, 2014). These linguistic insights highlight where switching boundaries are likely to appear, making them directly relevant to token-level detection.

Beyond understanding the structure of CS, it is also important to quantify its extent. The Code-Mixing Index (CMI) (Das and Gambäck, 2014) is a widely used metric that measures the proportion of tokens belonging to the embedded language within an utterance. Formally, for a sentence with  $N$  tokens and  $n_E$  tokens from the embedded language, CMI is defined as:

$$\text{CMI} = \frac{n_E}{N} \times 100$$

A CMI value of zero corresponds to purely monolingual text, while higher values indicate increased code-mixing. In this work, we use CMI to characterize dataset mixing intensity and to analyze model behavior under varying degrees of code-switching.

With the growth of NLP for multilingual communities, computational approaches progressed from early rule-based systems to modern Deep Learning methods. CS datasets now exist for several tasks, including POS tagging (Jamatia et al., 2018), Sentiment Analysis (Patwa et al., 2020), and Machine Translation (MT) (Dhar et al., 2018). Token-level Language Identification (LID), a crucial preprocessing step for many CS tasks, has benefited from multilingual encoders such as IndicBERT (Kakwani et al., 2020) and MuRIL (Khanuja et al., 2021), showing that fine-tuning on CS corpora improves performance (Santy et al., 2021). Multitask Learning (MTL) with POS has also been effective in uncovering structural clues (Winata et al., 2018; Pitale and Malapati, 2023).

Recent CSD approaches such as AnE (Sterner, 2024) aggregate CS corpora for multilingual token-level LID and identifies English even in unseen language pairs (e.g., Indonesian–English), achieving 2.3–4.6% F1 gains over pair-specific SoTA using XLM-RoBERTa. While these multilingual approaches demonstrate strong generalization, Indic-specific challenges such as data scarcity and romanization variations remain underexplored.

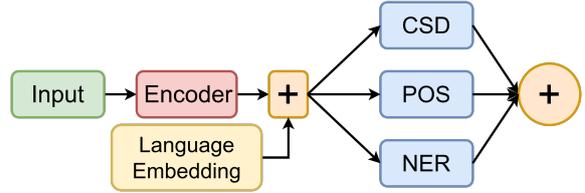


Figure 1: Architecture of CSDI. A shared encoder processes romanized input tokens. A trainable matrix-language embedding is added to the encoder representations to condition them on the base Indic language. The conditioned representations are passed to three task-specific heads (CSD, POS, NER). During training, losses from the three tasks are combined additively and jointly backpropagated to the shared encoder.

Although CS is a relatively frequent phenomenon, obtaining labelled CS data for supervised training is particularly challenging, as naturally mixed utterances are sparsely distributed across large monolingual corpora. Additionally, inconsistent orthography and informal romanization introduce noise and ambiguity, especially in user-generated content. These issues are not fully addressed by existing multilingual LID systems, which often struggle with Indic-specific variation. This highlights the need for more robust and adaptable approaches to token-level detection in romanized Indic–English settings.

### 3 CSDI – Code-Switching Detection for Indic Languages

In this section, we present CSDI (Code-Switching Detection for Indic Languages), our proposed multi-task learning (MTL) framework designed to detect code-switching in Indic text.

CSDI jointly learns CSD, NER, and POS tagging through a shared encoder. This design allows the model to leverage syntactic and semantic cues from auxiliary tasks. Additionally, the framework incorporates non-code-switched (monolingual) data to improve robustness and generalization.

The following subsections describe the overall architecture and training strategies applied to CSDI.

#### 3.1 Architecture

Figure 1 summarizes the flow of information through the model, from romanized input to the multitask outputs. CSDI is built around a shared multilingual transformer encoder that learns a uni-

fied representation for CSD, NER, and POS tagging. This shared design enables linguistic information learned by auxiliary tasks to support CSD, which is typically the lowest-resource component in Indic settings.

### Shared Encoder

We experimented with IndicBERT, XLM-RoBERTa, mBERT, and MuRIL, and found MuRIL to offer the best performance on romanized Indic text, especially for languages with sparse training data. These results are discussed in more detail in Section 6.

### Matrix-Language Conditioning

Although all input is romanized, the underlying Indic languages exhibit different syntactic and lexical tendencies, which influence switching behaviour. To encode this information, CSDI includes a trainable matrix-language embedding. The '+' operation in the architecture denotes element-wise addition. For each input sentence, the embedding corresponding to its base Indic language is added to the encoder’s output. This conditioning shifts the internal representation to reflect language-specific tendencies—for example, the higher likelihood of English verb insertions in Hindi compared to Kannda.

### Task-Specific Heads

Three lightweight linear classifiers operate on the conditioned encoder output (Figure 1):

- **CSD head:** predicts English vs. Indic token identity.
- **NER head:** predicts BIO entity types.
- **POS head:** predicts universal POS tags.

Each head performs word-level classification using the first token in the token sequence corresponding to each word. The auxiliary signals from POS and NER help the encoder distinguish structural and semantic contexts in which code-switching is more likely.

The multi-task setup encourages the encoder to learn syntactic boundaries (from POS) and semantic constraints (from NER) that correlate with switching events, reducing confusion in ambiguous romanized contexts. For example, patterns such as English nouns within Hindi noun phrases or English verbs following Hindi auxiliaries are reinforced across tasks.

## 3.2 Training Strategy

CSDI is optimized with a linearly weighted multi-task loss:

$$\mathcal{L} = w_{csd} \mathcal{L}_{csd} + w_{ner} \mathcal{L}_{ner} + w_{pos} \mathcal{L}_{pos}.$$

Here,  $w_{csd} = w_{ner} = w_{pos} = 1$  in the main experiments (ablation studies vary these weights).

All three task heads are trained with token-level cross-entropy loss. The '+' operation at the output side of the architecture represents additive loss aggregation rather than a combination of task predictions. Each task head produces an independent token-level loss, which are summed (optionally weighted) up to form a single scalar objective. The combined loss is then backpropogated through the task heads and the shared encoder.

To prevent high-resource NER/POS tasks from dominating the low-resource CSD signal, we employ a *two-phase strategy*:

1. **Warm-up** (first epoch): the model is trained on a subset ( $\approx 20\%$ ) of the CSD data with CSD loss only.
2. **Joint optimization** (remaining epochs): all three losses are combined.

This mitigates catastrophic forgetting while preserving the focus on CSD.

We conduct a set of ablation studies to understand the contribution of each component of CSDI. These include: (i) task ablations (CSD-only, CSD+POS, CSD+NER, full MTL), (ii) encoder selection (IndicBERT, mBERT, XLM-RoBERTa, MuRIL), (iii) data ablations (removing dictionary data, removing monolingual data, disabling romanization augmentation), and (iv) language-transfer settings (single-language models, family-grouped models, and all-language joint training). These ablations isolate the effect of auxiliary tasks, encoder choice, and cross-lingual transfer. POS and NER objectives serve only as auxiliary supervision during training; all reported results correspond to the CSD task.

The family-grouped and all-language models test the hypothesis that typologically related languages share switching patterns, enabling low-resource languages (e.g., Marathi) to benefit from high-resource members of the same family (e.g., Hindi). Future extensions could explore dynamic loss weighting for finer task balancing.

Dataset	English	Bengali	Hindi	Kannada	Malayalam	Marathi	Telugu
Dictionary Dataset	english-words	Aksharantar					
Multitask Dataset	Code-Switching	openslr	L3Cube-HingLD, LinCE, CoSTA	Kanglishicon 20233	WoLLal	Marathi-Dataset	Word LID CoSTA
	NER	Bangla-Complex-NER	LinCE			Marathi-Dataset	
	POS	KCIS Bangla NLP	LinCE	KCIS	KCIS	KCIS	

Figure 2: Overview of the datasets used for training CSDI. The table lists all dictionary, CSD, NER, and POS resources included for each Indic language. Complete references are listed in the Appendix.

## 4 Used Datasets

In this section, we describe the datasets used for training and evaluation. We make use of both publicly available and internal resources. The training data for CSDI contains a diverse collection of Indic–English code-switched data. All texts are transliterated into Latin script to eliminate orthographic variation and enable cross-language generalization. The training datasets include a dictionary-based dataset and a multitask dataset (for POS, NER, and CSD). Figure 2 summarizes the training data composition. A more detailed list of the used dataset can be found in Table 8.

### 4.1 Dictionary Train Dataset

This dataset consists of romanized word lists for multiple languages. For English, we use the *english-words* database<sup>1</sup>, while for Indic languages we extract the top 10,000 most frequent words from the Aksharantar corpus (Madhani et al., 2023) and convert them into Latin script using indic-transliteration library<sup>2</sup>. These lists provide broad lexical coverage and help the model handle unseen romanized forms.

### 4.2 Multitask Train Dataset

The Multitask Dataset includes CSD, NER, and POS components. The NER and POS datasets are classical NLP datasets, while the CSD component consists of token-labeled sentences. Figure 3 provides an overview.

Most datasets are manually annotated, but for resources such as CoSTA, OpenSLR, and Kanglish, language labels are assigned automatically using a script heuristic: tokens in Latin script are tagged

<sup>1</sup><https://github.com/dwyl/english-words>

<sup>2</sup><https://pypi.org/project/ai4bharat-transliteration/>

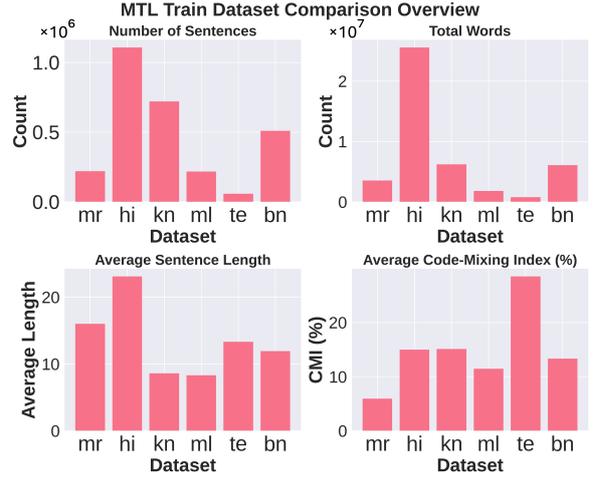


Figure 3: Overview of the MTL training data across six Indic languages. The figure summarizes the number of sentences, total words, average number of words in a sentence, and average CMI for each language.

as English, and tokens in Indic scripts as the corresponding Indic language. This silver labeling introduces noise — for example, Indic Named Entities written in Latin script may be incorrectly tagged as English, while English-origin borrowed words (e.g., बस (bus) or ट्रेन (train)) written in Indic scripts may be mislabeled as Indic — but still provides valuable weak supervision.

As seen in Fig. 3, the Code-Mixing Index (CMI) varies across languages, reflecting domain differences. Marathi data, drawn mostly from formal text, contains fewer English insertions, while Telugu data, sourced largely from social media, shows more conversational code-switching.

### 4.3 Multitask Test Dataset

Although CSDI is trained in a multitask setting using CSD, POS, and NER supervision, evaluation is performed exclusively on the CSD task (token-level language identification). The General and Internal test sets therefore contain only token-level language identification labels. Accordingly, both the General and Internal test sets contain only token-level language identification labels too. The CSD part of the Multitask Test Dataset has the same composition as the Multitask Train Dataset, a mixture of gold- and silver-labeled data. All test data are held out and not used during training.

For each language pair, we reserve 1,000 sentences for the test split, except for Hindi, for which we use 6,000 held-out sentences. This results in a slightly uneven distribution across languages, but it reflects the availability of annotated

resources and ensures sufficient evaluation coverage for Hindi. Datasets that provide POS or NER annotations are used solely as auxiliary learning signals during training and are not evaluated at test time.

#### 4.4 Pre-processing

Since the datasets originate from various domains and scripts, several normalization steps are applied prior to generating augmented romanized variants:

- removal of URLs, mentions, hashtags, emojis, numbers, and punctuation;
- explicit token-level language tagging: English tokens remain in Latin script, while Indic tokens are transliterated into their native script;
- treatment of intra-word code-switching (common in Malayalam) by assigning the embedded English label to the whole token;
- cleaning of the Aksharantar dictionary to remove English-origin words listed under Indic vocabularies.

The Aksharantar Indic dictionaries occasionally contain transliterated English-origin words (e.g., quarantine, audience) that are either direct borrowings or near-phonetic renderings in Indic script. Retaining such entries as Indic tokens can introduce systematic noise in code-switching detection, as these forms are likely to appear as English insertions in naturally mixed text.

To mitigate this, we identify and remove English-origin entries from the Indic dictionaries using a semi-automatic, overlap-based filtering procedure. Details of the filtering criteria and overlap analysis are provided in the Appendix.

These steps standardize the heterogeneous corpora and ensure consistent input before augmentation and training.

#### 4.5 Augmented Dataset

The above datasets do not fully capture the variation introduced by informal romanization. To address this, we introduce an *Augmented Dataset* modeling cross-script and dialectal spelling variation. Romanization across speakers is highly inconsistent: for example, the English word *school* may appear as *school*, *skool*, or *iskul*, while Hindi किताब (book) may be written as *kitab* or *keetab*. As

no resource exhaustively covers these forms, we generate variants automatically.

All training sentences are first transliterated into their native scripts and then back-transliterated into Latin script using three transliteration libraries:<sup>345</sup>. Because each library follows different conventions, this round-trip produces multiple plausible romanizations. To avoid excessive duplication, roughly 40% of tokens in each sentence are replaced with their variants, yielding up to five augmented versions per sentence. This improves robustness to spelling noise and non-standard romanization.

#### 4.6 Internal Movie Dialogue Test Dataset

To evaluate performance on naturally occurring conversational speech, we use an *Internal Movie Dialogue Test Dataset* consisting of transcribed episodes from SonyYAY!<sup>6</sup> and SonyLIV<sup>7</sup>. Each episode is transcribed in the script of the audio track (e.g., Devanagari for Hindi), and the code-switched segments follow the same script.

SonyYAY! provides one animated episode each for Hindi, Marathi, and Telugu. SonyLIV provides one live-action episode each for Hindi, Malayalam, Marathi, and Telugu. Unlike the Multitask Test set, this dataset is smaller and reflects real dialogue rather than curated text.

Malayalam exhibits a lower CMI, partly due to annotator treatment of intra-word mixing: forms such as *drive-cheythu* or *call-um* were labeled as Malayalam, while during training such hybrids were treated as English based on their lexical roots. This mismatch contributes to lower Malayalam performance on this internal dataset.

## 5 Experiments

We evaluate CSDI on six Indic–English language pairs: Bn, Hi, Kn, Ml, Mr, and Te. All models are trained on romanized text with a weighted multi-task loss and fine-tuned for 3 epochs using a learning rate of  $1 \times 10^{-5}$  and a batch size of 8. Performance is reported using macro F1, English F1 (the primary metric), and  $\Delta$ CMI.

<sup>3</sup><https://indic-trans.readthedocs.io/>

<sup>4</sup><https://pypi.org/project/ai4bharat-transliteration/>

<sup>5</sup>[https://github.com/indic-transliteration/indic\\_transliteration\\_py](https://github.com/indic-transliteration/indic_transliteration_py)

<sup>6</sup>[https://en.wikipedia.org/wiki/Sony\\_YAY!](https://en.wikipedia.org/wiki/Sony_YAY!)

<sup>7</sup><https://en.wikipedia.org/wiki/SonyLIV>

## 5.1 Evaluation Setup

We use two held-out test sets: (i) a manually curated general test set, and (ii) a noisy internal movie-dialogue test set containing naturally occurring code-switching. These allow us to evaluate both clean, annotation-controlled performance and real-world robustness. We compare CSDI against AnE (Sterner, 2024), a recent multilingual CS detector based on XLM-RoBERTa.

Beyond token-level accuracy, it is important that a code-switching detector preserves the overall degree of mixing present in the input. We therefore report  $\Delta\text{CMI}$ , defined as the difference between the predicted and gold CMI values (Section 2), to assess how faithfully a model reproduces the original mixing density.

A low absolute  $\Delta\text{CMI}$  indicates that the model does not systematically over-predict or suppress English tokens, even when token-level errors occur. This property is particularly important for downstream tasks such as code-switched machine translation, where preserving the proportion of embedded language is often critical. In this sense,  $\Delta\text{CMI}$  complements F1-based metrics by measuring distributional faithfulness rather than local accuracy alone.

## 5.2 Ablation Settings

To isolate the contribution of different components, we evaluate several model variants differing in task configuration, encoder choice, data composition, and cross-lingual training setup. The specific ablations and their results are presented in Section 6.

# 6 Results and Discussion

## 6.1 Main Results

The results in Table 1 correspond to the best-performing CSDI configuration, which uses the MuRIL encoder and non-uniform task weights (CSD: 2.0, POS: 1.5, NER: 1.0). This setup emerged as optimal in our ablation studies and is therefore used for all reported main results. Under this setup, CSDI consistently outperforms the AnE baseline (Sterner, 2024) across all six language pairs, with substantial improvements in English F1, macro F1, and  $\Delta\text{CMI}$ . In contrast to AnE, however, who can work for any base language, our system is confined to those languages for which it was trained.

$\Delta\text{CMI}$  captures how faithfully a model reproduces the global degree of code-mixing in an utter-

ance. CSDI obtains near-zero  $\Delta\text{CMI}$  on the General test set, indicating that it not only identifies English tokens accurately but also models the overall mixing behaviour, an essential property for downstream MT systems that rely on retaining English insertions.

Since CSDI formulates detection as a binary task (English vs. L1), English F1 is the primary metric: English tokens are sparse yet carry most of the switching signal. CSDI shows the largest improvements for lower-resource languages such as Marathi and Bengali, demonstrating strong cross-lingual transfer from high-resource counterparts.

To assess robustness under noisier, real-world conditions, we additionally evaluate CSDI on the Internal Movie Dialogue test set. As expected, performance on this dataset is lower due to inconsistent automatic romanization and conversational style variation. Nevertheless, CSDI maintains strong performance across languages, showing reliable generalization despite domain mismatch. Full internal dataset results are provided in Table 13.

## 6.2 Encoder Selection

We compare four pretrained encoders—IndicBERT, mBERT, XLM-RoBERTa, and MuRIL—to determine which best models romanized Indic–English text. On the Internal test set (Table 2), MuRIL achieves the strongest overall performance, with higher English F1 and more stable  $\Delta\text{CMI}$  than the other encoders; similar trends are observed on the General test set. This advantage likely stems from MuRIL’s exposure to romanized Indic data during pretraining, which improves its handling of spelling variability and noisy transliterations. In contrast, mBERT performs weakest, and XLM-R shows reduced robustness to romanization noise. Based on these results, we use MuRIL as the encoder for all subsequent experiments.

## 6.3 Task Contribution

Table 3: Task contribution comparison on internal test data.

Task Combination	Macro F1	Prec. EN	Prec. L1	$\Delta\text{CMI}$
CSD + POS	0.7731	0.5063	0.9711	-9.83
CSD + NER + POS	<b>0.8758</b>	<b>0.7300</b>	<b>0.9760</b>	<b>-2.65</b>
CSD + NER	0.7027	0.3977	0.9663	-16.09

Table 3 & 4 shows that auxiliary tasks substantially improve CSD performance, with POS providing

Table 1: Final results for all language pairs on the General test data. L1 denotes the base Indic language in each pair.

Lang. Pair	System	English			L1			Overall		
		P	R	F1	P	R	F1	Acc	MF1	$\Delta$ CMI
Hi-En	AnE	0.93	0.70	0.80	0.89	0.98	0.93	0.90	0.86	+7.26
	CSDI	0.98	0.96	<b>0.97</b>	0.98	0.99	0.99	0.98	<b>0.98</b>	<b>+0.48</b>
Mr-En	AnE	0.83	0.95	0.88	0.98	0.95	0.97	0.95	0.93	+2.65
	CSDI	0.93	0.94	<b>0.94</b>	0.98	0.98	0.98	0.98	<b>0.96</b>	<b>-0.70</b>
Bn-En	AnE	0.79	0.89	0.84	0.94	0.88	0.91	0.88	0.87	-4.33
	CSDI	0.95	0.95	<b>0.95</b>	0.97	0.97	0.97	0.97	<b>0.96</b>	<b>-0.29</b>
Kn-En	AnE	0.88	0.96	0.92	0.93	0.81	0.86	0.90	0.89	+5.12
	CSDI	0.92	0.95	<b>0.93</b>	0.93	0.87	0.90	0.92	<b>0.91</b>	<b>+2.40</b>
Ml-En	AnE	0.99	0.81	0.89	0.66	0.98	0.79	0.86	0.84	-13.30
	CSDI	0.99	0.99	<b>0.99</b>	0.99	0.99	0.99	0.99	<b>0.99</b>	<b>+0.08</b>
Te-En	AnE	0.93	0.93	0.93	0.92	0.92	0.92	0.93	0.93	-0.07
	CSDI	0.98	0.99	<b>0.99</b>	0.99	0.98	0.98	0.98	<b>0.98</b>	+0.32

Table 2: Encoder comparison on the internal dataset (English F1 only).

Encoder	HI	MR	ML	TE
IndicBERT	0.728	0.551	0.437	0.757
mBERT	0.691	0.467	0.443	0.763
MuRIL	<b>0.796</b>	<b>0.563</b>	0.245	<b>0.764</b>
XML-R	0.715	0.521	<b>0.480</b>	0.762

Table 4: Task contribution comparison on General test performance for Hi-En.

Task Combination	Macro F1	Prec. EN	Prec. L1	$\Delta$ CMI
CSD + NER	0.9827	0.9836	0.9867	0.48
CSD + NER + POS	0.9747	0.9655	0.9847	<b>0.09</b>
CSD + POS	<b>0.9843</b>	0.9819	0.9893	0.24

the strongest gains. This likely stems from POS cues marking syntactic boundaries where switches frequently occur (e.g., Hindi noun before an English verb: “मैं गाना *play* करता हूँ”). In contrast, adding large amounts of non-code-switched NER data can reduce accuracy by biasing the model toward monolingual behavior. The full multitask configuration (CSD + NER + POS) achieves the best overall performance, demonstrating that combining syntactic and semantic signals is beneficial for modeling code-switching.

## 6.4 Data Contribution

Table 5 summarizes how different data sources affect model performance. Removing the dictionary hurts generalization to unseen romanized forms, and dictionary-only training leads to skewed predictions. Monolingual data provides small but

consistent gains by improving contextual modeling. Augmented romanization variants, especially for English tokens in non-standard forms such as “*iskul*” for “*school*”, improve recall on noisy internal data.

Table 5: Data contribution comparison on General test data.

Dataset Config	Macro F1	Prec. EN	Prec. L1	$\Delta$ CMI
All Data	<b>0.9723</b>	<b>0.9647</b>	<b>0.9823</b>	<b>+0.25</b>
Dictionary Data	0.2572	0.2897	0.7061	+25.33
CSD + Monolingual	0.2868	0.2851	0.6590	+20.50
Only CSD Data	0.4025	0.2938	0.7218	+3.42

Table 6: Impact of different training data components on Internal test performance for Hi-En.

Dataset Config	Macro F1	Prec. EN	Prec. L1	$\Delta$ CMI
All Data	<b>0.8717</b>	0.7102	0.9783	-3.41
Only Dictionary	0.1678	0.1193	0.6224	4.53
CSD + Monolingual	0.2392	0.1382	0.8157	-0.41
Only CSD	0.2803	0.1375	0.8269	-6.67

## 6.5 Language Grouping and Transfer Learning

To assess cross-lingual transfer, we compare three training strategies: (i) training each language independently, (ii) grouping languages by language family (Indo-Aryan or Dravidian), and (iii) training all six languages jointly. The impact of family-based grouping relative to individual training is summarized in Table 7, while the comparison between individual and fully joint training is reported in the supplementary material (Table 12).

Table 7: Effect of family-level grouping on the General test set. "Group-IA" = Indo-Aryan grouped; "Group-DR" = Dravidian grouped.

Language	Model Type	Macro F1	Prec. EN	Prec. L1
Hindi	Individual	<b>0.9723</b>	0.9647	0.9823
	Group - IA	0.9718	0.9642	0.9819
Bengali	Individual	0.8072	0.8154	0.8405
	Group - IA	<b>0.8243</b>	0.8235	0.8562
Marathi	Individual	0.6711	0.4194	0.9108
	Group - IA	<b>0.8875</b>	0.8723	0.9409
Telugu	Individual	<b>0.8586</b>	0.9020	0.8365
	Group - DR	0.7129	0.6216	0.9605
Kannada	Individual	0.5854	0.5652	0.6056
	Group - DR	<b>0.7920</b>	0.7734	0.8102
Malayalam	Individual	<b>0.7977</b>	0.8585	0.9688
	Group - DR	0.7677	0.8429	1.0000

Within-family grouping produces the strongest benefits, particularly for low-resource Marathi and Kannada. These improvements stem from shared English borrowing patterns and similar syntactic environments for mixing within the Indo-Aryan and Dravidian families. High-resource Hindi remains stable, showing only a minimal reduction due to mild negative transfer. Malayalam exhibits a small drop, which we attribute to inconsistent intra-word annotation conventions in the available training data.

We additionally evaluated parameter-efficient transfer through encoder freezing and adapter-based learning (results summarized in the supplemental material A.2). Freezing most of the encoder while fine-tuning only shallow task heads already yields competitive performance, especially for Marathi and Telugu, indicating that substantial cross-lingual transfer arises from shared subword representations. Adapter-based transfer further shows that lightweight modules trained jointly with multitask supervision recover a large portion of the full-model performance, making them attractive for scaling CSDI to new Indic languages with minimal additional parameters.

Together, these findings support strong within-family transfer and demonstrate that CSDI can be adapted effectively to new languages—either through full multitask fine-tuning or more parameter-efficient techniques.

## 6.6 Comparison with AnE

The final results show that CSDI matches or outperforms AnE across most language pairs in macro F1 while maintaining a smaller  $\Delta$ CMI, with the

largest gains on low-resource pairs (Mr–En, Bn–En). This indicates a more accurate estimation of code-mixing intensity, which is crucial for maintaining switch balance during translation.

A common weakness observed in the AnE system is that it often misclassifies homographic tokens between English and Indic in romanized form. For example, the Hindi phrase “के लिए” can appear as “*ke liye*” or “*key lie*”, overlapping with English dictionary entries and leading AnE to falsely predict English. CSDI, by contrast, leverages contextual embeddings and multitask signals (NER, POS) to correctly disambiguate many such cases.

However, both models still struggle with context-dependent named entities. For example, in “*Honey, kal mat aana*” (“Honey, don’t come tomorrow”), “*Honey*” is a person’s name rather than the English noun, yet both models misclassify it as English due to limited context within a single sentence.

## 6.7 Language-wise Analysis

- **Hi:** CSDI attains near-perfect F1, largely due to the abundance of training data for this pair.
- **Mr:** Marathi shows the largest  $\Delta$ CMI improvement, suggesting better modeling of code-switching boundaries despite limited data.
- **Bn, Kn, Te:** These languages show strong generalization with improved or comparable  $\Delta$ CMI and stable macro F1, indicating robust behavior across both Indo-Aryan and Dravidian families.
- **Ml:** CSDI performs poorly on the Malayalam–English pair on the internal dataset but nearly perfectly on the General dataset. This discrepancy is driven by inconsistent intra-word labeling: annotators labeled English-origin stems (e.g., “*drive-cheythu*”, “*call-um*”) as Malayalam, whereas our training scheme treated English lexical roots as English. This mismatch inflates false negatives and highlights the need for standardized intra-word code-switching annotation guidelines.

## 7 Conclusion

CSDI introduces a multitask learning framework for Indic–English code-switching detection that jointly models CSD, NER, and POS through a shared encoder with matrix-language conditioning.

Across six Indic language pairs, the model delivers strong and consistent gains, achieving state-of-the-art macro F1 and near-zero  $\Delta$ CMI, and outperforming the AnE baseline in both accuracy and code-mixing density estimation.

The effectiveness of CSDI stems from its ability to exploit syntactic and semantic cues from auxiliary tasks—particularly POS—to identify linguistically plausible switching boundaries. This architecture also enables meaningful cross-lingual transfer: high-resource languages such as Hindi provide structural signals that substantially benefit lower-resource languages like Marathi. The model’s stability across both clean and noisy evaluation settings underscores its practical utility for token-level language identification in code-mixed Indic text.

Overall, CSDI offers a robust and extensible foundation for scalable multilingual code-switching detection and serves as a reliable component for downstream applications, including code-switched machine translation.

## Limitations

CSDI, in its current form, handles only full-word code-switching. Intra-word mixing is assigned a single label based on the lexical origin of the token, which simplifies annotation but prevents the model from capturing finer grained subword structure — an issue particularly relevant for languages such as Malayalam. Addressing this would require tokenizers or encoders operating at character or byte levels.

The model also relies on sentence-level context and therefore struggles with discourse-dependent cases, including person names and contextually grounded entities that cannot be disambiguated from local information alone. Similarly, Romanized spellings that differ substantially from the training distribution reduce recall for English tokens, a consequence of inconsistent orthography and social-media spelling variation.

CSDI further assumes a single embedded language (English) within a base Indic language. While this reflects the predominant pattern in our datasets, it does not capture multilingual settings where more than one embedded language may appear.

Finally, the approach benefits from auxiliary supervision via NER, POS, and dictionary resources. For extremely low-resource languages where such

datasets are limited or unavailable, the gains from multitask learning may diminish, suggesting the need for task-agnostic or unsupervised alternatives.

## References

- Gustavo Aguilar, Sudipta Kar, and Tamar Solorio. 2020. [Lince: A centralized benchmark for linguistic code-switching evaluation](#). *Preprint*, arXiv:2005.04322.
- A. Bolonyai. 2009. Code-switching, imperfect acquisition, and attrition. In Barbara E. Bullock and Almeida Jacqueline Toribio, editors, *The Cambridge Handbook of Linguistic Code-switching*, Cambridge Handbooks in Language and Linguistics, pages 253–269. Cambridge University Press, Cambridge.
- Amitava Das and Björn Gambäck. 2014. [Identifying languages at the word level in code-mixed Indian social media text](#). In *Proceedings of the 11th International Conference on Natural Language Processing*, pages 378–387, Goa, India. NLP Association of India.
- Anik Dey and Pascale Fung. 2014. [A Hindi-English code-switching corpus](#). In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).
- Mrinal Dhar, Vaibhav Kumar, and Manish Shrivastava. 2018. [Enabling code-mixed translation: Parallel corpus creation and MT augmentation approach](#). In *Proceedings of the First Workshop on Linguistic Resources for Natural Language Processing*, pages 131–140, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- François Grosjean. 2010. *Bilingual: Life and Reality*. Harvard University Press.
- Janice Jake and Carol Myers-Scotton. 1997. [Codeswitching and compromise strategies: Implications for lexical structure](#). *International Journal of Bilingualism - INT J BILING*, 1:25–39.
- Anupam Jamatia, Amitava Das, and Björn Gambäck. 2018. [Deep learning-based language identification in english-hindi-bengali code-mixed social media corpora](#). *Journal of Intelligent Systems*, 28.
- Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. [IndicNLP Suite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961. Association for Computational Linguistics.
- Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip

- Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Gali, Vish Subramanian, and Partha Talukdar. 2021. [MuriL: Multilingual representations for indian languages](#). Preprint, arXiv:2103.10730.
- Yash Madhani, Sushane Parthan, Priyanka Bedekar, Gokul Nc, Ruchi Khapra, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Khapra. 2023. [Aksharantar: Open Indic-language transliteration datasets and models for the next billion users](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 40–57, Singapore. Association for Computational Linguistics.
- Pieter Muysken. 1995. *Code-switching and grammatical theory*, pages 177–198. Cambridge University Press, Cambridge.
- Parth Patwa, Gustavo Aguilar, Sudipta Kar, Suraj Pandey, Srinivas PYKL, Björn Gambäck, Tanmoy Chakraborty, Tamar Solorio, and Amitava Das. 2020. [SemEval-2020 task 9: Overview of sentiment analysis of code-mixed tweets](#). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 774–790, Barcelona (online). International Committee for Computational Linguistics.
- Omkar Pitale and Aruna Malapati. 2023. [Leveraging multi-task learning in code-switched setting](#). Preprint.
- Shana Poplack. 1981. Syntactic structure and social function of code-switching. In Richard P. Durán, editor, *Latino Language and Communicative Behavior*, pages 169–184. Ablex, Norwood, NJ.
- Adithya Pratapa and Monojit Choudhury. 2017. [Quantitative characterization of code switching patterns in complex multi-party conversations: A case study on Hindi movie scripts](#). In *Proceedings of the 14th International Conference on Natural Language Processing (ICON-2017)*, pages 75–84, Kolkata, India. NLP Association of India.
- Sebastin Santy, Anirudh Srinivasan, and Monojit Choudhury. 2021. [BERTologiCoMix: How does code-mixing interact with multilingual BERT?](#) In *Proceedings of the Second Workshop on Domain Adaptation for NLP*, pages 111–121, Kyiv, Ukraine. Association for Computational Linguistics.
- HAZ Sameen Shahgir, Ramisa Alam, and Md. Zarif Ul Alam. 2023. [Banglaconer: Towards robust bangla complex named entity recognition](#). Preprint, arXiv:2303.09306.
- Igor Sterner. 2024. [Multilingual identification of English code-switching](#). In *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024)*, pages 163–173, Mexico City, Mexico. Association for Computational Linguistics.
- Genta Indra Winata, Andrea Madotto, Chien-Sheng Wu, and Pascale Fung. 2018. [Code-switching language modeling using syntax-aware multi-task learning](#). In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, pages 62–67, Melbourne, Australia. Association for Computational Linguistics.

## A Appendix

### A.1 Dataset Supplementary Material

#### A.1.1 Training and Test Set Overviews

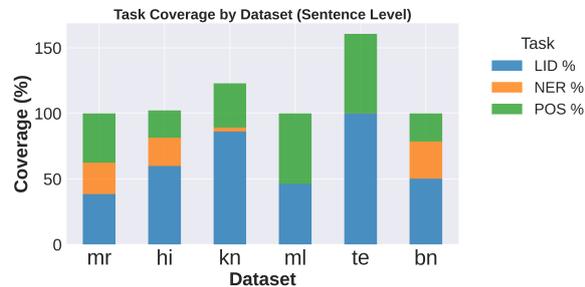


Figure 4: Task coverage across training datasets, showing the proportion of sentences annotated with CSD (LID), NER, and POS for each language.

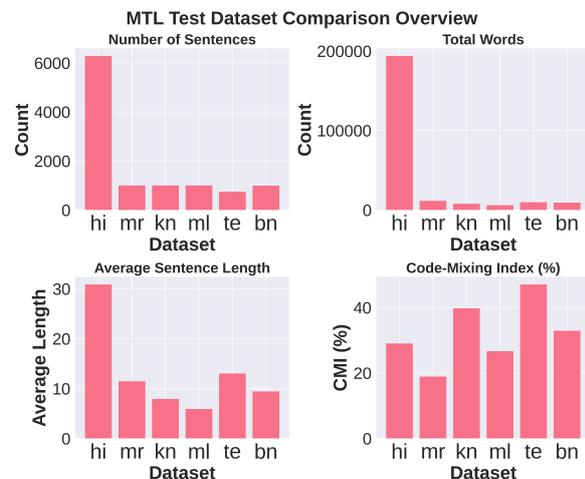


Figure 5: General Test Dataset statistics: number of sentences, total word count, average sentence length, and CMI for each language.

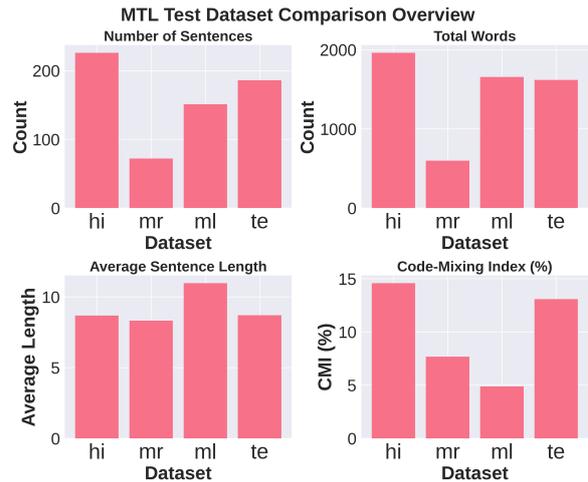


Figure 6: Internal Test Dataset statistics: number of sentences, total words, average sentence length, and CMI across languages.

#### A.1.2 Dataset Inventory

Table 8 provides a consolidated inventory of all datasets used in this work.

Table 8: Dataset used with descriptions and sources.

Dataset	Description	Source
english-words	Curated English word list used as a lexical resource	<a href="https://github.com/dwyl/english-words">https://github.com/dwyl/english-words</a>
Aksharantar	Indic transliteration corpus	(Madhani et al., 2023)
OpenSLR	OpenSLR-104 code-switched Indic transcripts	<a href="https://www.openslr.org/104/">https://www.openslr.org/104/</a>
L3Cube-HingLID	Hindi-English code-mixed language identification dataset	<a href="https://github.com/l3cube-pune/code-mixed-nlp">https://github.com/l3cube-pune/code-mixed-nlp</a>
Kanglishicon	Kannada-English code-mixed text dataset	<a href="https://sites.google.com/view/kanglishicon2022/dataset">https://sites.google.com/view/kanglishicon2022/dataset</a>
WoLLaI	WoLLaI Mal-Eng	<a href="https://data.mendeley.com/datasets/tzrcrrwz4n/1">https://data.mendeley.com/datasets/tzrcrrwz4n/1</a>
L3Cube-MahaNLP	Marathi dataset subset used in prior code-mixed NLP releases	<a href="https://github.com/Harry262000/Marathi-Dataset">https://github.com/Harry262000/Marathi-Dataset</a>
CoSTA	Silver labelled dataset for Bn, Hi, Mr, Te	<a href="https://github.com/csalt-research/CoSTA">https://github.com/csalt-research/CoSTA</a>
Bangla-Complex-NER	Bangla complex named entity recognition dataset	(Shahgir et al., 2023)
LINCE	LINCE benchmark providing Hindi NER and POS annotations	(Aguilar et al., 2020)
KCIS (Bn, Kn, Mr, Ml)	POS-tagged Indian language resources released by IIIT-Hyderabad	<a href="https://ltrc.iiit.ac.in/showfile.php?filename=downloads/kolhi/">https://ltrc.iiit.ac.in/showfile.php?filename=downloads/kolhi/</a>
Telugu-POS	Telugu-English POS-tagged code-mixed dataset	GitHub: SunilGundapu / Word-Level-LID-TE

#### A.1.3 Filtering English-Origin Entries in Indic Dictionaries

Aksharantar Indic dictionaries contain a small but non-negligible number of English-origin words transliterated into Indic scripts. While such forms

may be considered lexical borrowings, treating them as Indic tokens introduces ambiguity for token-level code-switching detection, where the goal is to identify English insertions in Indic text.

To quantify this overlap, we compare transliterated dictionary vocabularies across languages against an English word list and computed lexical overlap statistics. We observe a non-trivial overlap between English and Indic dictionaries, as well as homographic collisions across Indic languages themselves. Figure 7 illustrates the pairwise Jaccard similarity of dictionary vocabularies across languages and English.

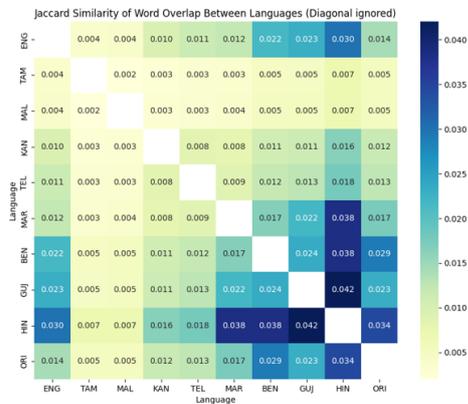


Figure 7: Pairwise Jaccard similarity of word overlap between transliterated dictionary vocabularies across languages and English (diagonal ignored). Higher overlap between English and Indo-Aryan languages reflects a greater presence of shared or borrowed romanized forms.

We observe substantially higher overlap between English and Indo-Aryan languages (e.g., Hindi, Marathi, Bengali, Gujarati) than with Dravidian languages. This pattern motivates more aggressive filtering of English-origin entries for Indo-Aryan dictionaries in order to reduce systematic noise during training.

Based on this analysis, dictionary entries with high-confidence overlap with English forms were removed from the Indic dictionaries prior to training. Table 9 shows representative examples of English-origin entries removed from the Hindi dictionary; similar patterns are observed for other languages.

Table 9: Examples of English-origin entries removed from the Hindi dictionary.

Indic form	Romanized form	English gloss
क्रैडल	cradle	cradle
क्वॉरिन्टाइन	quarantine	quarantine
इअर	ear	ear
ऑडिऐंस	audience	audience
ऐंजल	angel	angel
लूवे	loove	love
ऑफर:	offerh	offer

#### A.1.4 Example of Romanization Augmentation

To illustrate the effect of our round-trip transliteration augmentation, consider the Hindi sentence:

*kal school mai new teacher aai thi*

Augmentation yields several plausible variants:

- *kal skool mai new techar ayi thi*
- *kal iskool me new teacher aayi thi*
- *kal school mey nyu teacher aaii thi*
- *kal skul mai new teacher ayi thi*
- *kal skool me new ticher aayee thi*

These reflect natural variation found in social media and informal writing.

## A.2 Transfer Learning

### A.2.1 Encoder Freezing Experiments

Table 10 reports performance on the Internal test set for multiple encoder-freezing strategies compared against full fine-tuning during sequential transfer from Hindi to Marathi and from Kannada to Telugu and Malayalam. The evaluated variants include partial freezing of different encoder segments (Low–Mid: layers 1–6, Mid–High: layers 6–12, and Mid: layers 4–9), freezing the entire encoder (“Encoder”), and freezing both the encoder and the POS head while training only the remaining task heads (“Encoder+POS”). Across all languages, freezing the full encoder consistently outperforms both partial freezing strategies and full fine-tuning.

Table 10: Performance of partial encoder freezing vs. full finetuning for Mr, Te, MI on internal test data

Lang	Low-Mid	Mid-High	Mid	Encoder	Enc+POS	Full FT
MR	0.609	0.500	0.566	<b>0.660</b>	0.604	0.569
TE	0.794	0.807	0.776	<b>0.811</b>	0.806	0.803
ML	0.228	0.494	0.213	<b>0.506</b>	0.447	0.456

### A.2.2 Adapter-Based Transfer

Table 11 presents a comparison between adapter-based transfer strategies and full multitask fine-tuning for Marathi and Telugu on Internal test set. Language-specific adapters trained in isolation perform poorly, while zero-shot transfer using a Hindi-trained adapter provides a clear improvement. Combining language adapters with multitask learning substantially improves performance, narrowing the gap to full multitask fine-tuning. Overall, the results indicate that adapters offer a viable parameter-efficient alternative, although full fine-tuning remains the strongest setting.

Table 11: Adapter-based transfer vs. full multitask fine-tuning for Mr, Te.

Lang	Lang Adapter	Lang+MTL	Zero-shot	Full MTL
MR	0.3418	0.5208	0.4951	<b>0.6078</b>
TE	0.5737	0.7921	0.7625	<b>0.7915</b>

### A.2.3 Six-Language Joint Training

Earlier, we compared individual-language training with family-based grouping of related languages. Here, we extend this analysis by comparing individual-language models with joint training across all six languages.

Table 12: Comparison of individual-language models and joint six-language multitask training on the General test set.

Language	Model Type	Macro F1	Prec. EN	Prec. L1
Hindi	Individual	0.9723	0.9647	0.9823
	Joint	0.9721	0.9655	0.9818
Bengali	Individual	0.8072	0.8154	0.8405
	Joint	0.8056	0.7857	0.8481
Marathi	Individual	0.6711	0.4194	0.9108
	Joint	<b>0.8541</b>	0.7143	0.9572
Telugu	Individual	0.8586	0.9020	0.8365
	Joint	<b>0.9227</b>	0.9153	0.9301
Kannada	Individual	0.5854	0.5652	0.6056
	Joint	<b>0.8226</b>	0.7826	0.8661
Malayalam	Individual	0.7977	0.8585	0.9688
	Joint	<b>0.8112</b>	0.8634	1.0000

## A.3 Full Results for All Test Sets

Table 13: Results on both the General test set and the Internal Movie Dialogue test set.

Lang. Pair	System	English			L1			Overall		
		P	R	F1	P	R	F1	Acc	MF1	$ \Delta\text{CMI} $
<i>General Test Set</i>										
Hi-En	AnE	0.93	0.70	0.80	0.89	0.98	0.93	0.90	0.86	7.26
	CSDI	<b>0.98</b>	0.96	<b>0.97</b>	0.98	0.99	0.99	0.98	<b>0.98</b>	<b>0.48</b>
Mr-En	AnE	0.83	0.95	0.88	0.98	0.95	0.97	0.95	0.93	2.65
	CSDI	<b>0.93</b>	0.94	<b>0.94</b>	0.98	0.98	0.98	0.98	<b>0.96</b>	0.70
Bn-En	AnE	0.79	0.89	0.84	0.94	0.88	0.91	0.88	0.87	4.33
	CSDI	<b>0.95</b>	0.95	<b>0.95</b>	0.97	0.97	0.97	0.97	<b>0.96</b>	<b>0.29</b>
Kn-En	AnE	0.88	0.96	0.92	0.93	0.81	0.86	0.90	0.89	5.12
	CSDI	<b>0.92</b>	0.95	<b>0.93</b>	0.93	0.87	0.90	0.92	<b>0.91</b>	<b>2.40</b>
Ml-En	AnE	0.99	0.81	0.89	0.66	0.98	0.79	0.86	0.84	13.30
	CSDI	0.99	0.99	<b>0.99</b>	0.99	0.99	0.99	0.99	<b>0.99</b>	<b>0.08</b>
Te-En	AnE	0.93	0.93	0.93	0.92	0.92	0.92	0.93	0.93	0.07
	CSDI	<b>0.98</b>	0.99	<b>0.99</b>	0.99	0.98	0.98	0.98	<b>0.98</b>	0.32
<i>Internal Movie Dialogue Test Set</i>										
Hi-En	AnE	0.54	0.83	0.66	0.96	0.88	0.92	0.87	0.79	7.54
	CSDI	<b>0.74</b>	0.88	<b>0.80</b>	0.98	0.95	0.96	0.94	<b>0.88</b>	<b>2.65</b>
Mr-En	AnE	0.49	0.67	0.56	0.97	0.94	0.95	0.92	0.76	2.83
	CSDI	<b>0.60</b>	0.74	<b>0.66</b>	0.96	0.98	0.97	0.94	<b>0.81</b>	<b>0.00</b>
Ml-En	AnE	0.90	0.67	0.79	0.97	0.99	0.98	0.97	0.87	1.95
	CSDI	0.43	0.81	0.56	0.98	0.92	0.95	0.91	0.76	6.56
Te-En	AnE	0.75	0.82	0.78	0.97	0.96	0.96	0.94	0.87	1.05
	CSDI	<b>0.80</b>	0.79	<b>0.79</b>	0.97	0.97	0.97	0.95	<b>0.88</b>	<b>0.24</b>

# Vinclat: Evaluating Reasoning, Cognition and Culture in One Game

Marc Pàmies   Javier Aula-Blasco   Aitor Gonzalez-Agirre   Marta Villegas  
Barcelona Supercomputing Center  
mpamies@bsc.es

## Abstract

This paper introduces *Vinclat*, a novel evaluation dataset for Catalan carefully designed to assess the reasoning capabilities and cultural knowledge of LLMs. It comprises 1,000 high-quality instances, meticulously crafted and reviewed by human annotators. Each instance presents a complex riddle that requires a two-step reasoning process involving inferential and abductive reasoning, along with other cognitive skills such as lexical retrieval, paraphrasing, flexibility in interpretation, pattern recognition, and associative thinking. Given four independent clues, models should infer intermediate concepts which, despite being seemingly unrelated, can be creatively connected to reach a final solution. The task targets a unique blend of capabilities, distinguishing it from existing NLP benchmarks. Our evaluation of state-of-the-art models reveals that these still fall significantly short of human-level reasoning, although scaling trends suggest that the performance gap may narrow over time. This indicates that *Vinclat* provides a robust and long-term challenge, resisting the rapid saturation that is commonly observed in many existing evaluation datasets.

## 1 Introduction

In recent years, large language models (LLMs) have made groundbreaking progress in natural language processing (NLP), achieving state-of-the-art performance in tasks ranging from straightforward factual knowledge (Krathwohl, 2002) question answering (Mihaylov et al., 2018; Petroni et al., 2021), to more complex tasks at the human cognitive level such as multi-step reasoning (Cobbe et al., 2021; Srivastava et al., 2023). More recently, increasing attention has been directed towards enhancing the multilingual reasoning capabilities of these models, as demonstrated by recent releases that highlight the significant impact of improved reasoning in solving complex problems across languages (Guo et al., 2025; Zheng et al., 2025).

Despite these advances, assessing further complex reasoning capabilities entangled with deep contextual knowledge remains a significant challenge, particularly in creative and multilingual contexts, where conventional benchmarks often fall short (Ghosh et al., 2025; Shojaee et al., 2025). There is a notable scarcity of downstream tasks that have been explicitly crafted to evaluate the equivalent to complex human cognitive skills in LLMs, leading to the widespread use of coding (Gu et al., 2024; Zhuo et al., 2024) and mathematical (Mishra et al., 2022; Glazer et al., 2024) datasets as proxies. While these datasets can provide some insight, they are insufficient to comprehensively capture the nuanced and multifaceted reasoning abilities required for broader cognitively complex tasks. This gap underscores the need for a more diverse suite of reasoning-intensive NLP tasks that pave the way for holistic evaluation frameworks, which is critical to advance LLMs towards more human-like intelligence.

We believe that problem-solving datasets derived from games can serve as effective tools to evaluate progress in this area. Despite being often regarded as mere entertainment, games have a long history of being used as valuable test-beds for artificial intelligence research (Hu et al., 2024; Silver et al., 2016, 2017; Vinyals et al., 2019). This is primarily due to the fact that games offer a safe, controlled environment that simulates real-world scenarios and can often be generalized to more complex domains. Moreover, the well-defined rules and objectives inherent in games make it relatively straightforward to assess success and measure performance. Building on this belief, we introduce *Vinclat*<sup>1</sup>, a Catalan-language dataset for multi-step problem solving that employs a game-based structure to evaluate both the reasoning capabilities of LLMs and the depth of their cultural awareness.

<sup>1</sup><https://hf.co/datasets/projecte-aina/vinclat>

## 2 The Task

### 2.1 Task Definition

The main objective is to guess a target word or term (*solution*) based on a set of *keywords* that are obtained from four given *hints* (see Figure 1). Notably, the length of the target word(s) is provided in advance, which introduces a structural constraint that guides the reasoning process. However, successfully solving the task is more complex than it might appear at first glance, as it involves two distinct steps, each requiring different types of reasoning abilities (see §2.2).

- **Step 1.** First, a plausible term must be inferred for each hint. These intermediate terms, or *keywords*, are not constrained by a fixed length, and it is not strictly necessary to identify all of them correctly to arrive at the final solution (see Table 1).
- **Step 2.** In this phase, the set of seemingly unrelated terms obtained in Step 1 must be jointly interpreted to uncover a hidden semantic or conceptual link. This requires synthesizing the intermediate answers into a single unifying term that captures their commonality, often involving associative and categorical reasoning (see Table 2).

### 2.2 Involved Cognitive Abilities

Solving a *Vinclat* requires a broad set of human-like cognitive skills. LLMs evaluated with *Vinclat* need to be able to jointly demonstrate all of these

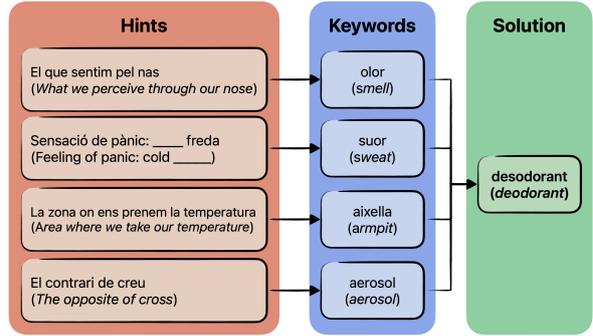


Figure 1: Example of a *Vinclat*. Here, only the *Hints* and number of letters of the *Solution* are provided. The *Keywords* and the actual *Solution* are to be guessed.

capabilities in order to successfully complete the task. First, players need to access precise lexical items from memory, especially when clues are vague or indirect—a process known as lexical retrieval (Levett, 1989), often mirrored in factual recall benchmarks like LAMA (Petroni et al., 2019) or MMLU (Hendrycks et al., 2020). Beyond direct recall, *Vinclat* challenges players to draw inferences. For instance, inferring that “*fa llum però no crema*” (it lights up but doesn’t burn) refers to an “LED bulb” requires bridging gaps between literal meaning and plausible explanation, similar to what ARC (Clark et al., 2018) or StrategyQA (Geva et al., 2021) test in LLMs. These clues often follow implicit patterns such as recognizing that “*martell*” (hammer), “*trepant*” (drill), “*tornavís*” (screwdriver) and “*serra*” (saw) all belong to the category “*eines*” (tools), demanding a form of analogy-making and semantic pattern recognition

Hint	Keyword	Explanation
La filla de la paciència (The daughter of patience)	ciència (science)	Based on the Catalan proverb “ <i>La paciència és la mare de la ciència</i> ” (which translates to “patience is the mother of science”), it should be inferred that the daughter of <i>patience</i> is <i>science</i> .
Roma cap per avall (Rome upside down)	amor (love)	The word “ <i>Roma</i> ” (Rome) should be read backwards as “ <i>amor</i> ” (Love), despite not being explicitly stated in the hint.
El 25 de desembre, ____, ____, ____ (On December 25, ____ ____)	fum (smoke)	Successfully deciphering this fill-in-the-blank hint requires knowledge of the lyrics of a traditional Catalan Christmas carol.
De sol, digital o d’agulles (sun, digital or needle)	Rellotge (clock)	All three words refer to types of clocks.

Table 1: Examples showcasing the reasoning involved in *Vinclat*’s first step.

Keywords	Solution
Reina ( <i>queen</i> ) Egipte ( <i>Egypt</i> ) Llet ( <i>milk</i> ) Serp ( <i>snake</i> )	Cleopatra
Arbre ( <i>tree</i> ) Bombardeig ( <i>bombing</i> ) Quadre ( <i>painting</i> ) País Basc ( <i>Basque Country</i> )	Guernica

Table 2: Examples illustrating the associative and categorical reasoning required in *Vinclat*’s second step.

that benchmarks like ConceptARC (Lörer et al., 2023) and some BIG-bench sub-tasks (Srivastava et al., 2022) aim to capture. *Vinclat* also encourages players to mentally test possible solutions before committing: this trial-and-error process reflects hypothesis generation and testing (Klahr and Dunbar, 1988), a capability explored in datasets such as GSM8K (Cobbe et al., 2021) and CLUTRR (Sinha et al., 2019). Interestingly, interpreting figurative or idiomatic clues calls for metalinguistic awareness, the capacity to analyze language abstractly (Bialystok, 1986), a skill also critical for disambiguation tasks like the Winograd Schema Challenge in SuperGLUE (Wang et al., 2019). Lastly, when one or more clues are unclear, *Vinclat* players often rely on semantic guessing—filling in gaps based on context and heuristics, a behavior echoed in masked language models like BERT (Devlin et al., 2019) and rooted in classic work on memory interpolation (Bartlett, 1932). All these cognitive mechanisms reflect a rich interplay of language, reasoning, and abstraction that makes *Vinclat* a considerably challenging task.

### 3 The Dataset

The *Vinclat* dataset contains 1,000 instances that were carefully conceived by native Catalan speakers. As mentioned, every instance entails four independent hints, each pointing to an intermediate term or keyword. The model must first infer these intermediate terms and then uncover a final solution term, whose number of letters and words is provided as an additional hint. The final solution is designed so that its specified word and letter pattern is unique. Other plausible answers may exist, but none match the given structure.

### 3.1 Data Creation and Curation

The dataset was created by the original authors of the *Vinclat* game, who released a new instance per day on the official website<sup>2</sup>. The authors met on a weekly basis in collaborative sessions to produce the puzzles for the upcoming period. During these sessions, they aimed to:

- ensure topical diversity, avoiding repetition across instances,
- vary difficulty levels by including at least one relatively easy clue to support solvability,
- maintain linguistic and semantic coherence while encouraging creative associations.

We manually revised the original dataset to ensure that all instances can be interpreted in isolation, improving the fairness and reproducibility of the task. During this curation, we revised instances that were time-dependent or game-specific, as these relied on information unavailable to both LLMs and human annotators (see §4.5). For example, “*Les festes que gaudirem en un mes*” (“The holidays we’ll enjoy in a month”) was rewritten as “*Les festes que gaudirem el mes de desembre*” (The holidays we’ll enjoy in December) to remove the need for knowing the date. Similarly, clues like “*La resposta d’ahir*” (Yesterday’s answer)—which originally referred to “*cervell*” (brain)—were replaced with self-contained alternatives such as “*La part més gran de l’encèfal humà*” (The largest part of the human encephalon). This handcrafted nature of the dataset, along with its careful attention to cultural relevance and cognitive challenge, makes it a unique and high-quality resource for evaluating reasoning and association capabilities in LLMs.

Crucially, the dataset was never published in bulk or made publicly available in machine-readable format prior to this study. The only access point to individual instances was through the daily publication of a single puzzle on the website, which limits their presence in large-scale training corpora typically used for LLM pre-training. Since December 2nd, 2024, no new instances have been added, and all previously published puzzles have been removed from the website, making them unavailable online. Furthermore, the highly localized and manually authored nature of the dataset—including its use of Catalan cultural references, idiomatic expressions, and unpublished formulations—makes

<sup>2</sup><https://vinclat.cat>

it extremely unlikely that LLMs have been exposed to it or anything similar during pre- or post-training. For these reasons, we consider the risk of data contamination to be negligible. To preserve long-term validity, we release only instance-level hints and maintain an external leaderboard<sup>3</sup> where models are evaluated and added over time.

### 3.2 Content Domains and Cultural Context

The dataset covers a wide range of topics to ensure both variation in difficulty and broad appeal. Hints and solutions touch upon areas such as arts and culture (e.g., theatre, literature, music, cinema), general knowledge (e.g., science, history, geography), daily life and common sense associations, politics and society, philosophical or abstract concepts, famous personalities, religion and spirituality, and Catalan popular sayings, idioms, traditions and folklore, among others. This topical diversity contributes to the complexity and richness of the task, as it requires both semantic flexibility and cultural awareness.

Although the dataset is deeply rooted in the Catalan cultural and linguistic context, not all instances rely exclusively on local knowledge. Many hints reflect Catalonia’s traditions, idiomatic expressions, social references, and shared historical memory, which require a certain level of cultural grounding for successful resolution. At the same time, the dataset deliberately includes hints from more global or universal domains. This careful balance between local and global content reflects an intent to design a game that is culturally situated yet broadly accessible, showcasing the linguistic and conceptual nuances of Catalan while remaining solvable by educated native speakers. As a result, *Vinclat* provides a particularly rich resource for evaluating the interplay between language, culture, and reasoning in LLMs.

## 4 Evaluation

### 4.1 Baselines

As this task involves multi-step reasoning guided by explicit instructions, it fundamentally relies on the model’s ability to understand and follow natural language directives. Recent findings suggest that *base* models (i.e., pre-trained solely via next-token prediction) often fail on instruction-heavy datasets, particularly when answers must conform to specific

formats or require tool-like reasoning behaviors (Liang et al., 2023; Zhou et al., 2023). As such, we do not evaluate *base* models, as they have not been optimized for instruction-following behavior.

For baseline evaluation, we select the top-performing open-source models with demonstrated reasoning capabilities, as preliminary experiments revealed that non-reasoning LLMs, in spite of being instructed, yield remarkably low results. In particular, we evaluate the performance of DeepSeek-V3-0324 (Liu et al., 2024), DeepSeek-R1 (Guo et al., 2025) and the Qwen3 family of models (Zheng et al., 2025). The latter offers an especially compelling case study, as it allows us to analyze if performance scales with model size. Furthermore, Qwen3’s ability to turn thinking on and off at will provides a controlled setup to quantify the benefits of forcing an LLM to deliberate before answering. The inclusion of DeepSeek-V3-0324, which can be considered the non-reasoning counterpart to DeepSeek-R1, allows us to investigate the performance gap attributable to reasoning capabilities within the same architecture. As new models become available, we plan to evaluate the most promising ones and publish their results on a public leaderboard.

### 4.2 Implementation Details

Evaluated models were served as endpoints using the vLLM library (v.0.8.5),<sup>4</sup> enabling seamless interaction with an OpenAI-compatible server. This setup allowed the models to function as if they were part of a hosted API service, facilitating inference requests in a standardized format compatible with the `openai` python library (v.1.60.2).<sup>5</sup> The compute nodes in the cluster used for evaluation are equipped with four 64GB H100 GPUs, which results in a limited VRAM capacity to accommodate the largest models in a single node. To address this, we use the dockerized version of vLLM<sup>6</sup> for distributed inference. Specifically, in such cases, four nodes were employed with both the tensor and pipeline parallel sizes set to 4, ensuring efficient distributed processing (Kwon et al., 2023).

For generation parameters, we adhered to the default configurations recommended by model providers, thereby eliminating the need for extensive parameter tuning. Regarding the number of generation tokens, the limit was dynamically set to

<sup>3</sup><https://hf.co/spaces/projecte-aina/vinclat-leaderboard>

<sup>4</sup><https://github.com/vllm-project/vllm>

<sup>5</sup><https://pypi.org/project/openai/>

<sup>6</sup><https://hub.docker.com/r/vllm/vllm-openai>

the model’s context window size minus the length of the largest prompt in the batch, providing the LLM with as many tokens as possible to carry out its reasoning process. It is worth noting that, given the nature of the dataset, prompt lengths exhibit minimal variation and are significantly shorter—around 300 tokens— than the available context window (i.e., 163,840 tokens for DeepSeek and 40,960 for Qwen3). The inference and evaluation scripts are fully available to the public.<sup>7</sup>

### 4.3 Prompt Templates

We designed two prompt templates that differ in how they convey the structure of the target solution: a fill-in-the-blanks format and a letter-count format. Each instance was evaluated using both templates, in two languages, Catalan and English, resulting in four total prompt variants per instance (see Appendix A). This allows us to analyze model performance across different representations of structural constraints and languages of instruction, while keeping the content (the four hints and the solution) consistently in Catalan.

Each prompt begins with the same task description, which provides models with a clear breakdown of the two-stage reasoning process required by the game. The prompt then diverges into one of two formats:

- **Letter-count.** This variant presents the expected structure of the solution in natural language (e.g., “The solution has 2 words with the following structure: 2 letters for the first word and 5 letters for the second word.”). This evaluates whether the model can reason over symbolic constraints and integrate them with thematic associations from the hint words.
- **Fill-in-the-blanks.** This *cloze-style* variant encodes the structure of the solution using underscore placeholders for each letter in the solution. For example, for a two-word solution with two letters in the first word and three in the second, the prompt would include “The solution should fit here: \_\_ \_\_\_” (i.e., two underscores followed by a whitespace and three underscores). This evaluates whether the model can interpret and use visual or subword-level cues to generate appropriate candidate solutions.

Using both formats allows us to assess whether models rely on abstract understanding of word constraints (e.g., word counts and lengths in text) versus those that leverage pattern recognition or visual structure matching (Bubeck et al., 2023; Li et al., 2024). As mentioned, both prompt formats were used in Catalan and English. This design choice enables us to evaluate models’ ability to follow instructions in Catalan, a relatively low-resource language, while also isolating the effect of prompt language on performance. Given that many large language models are primarily post-trained in English, this comparison allows us to examine how language choice in prompting influences model behavior. At the same time, it reflects realistic usage scenarios in multilingual regions, where users may interact with systems in more than one language or receive instructions in one language while processing content in another.

### 4.4 Scoring Criteria

An instance is considered to be correctly solved if and only if the model produces the final solution, regardless of the correctness and number of keywords identified in the intermediate step. The rationale for not penalizing keywords is that they are not necessarily fixed terms; variations or synonyms can also contribute to a successful solution. Additionally, no penalty is applied for failing to identify all four keywords, as omitting unclear or misleading clues can be advantageous. In such cases, disregarding a specific hint is preferable to selecting an incorrect keyword that could divert the solution process away from the right final answer.

The final part of the prompt specifies that the response must be returned in JSON format, making it easier to parse the LLM’s output. After extracting the dictionary—which, in reasoning models, appears immediately after the `</think>` tag—the correctness of the solution was verified using exact match. All responses were lowercased prior to validation in order to ensure that correct answers were not penalized due to differences in letter casing.

### 4.5 Human Baseline

To establish a human baseline for the task, we conducted a targeted human evaluation on a sample of 300 instances randomly drawn from the full dataset. This subset was double-checked and confirmed to balance feasibility and diversity while capturing a representative cross-section of the task’s difficulty and content variety.

<sup>7</sup>[https://github.com/mapama247/vinclat\\_eval](https://github.com/mapama247/vinclat_eval)

We recruited three human annotators to each complete all 300 selected instances under different conditions designed to simulate varying levels of resource availability and knowledge access:

- **Annotator A** completed the task without access to the internet, relying solely on personal knowledge and reasoning.
- **Annotator B** had full access to the internet, simulating an open-book approach to answering each clue and inferring the final keyword.
- **Annotator C** was allowed to use the internet for 50% of the instances (150 out of 300). These instances were selected randomly and balanced across the sample to mitigate potential order or topic bias.

This design allows us to assess not only the upper-bound performance of knowledgeable human participants with external support, but also the limits of unaided reasoning, and the mixed conditions in between. All annotators were native Catalan speakers, born and raised in Catalonia, and all held university degrees (with some having completed postgraduate studies). This ensured both linguistic and cultural familiarity with the game and its underlying associations. Annotators were asked to leave instances unanswered if they failed to find a solution within 5-7 minutes of trying, as this avoided prolonged effort on especially tricky instances.

Human performance varied substantially depending on internet access. In the offline condition, annotators answered on average 61% of the instances, with an accuracy of 82% on answered items, yielding an overall accuracy of 50% when accounting for unanswered cases. In the online condition, response rates increased to an average of 84%, with an accuracy of 85% on answered items and an overall accuracy of 71%. These results show that access to online resources markedly improves both coverage and total accuracy, primarily by reducing the number of skipped items.

To quantify the uncertainty of these estimates, we computed 95% confidence intervals for the proportion of correct responses using the standard normal approximation for proportions:

$$CI = \hat{p} \pm Z \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

Model	EN_LC	EN_FB	CA_LC	CA_FB	Avg.	UNKs
<b>Without Reasoning</b>						
Qwen3-0.6B	0	0	0	0	<b>0</b>	2,575
Qwen3-1.7B	0	0.1	0	0.1	<b>0.05</b>	98
Qwen3-4B	0.2	0.3	0.2	0.4	<b>0.28</b>	2.25
Qwen3-8B	0.7	0.5	0.6	1.0	<b>0.7</b>	0
Qwen3-14B	0.7	1.4	1.9	1.4	<b>1.35</b>	13.25
Qwen3-32B	1.3	2.0	1.8	2.0	<b>1.78</b>	0.25
Qwen3-30B-A3B	1.3	1.3	1.2	1.4	<b>1.3</b>	3.75
Qwen3-235B-A22B	11.5	11.0	7.6	6.3	<b>9.1</b>	9.75
DeepSeek-V3-0324	13.1	7.6	9.4	8.3	<b>9.6</b>	29.25
<b>With Reasoning</b>						
Qwen3-0.6B	0.1	0	0	0.1	<b>0.05</b>	767.5
Qwen3-1.7B	0.4	0	0.1	0.1	<b>0.15</b>	402.25
Qwen3-4B	0.7	0.5	1.0	0.8	<b>0.75</b>	28.25
Qwen3-8B	2.6	1.0	1.7	1.2	<b>1.63</b>	122.25
Qwen3-14B	4.2	3.1	2.5	2.3	<b>3.03</b>	279.5
Qwen3-32B	7.0	5.7	6.5	5.2	<b>6.1</b>	98.67
Qwen3-30B-A3B	2.9	1.8	2.9	1.9	<b>2.38</b>	125
Qwen3-235B-A22B	22.3	13.4	21.2	13.0	<b>17.48</b>	103.5
DeepSeek-R1	33.0	22.7	33.5	23.8	<b>28.25</b>	109

Table 3: Accuracy per model. "EN" and "CA" refer to the prompt language; "LC" refers to the letter-count format and "FB" to fill-in-the-blanks. Column "UNKs" refers to the amount of keywords tagged as "unknown" by the model, averaged across prompts.

where  $\hat{p}$  is the observed accuracy,  $n$  the number of instances annotated, and  $Z = 1.96$  corresponds to a 95% confidence level (Agresti and Coull, 1998). The resulting intervals are 47-59.4% for offline, and 66.9-75.5% for online, supporting that the measured performance provides a statistically robust estimate of human capabilities on the full dataset. It is worth noting that all original *Vinciat* instances were confirmed to be solvable by the original authors, as records from the online game show that there was always a subset of players who successfully solved each instance.

To complement the quantitative evaluation, a human annotator different to the three above but with a similar profile conducted a qualitative error analysis of model outputs. For each model-language-prompt configuration, a minimum of 25 incorrectly solved instances were randomly selected. The annotator was allowed to explore more instances for any configuration if they deemed it necessary to complete their analysis, totaling over 1,800 cases analyzed. The annotator examined the reasoning traces (when available) and final answers, identifying recurrent error types such as hint misinterpretation, linguistic or cultural misunderstanding, hallucination, and structural inconsistency. This manual review provided fine-grained insights into the nature of model failures and supported the interpretation of quantitative trends discussed in §5.

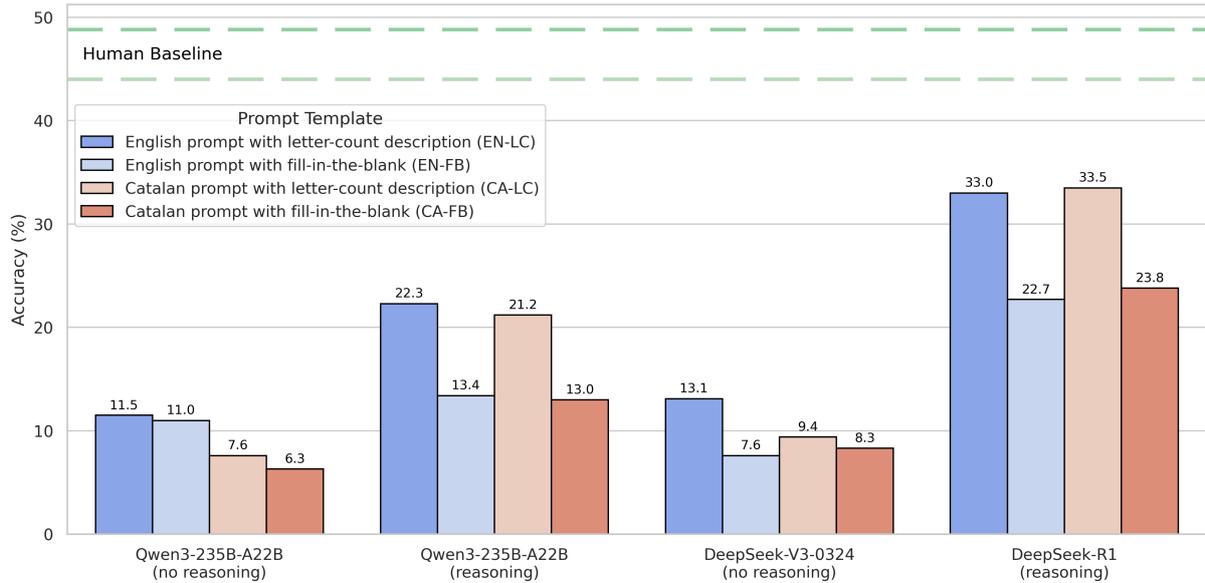


Figure 2: Accuracy scores based on prompt format and reasoning capabilities.

## 5 Results and Discussion

**Comparison to human baselines** Across all evaluations, large language models perform significantly worse than human annotators, regardless of internet access (see Table 3 and Figure 2). The highest-performing model only reaches 33.5% accuracy, while the majority of evaluated models struggle to exceed 20%, and several models remain below 10%. In contrast, human annotators, both with and without full access to online resources, performed considerably better, confirming that human-level commonsense reasoning and cultural background knowledge—especially in a task deeply rooted in local culture—remain out of reach for current LLMs. This mirrors findings in other abductive or commonsense-heavy tasks, where humans considerably outperform models (Bhagavatula et al., 2020; Loredó Lopez et al., 2025), suggesting that such deficits are not merely language- or culture-specific but reflect broader limitations in the reasoning mechanisms of current large language models.

**Impact of model size and reasoning** There is a clear correlation between model size and performance in the *Vinclat* task, as shown in Figure 3. Larger models consistently achieve higher accuracy than smaller ones, although the overall performance remained modest. This trend is more pronounced in models equipped with explicit reasoning capabilities, which show a more noticeable upward slope in performance with increasing size

than non-reasoning models, consistent with prior work showing that emergent abilities such as abstraction and multi-hop reasoning often surface at scale (Wei et al., 2022a). By comparing identical architectures with and without reasoning enabled, it becomes clear that reasoning mechanisms enhance the model’s ability to handle complex tasks like *Vinclat* (see examples in Appendix B).

Another interesting finding is that reasoner models often label hint keywords as “unknown”, suggesting a higher degree of calibration and a learned tendency to abstain when uncertain—an ability associated with more robust and reliable models (Kadavath et al., 2022). The only outlier to this trend is Qwen3-0.6B, but we suspect this may be due to its limited capabilities, as the model is relatively small in size.

**Impact of prompt format** The choice of prompt format affects model performance in systematic ways (see Figure 2). In general, models tend to achieve higher accuracy with the “Letter-count” format, which explicitly states the number of words and letters, than with the “Fill-in-the-blanks” (underscore-based) version. A likely reason for this is that the former format provides a more explicit constraint, reducing ambiguity and helping models focus on finding a semantically coherent solution of the correct structure. In contrast, the latter requires symbolic interpretation and spatial reasoning, which appears to challenge smaller or less specialized models.

Interestingly, reasoner models handle the “Fill-in-the-blanks” format better than non-reasoning models, suggesting they can internalize structural constraints more effectively. These results align with findings that structured or symbolic formats require planning-like behavior (Yao et al., 2023) and that standard models struggle when format interpretability is low (Wei et al., 2022b). This highlights the importance of prompt format in evaluating model capabilities and suggests that different prompting strategies may probe distinct aspects of model reasoning and language understanding.

**Impact of architecture** Mixture-of-Experts (MoE) models consistently outperformed their dense counterparts of comparable nominal size (see Figure 3). While MoE architectures dynamically route tokens through different expert sub-networks, their actual parameter count is larger in aggregate, offering greater representational capacity (Shazeer et al., 2017). The advantage was especially notable in reasoner MoEs, suggesting a possible synergy between structured reasoning components and sparse activation patterns. These findings support recent literature showing that MoEs can yield high performance on diverse tasks when well-routed and supported by instruction-following objectives (Du et al., 2022). Nevertheless, they also raise open questions about efficiency and cultural robustness, as the better performance of MoEs does not yet translate to human-like capabilities in culturally rich or open-ended inference tasks like *Vinclat*.

**Impact of prompt language** Prompt language had only a minor effect on performance (see Table 3 and Figure 2). Differences between English and Catalan versions were small compared to those driven by reasoning capabilities or prompt format. This suggests that once models understood the task structure, their performance depended more on internal reasoning mechanisms than on the surface language of the instructions. The minimal gap may also indicate that high-performing multilingual or instruction-tuned models can transfer comprehension of well-specified tasks across languages (Scao et al., 2023). However, this apparent robustness should be interpreted with caution. Since *Vinclat* itself is culturally rooted in Catalan, limited variation by prompt language may simply reflect the dominant role of cultural and associative reasoning over linguistic comprehension.

**Qualitative error analysis** The error typology revealed consistent patterns across architectures and prompting conditions. The most frequent errors were misinterpretation of hints and incorrect final solutions despite reasonable partial reasoning, each accounting for roughly 40–60% of observed cases. Linguistic or cultural misunderstandings such as mistranslating idiomatic expressions or missing references to Catalan media, literature, or humor were also highly prevalent, especially in non-reasoning and English-prompted runs.

All models exhibited systematic reasoning artifacts. In DeepSeek-R1, a primacy effect was observed when prompted in Catalan (the model fixating on its first hypothesis even when incorrect), whereas the English version showed a recency effect, tending to choose the last proposed answer. Models prompted in English also displayed a tendency to overthink, producing excessively long and convoluted chains of thought that ultimately led it away from the correct answer. Both DeepSeek versions suffered from semantic overgeneralization, proposing overly broad concepts (e.g., *autism* for *Asperger*) and frequent cultural hallucinations when literally translating figurative Catalan.

The Qwen3-A22B model family showed a different profile. While they displayed stronger structural understanding—accurately reproducing the letter and word constraints—they often produced linguistic hallucinations, fabricating Catalan words from Spanish or English cognates (e.g., *meixó* from *mechón*). Orthographic errors and culturally induced failures were common, particularly when interpreting tongue-twisters, idioms, or jokes well known in Catalan society. Interestingly, several “failed” hints were semantically valid alternatives rather than true errors (e.g., guessing *tabú* instead of *secret*), suggesting that some model reasoning chains were plausible but misaligned with the original puzzle key.

Across models, reasoner variants generally produced more logically coherent explanations, whereas non-reasoning versions were more erratic and hallucination-prone. However, even the best-performing setups struggled with cultural grounding, reinforcing the view that commonsense and culturally situated reasoning remain unsolved challenges for current LLMs.

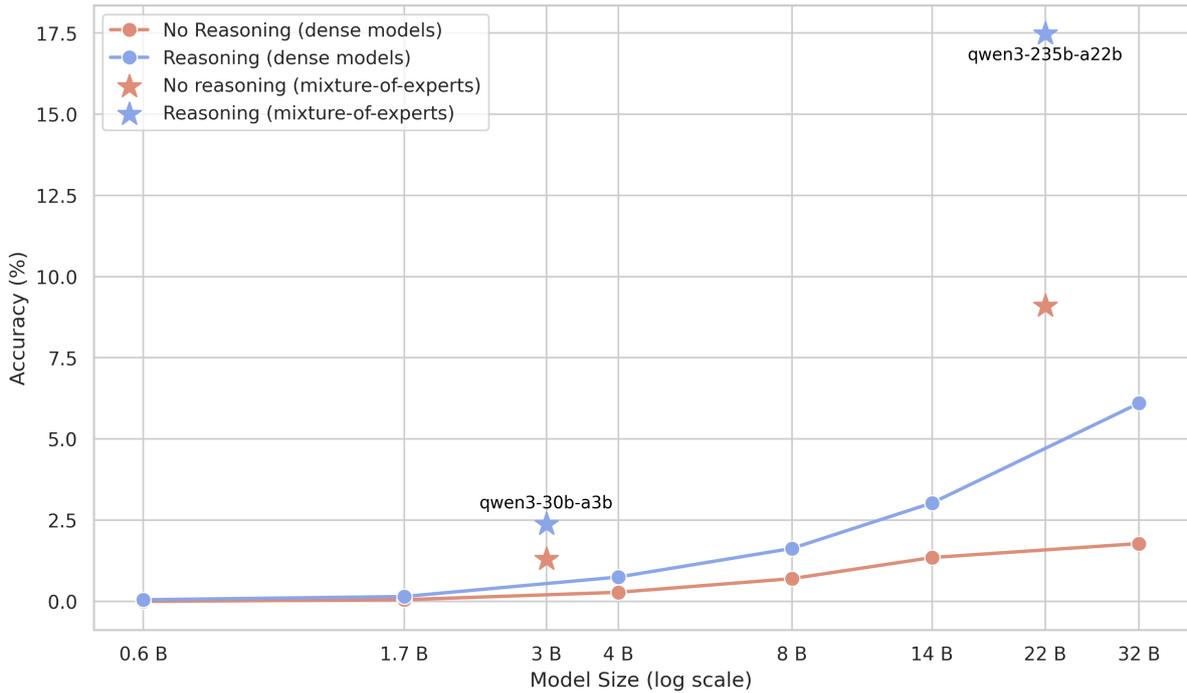


Figure 3: Accuracy scores based on model size and architecture for Qwen3 models.

## 6 Conclusion

This paper introduced a novel evaluation task for language models based on *Vinclat*, a Catalan word-association game that requires interpreting multiple hints and synthesizing them to retrieve a final solution under structural constraints. The resulting dataset of 1,000 high-quality instances—spanning domains such as culture, science, idioms, and history—offers a rich, culturally grounded benchmark for evaluating multi-hop reasoning, abstraction, and lexical-semantic association in LLMs.

Despite recent advances in large-scale language modeling, our results reveal that *Vinclat* remains a difficult challenge: even the strongest models tested achieved only modest accuracy, and performance fell significantly short of human baselines. The qualitative error analysis further showed that models frequently fail due to misinterpretation of clues, hallucinations, and a lack of cultural grounding, even when their reasoning structure appears coherent. Models also showed sensitivity to prompt format and language, with reasoners benefiting from clear structural prompts and showing less degradation when prompted in Catalan. Mixture-of-expert and reasoning-enabled models outperformed both their dense and non-reasoning counterparts, further supporting the need for explicit planning and compositional capabilities in solving complex tasks.

These findings point to clear limitations in current LLMs’ abilities to integrate indirect clues, deal with partial information, and apply structural constraints over multiple reasoning steps, all essential capabilities for nuanced language understanding. *Vinclat* offers a promising diagnostic tool for probing such skills, especially in multilingual and culturally diverse contexts.

Future work may explore whether task-specific training, retrieval augmentation, or fine-tuned reasoning strategies can close the performance gap. Incorporating targeted cultural or linguistic knowledge could also improve model grounding and interpretability. Finally, expanding the benchmark to other languages or game formats could help test generalization and transfer capabilities in multilingual and cross-cultural LLM evaluation.

## Limitations

**Cultural and linguistic specificity** *Vinclat* is deeply rooted in the Catalan language and culture. While this enriches its value as a culturally-grounded evaluation benchmark, it also poses challenges. Many hints rely on background knowledge that may not be well represented in LLM pretraining corpora, especially for mid-resource languages like Catalan. This makes it difficult to disentangle whether poor performance is due to lack of reasoning ability, insufficient cultural exposure, or both.

**Single-turn setup** The evaluation relied on a single-turn setting, which —although methodologically clean— does not reflect the interactive capabilities of modern LLMs. Multi-turn setups could allow models to revise initial guesses, ask for clarification, or reason iteratively. Especially for tasks involving ambiguity or multi-step reasoning, this restricted setup may underestimate true performance potential.

**Scope of model types** We focused on a representative set of open-weight models across different sizes and architectures. However, due to administrative and bureaucratic constraints, we did not include commercial, instruction-tuned models (e.g., GPT-4, Claude, Gemini), which may demonstrate stronger performance. In any case, the ultimate contribution of this work is the introduction of a new, complex dataset and to derive meaningful insights about the data, rather than to conduct an exhaustive model benchmarking exercise.

**Human baseline** While helpful as a reference point, the human baselines —especially the ones without internet access— may not fully reflect the capabilities of an average native speaker under realistic conditions. Although participants were trained on the task, they were not expert players and may have lacked familiarity with the types of strategies often needed to solve *Vinclat* puzzles. The internet-assisted baseline depends on the search strategies and tools used, which introduces variability. As such, these baselines are indicative rather than definitive upper bounds.

## Acknowledgments

We sincerely thank Quim Roura, Alba Ferran and Blai Samitier for their creativity and dedication in crafting the riddles included in the dataset. We also extend our gratitude to Carlota Alba Foz for her meticulous review and corrections, and to BSC’s Annotation Team for their role in establishing the human performance baseline.

This work was promoted and financed by *Generalitat de Catalunya* through the Aina project.

## 7 Bibliographical References

### References

- Alan Agresti and Brent A. Coull. 1998. [Approximate is better than "exact" for interval estimation of binomial proportions](#). *The American Statistician*, 52(2):119–126.
- Frederic C. Bartlett. 1932. *Remembering: A study in experimental and social psychology*. Cambridge University Press.
- Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Scott Wen tau Yih, and Yejin Choi. 2020. [Abductive commonsense reasoning](#). *Preprint*, arXiv:1908.05739.
- Ellen Bialystok. 1986. Factors in the growth of linguistic awareness. *Child development*, pages 498–510.
- Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrike, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. [Sparks of artificial general intelligence: Early experiments with gpt-4](#). *Preprint*, arXiv:2303.12712.
- Peter Clark, Isaac Cowhey, Oren Etzioni, and 1 others. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. [Training verifiers to solve math word problems](#). *Preprint*, arXiv:2110.14168.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *NAACL-HLT*.
- Nan Du, Yanping Huang, Andrew M. Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, Barret Zoph, Liam Fedus, Maarten Bosma, Zongwei Zhou, Tao Wang, Yu Emma Wang, Kellie Webster, Marie Pellat, Kevin Robinson, and 8 others. 2022. [Glam: Efficient scaling of language models with mixture-of-experts](#). *Preprint*, arXiv:2112.06905.
- Mor Geva, Daniel Khashabi, Wen-tau Chen, and Dan Roth. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics*.

- Akash Ghosh, Debayan Datta, Sriparna Saha, and Chirag Agarwal. 2025. The multilingual mind: A survey of multilingual reasoning in language models. *arXiv preprint arXiv:2502.09457*.
- Elliot Glazer, Ege Erdil, Tamay Besiroglu, Diego Chicharro, Evan Chen, Alex Gunning, Caroline Falkman Olsson, Jean-Stanislas Denain, Anson Ho, Emily de Oliveira Santos, and 1 others. 2024. Frontiermath: A benchmark for evaluating advanced mathematical reasoning in ai. *arXiv preprint arXiv:2411.04872*.
- Alex Gu, Baptiste Rozière, Hugh Leather, Armando Solar-Lezama, Gabriel Synnaeve, and Sida I Wang. 2024. Cruxeval: A benchmark for code reasoning, understanding and execution. *arXiv preprint arXiv:2401.03065*.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shitong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.
- Dan Hendrycks, Steven Burns, Sam Basart, Collin Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.
- Chengpeng Hu, Yunlong Zhao, Ziqi Wang, Haocheng Du, and Jialin Liu. 2024. Games for artificial intelligence research: A review and perspectives. *IEEE Transactions on Artificial Intelligence*.
- Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, and 17 others. 2022. [Language models \(mostly\) know what they know](#). *Preprint*, arXiv:2207.05221.
- David Klahr and Kevin Dunbar. 1988. The dual space search hypothesis: An alternative to the problem space hypothesis for scientific reasoning. In *Proceedings of the 10th Annual Conference of the Cognitive Science Society*.
- David R. Krathwohl. 2002. [A revision of Bloom’s Taxonomy: An overview](#). *Theory Into Practice*, 41(4):212–218.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. [Efficient memory management for large language model serving with pagedattention](#). *Preprint*, arXiv:2309.06180.
- Willem J.M. Levelt. 1989. *Speaking: From intention to articulation*. MIT Press.
- Jiawei Li, Yizhe Yang, Yu Bai, Xiaofeng Zhou, Yinghao Li, Huashan Sun, Yuhang Liu, Xingpeng Si, Yuhao Ye, Yixiao Wu, , Bin Xu, Ren Bowen, Chong Feng, Yang Gao, and Heyan Huang. 2024. [Fundamental capabilities of large language models and their applications in domain scenarios: A survey](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11116–11141, Bangkok, Thailand. Association for Computational Linguistics.
- Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, and 31 others. 2023. [Holistic evaluation of language models](#). *Preprint*, arXiv:2211.09110.
- Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.
- Angel Yahir Loredó Lopez, Tyler McDonald, and Ali Emami. 2025. [NYT-connections: A deceptively simple text classification task that stumps system-1 thinkers](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1952–1963, Abu Dhabi, UAE. Association for Computational Linguistics.
- Nikolai Lörer, Norman Lee, and Douwe Kiela. 2023. Conceptarc: Evaluating abstract reasoning in language models with the abstraction and reasoning corpus. *arXiv preprint arXiv:2306.01774*.
- Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. [Can a suit of armor conduct electricity? a new dataset for open book question answering](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium. Association for Computational Linguistics.
- Swaroop Mishra, Matthew Finlayson, Pan Lu, Leonard Tang, Sean Welleck, Chitta Baral, Tanmay Rajpurohit, Oyvind Tafjord, Ashish Sabharwal, Peter Clark, and 1 others. 2022. Lila: A unified benchmark for mathematical reasoning. *arXiv preprint arXiv:2210.17517*.
- Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. 2021. [KILT: a benchmark for knowledge intensive language tasks](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2523–2544, Online. Association for Computational Linguistics.

- Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*.
- Teven Le Scao, Angela Fan, Christopher Akiki, Elie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, and 373 others. 2023. [Bloom: A 176b-parameter open-access multilingual language model](#). *Preprint*, arXiv:2211.05100.
- Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. [Outrageously large neural networks: The sparsely-gated mixture-of-experts layer](#). *Preprint*, arXiv:1701.06538.
- Parshin Shojaee, Iman Mirzadeh, Keivan Alizadeh, Maxwell Horton, Samy Bengio, and Mehrdad Farajtabar. 2025. The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity. *arXiv preprint arXiv:2506.06941*.
- David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, and 1 others. 2016. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489.
- David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, and 1 others. 2017. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359.
- Koustuv Sinha, Shagun Sodhani, and 1 others. 2019. Clutr: A diagnostic benchmark for inductive reasoning from text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*.
- Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, and 432 others. 2023. [Beyond the imitation game: Quantifying and extrapolating the capabilities of language models](#). *Preprint*, arXiv:2206.04615.
- Aarohi Srivastava and 1 others. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.
- Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, and 1 others. 2019. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *nature*, 575(7782):350–354.
- Alex Wang, Yada Pruksachatkun, Nikita Nangia, and 1 others. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*.
- Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022a. [Emergent abilities of large language models](#). *Preprint*, arXiv:2206.07682.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022b. [Chain-of-thought prompting elicits reasoning in large language models](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.
- Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. [Tree of thoughts: Deliberate problem solving with large language models](#). *Preprint*, arXiv:2305.10601.
- Xingyu Zheng, Yuye Li, Haoran Chu, Yue Feng, Xudong Ma, Jie Luo, Jinyang Guo, Haotong Qin, Michele Magno, and Xianglong Liu. 2025. An empirical study of qwen3 quantization. *arXiv preprint arXiv:2505.02214*.
- Chunting Zhou, Pengfei Liu, Puxin Xu, Srinu Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. [Lima: Less is more for alignment](#). *Preprint*, arXiv:2305.11206.
- Terry Yue Zhuo, Minh Chien Vu, Jenny Chim, Han Hu, Wenhao Yu, Ratnadira Widayarsi, Imam Nur Bani Yusuf, Haolan Zhan, Junda He, Indraneil Paul, and 1 others. 2024. Bigcodebench: Benchmarking code generation with diverse function calls and complex instructions. *arXiv preprint arXiv:2406.15877*.

## A Prompt formats

### English prompt with fill-in-the-blanks format (EN-FB)

Let's play a game in Catalan! Your goal is to find a "solution" word or words of a specific length, which will be given to you. You will also receive four numbered hints. Try to solve each one to get a "hint word", which doesn't need to match the solution's length. Then, try to think about the common theme or connection between the "hint words" that you found. The final solution should fit the required letter count and is related to the "hint words" you identified.

It is important to note that you don't necessary need all the "hint words" to get to the final solution. If you're struggling with a hint or suspect your guess might be wrong, it's often better to focus on the "hint words" you are sure about. A wrong one can send you down the wrong path!

Here are your hints:

1. {hint\_1}
2. {hint\_2}
3. {hint\_3}
4. {hint\_4}

The "solution" should fit here: {underscores}. What's your guess? Return a JSON object with the following fields: *'hint\_word\_1'*, *'hint\_word\_2'*, *'hint\_word\_3'*, *'hint\_word\_4'*, *'solution'*. If you could not find the word associated to some hint, simply keep that field as *'unknown'*.

### Catalan prompt with fill-in-the-blanks format (CA-FB)

Et proposo jugar a un joc en català! El teu objectiu és trobar una paraula o paraules "solució" d'una longitud específica, que et serà donada. També rebràs quatre pistes numerades. Intenta resoldre cadascuna per obtenir una "paraula pista", que no té perquè tenir la mateixa longitud que la solució. Després, intenta pensar en el tema comú o la connexió entre les "paraules pista" que has trobat. La solució final ha de tenir la longitud requerida i estar relacionada amb les "paraules pista" que hakis identificat.

És important tenir en compte que no necessàriament necessites totes les "paraules pista" per arribar a la solució final. Si tens dificultats amb una pista o sospites que la teva conjectura pot ser incorrecta, sovint és millor centrar-se en les "paraules pista" de les quals estàs segur. Una resposta incorrecta pot ser contraproduent, ja que et farà anar pel camí equivocat!

Aquí tens les teves pistes:

1. {hint\_1}
2. {hint\_2}
3. {hint\_3}
4. {hint\_4}

La solució hauria d'encaixar aquí: {underscores}. Quina diries que és la resposta correcta? Retorna un objecte JSON amb els següents camps: *'hint\_word\_1'*, *'hint\_word\_2'*, *'hint\_word\_3'*, *'hint\_word\_4'*, *'solution'*. Si no has pogut trobar la paraula associada a alguna pista, simplement deixa aquest camp com a *'unknown'*.

### English prompt with letter-count format (EN-LC)

Let's play a game in Catalan! Your goal is to find a "solution" word or words of a specific length, which will be given to you. You will also receive four numbered hints. Try to solve each one to get a "hint word", which doesn't need to match the solution's length. Then, try to think about the common theme or connection between the "hint words" that you found. The final solution should fit the required letter count and is related to the "hint words" you identified.

It is important to note that you don't necessary need all the "hint words" to get to the final solution. If you're struggling with a hint or suspect your guess might be wrong, it's often better to focus on the "hint words" you are sure about. A wrong one can send you down the wrong path!

Here are your hints:

1. {hint\_1}
2. {hint\_2}
3. {hint\_3}
4. {hint\_4}

The "solution" has {letter\_count}. What's your guess? Return a JSON object with the following fields: 'hint\_word\_1', 'hint\_word\_2', 'hint\_word\_3', 'hint\_word\_4', 'solution'. If you could not find the word associated to some hint, simply keep that field as 'unknown'.

### Catalan prompt with letter-count format (CA-LC)

Et proposo jugar a un joc en català! El teu objectiu és trobar una paraula o paraules "solució" d'una longitud específica, que et serà donada. També rebràs quatre pistes numerades. Intenta resoldre cadascuna per obtenir una "paraula pista", que no té perquè tenir la mateixa longitud que la solució. Després, intenta pensar en el tema comú o la connexió entre les "paraules pista" que has trobat. La solució final ha de tenir la longitud requerida i estar relacionada amb les "paraules pista" que hakis identificat.

És important tenir en compte que no necessàriament necessites totes les "paraules pista" per arribar a la solució final. Si tens dificultats amb una pista o sospites que la teva conjectura pot ser incorrecta, sovint és millor centrar-se en les "paraules pista" de les quals estàs segur. Una resposta incorrecta pot ser contraproductiu, ja que et farà anar pel camí equivocat!

Aquí tens les teves pistes:

1. {hint\_1}
2. {hint\_2}
3. {hint\_3}
4. {hint\_4}

La "solució" té {letter\_count}. Quina diries que és la resposta correcta? Retorna un objecte JSON amb els següents camps: 'hint\_word\_1', 'hint\_word\_2', 'hint\_word\_3', 'hint\_word\_4', 'solution'. Si no has pogut trobar la paraula associada a alguna pista, simplement deixa aquest camp com a 'unknown'.

## B Selected instances

**Problem**

**Hint 1:** Pot ser de vàter, cuina, diari, seda...  
**Hint 2:** El cinema n'és el setè  
**Hint 3:** Els trencaclosques et fan unir-ne moltes  
**Hint 4:** El que pots fer si prems Ctrl+V

**Solution**

**Keyword 1:** Paper  
**Keyword 2:** Art  
**Keyword 3:** Peces  
**Keyword 4:** Enganxar  
**Final solution:** Collage

**LLM's reasoning trace**

Okay, let's tackle this Catalan puzzle. I need to find the solution word based on the four hints provided. The solution has a specific length, but the hint words don't necessarily need to match that. Let me go through each hint one by one.

- \*\*Pot ser de vàter, cuina, diari, seda...\*\*** The first hint mentions things that can be "de vàter" (toilet), "cuina" (kitchen), "diari" (daily), "seda" (silk). I'm thinking these are types of paper. Toilet paper, kitchen paper, daily paper (newspaper), silk paper. So the hint word here might be "paper".
- \*\*El cinema n'és el setè\*\*** The cinema is the seventh. Hmm, maybe referring to the "seventh art". Cinema is often called the seventh art. In Catalan, that's "el setè art". So the hint word here is "art".
- \*\*Els trencaclosques et fan unir-ne moltes\*\*** Puzzles (trencaclosques) make you put together many... Pieces? Parts? The Catalan word for pieces is "peces" or "peces de trencaclosques". So the hint word might be "peces".
- \*\*El que pots fer si prems Ctrl+V\*\*** If you press Ctrl+V, that's the paste function. In Catalan, "enganxar" means to paste. So the hint word could be "enganxar".

Now, looking at the hint words: paper, art, peces, enganxar. What's the common theme? Maybe combining these concepts. Paper is a material, art is creative, pieces are parts, and paste is sticking them together. Maybe the solution is related to collage. A collage involves pasting pieces of paper or other materials to create art. The Catalan word for collage is "collage" (same as English) but let me check the length. The solution blank is 7 letters. "Collage" has 7 letters. So that fits. Therefore, the solution should be "collage".

Figure 4: Response provided by DeepSeek-R1 to the instance with ID 453.

### Problem

**Hint 1:** Hi ha qui en té als cacauets, als gats o a la pols

**Hint 2:** No te'ls donen sense recepta

**Hint 3:** L'arrosseguen les abelles de flor en flor

**Hint 4:** Estació que comença la segona meitat de març

### Solution

**Keyword 1:** Al·lèrgia

**Keyword 2:** Medicaments

**Keyword 3:** Pol·len

**Keyword 4:** Primavera

**Final solution:** Antihistamínic

### LLM's reasoning trace

Okay, let's try to figure this out step by step. The goal is to find a solution word that's 14 letters long, based on four clues. Each clue gives a "hint word," and the solution is related to all of them. Let's tackle each clue one by one.

First clue: "Hi ha qui en té als cacauets, als gats o a la pols." Translating that from Catalan, it's talking about something you can find in peanuts, cats, or dust. Hmm. Common things in those... Allergies? Because people can be allergic to peanuts, cats (pet dander), and dust mites. So maybe "al·lèrgia" (allergy) is the hint word here. Let me check the length. "Al·lèrgia" has 8 letters, but maybe in plural? Not sure yet.

Second clue: "No te'ls donen sense recepta." Which translates to "They don't give them to you without a prescription." So this is likely referring to medications, like pills or drugs. The Catalan word for pills is "píndoles" or "medicaments." But "medicaments" is longer. Wait, "píndoles" is 8 letters. But maybe "recepta" itself? No, the answer here is probably "píndoles" since they require a prescription.

Third clue: "L'arrosseguen les abelles de flor en flor." Bees drag this from flower to flower. Bees collect pollen, so "pol·len" in Catalan. That's 6 letters. That seems right. Pollen is what they carry.

Fourth clue: "Estació que comença la segona meitat de març." The season starting in the second half of March. In the Northern Hemisphere, spring starts around March 20th or 21st. So "primavera" (spring) in Catalan. That's 9 letters.

Now, the solution needs to be 14 letters long and related to all these hints. Let's look at the hint words again: allergy, pills, pollen, spring. How do these connect? Spring is when pollen is high, causing allergies. People take pills (antihistamines) for allergies. So the common theme is allergies. The solution might be a word related to allergies or antihistamines. But a 14-letter Catalan word... Let's think. "Antihistamínic" (antihistamine) has 14 letters. Let's check: A-N-T-I-H-I-S-T-A-M-Í-N-I-C. Yes, 14 letters. That makes sense. The hint words are all related to allergies: allergens (peanuts, cats, dust), medication (pills), pollen (from flowers), and spring (when pollen is high). So the solution is "antihistamínic".

Figure 5: Response provided by Qwen3-235B-A22B to the instance with ID 965.

### Problem

**Hint 1:** Per molts l'homeopatia no ho és

**Hint 2:** La tercera pota que et falta d'un tamboret si ja tens els diners i l'amor

**Hint 3:** El percebem amb l'oïda

**Hint 4:** Sentir atentament

### Solution

**Keyword 1:** Medicina

**Keyword 2:** Salut

**Keyword 3:** So

**Keyword 4:** Escoltar

**Final solution:** Estetoscopi

### LLM's reasoning trace

Okay, let's tackle this puzzle step by step. The user wants me to find a solution word of a specific length (which isn't given here, but maybe the solution line has a blank that indicates the length) based on four clues. Each clue should lead to a "hint word," and the solution connects these hints.

Starting with clue 1: "Per molts l'homeopatia no ho és." Homeopathy is often considered a pseudoscience, so maybe the hint word here is "ciència" (science) because some people don't think homeopathy is scientific.

Clue 2: "La tercera pota que et falta d'un tamboret si ja tens els diners i l'amor." There's a saying that a stool needs three legs: "diner, amor, salut" (money, love, health). If they already have money and love, the missing third leg would be "salut" (health).

Clue 3: "El percebem amb l'oïda." That which we perceive with hearing is "so" (sound) or "música" (music). Maybe "so" is the hint here.

Clue 4: "Sentir atentament." To listen attentively could be "escoltar" (to listen) or "atenció" (attention). But "escoltar" is more directly related to listening carefully.

Now, connecting the hint words: ciència, salut, so, escoltar. The common theme might be something that combines science, health, sound, and listening. Maybe "salut" relates to health, "ciència" to medicine or biology, "so" to sound waves, and "escoltar" to using a stethoscope. The solution could be "estetoscopi" (stethoscope), which is a medical instrument used to listen to body sounds, related to health and science. Let's check the length: "estetoscopi" is 11 letters. If the blank is for an 11-letter word, that fits. Alternatively, maybe "fonendoscopi" which is another term, but less common. But "estetoscopi" is more widely known. So I think that's the solution.

Figure 6: Response provided by DeepSeek-R1 to the instance with ID 783.

### Problem

**Hint 1:** El 10 de novembre celebrem el dia mundial de la...

**Hint 2:** El paper de Harrison Ford amb un fuet

**Hint 3:** Estar ben mort i \_\_\_\_\_

**Hint 4:** L'explicació del passat

### Solution

**Keyword 1:** Ciència

**Keyword 2:** Indiana Jones

**Keyword 3:** Enterrat

**Keyword 4:** Història

**Final solution:** Arqueologia

### LLM's reasoning trace

Okay, let's tackle this step by step. I need to find four hint words based on the clues provided and then figure out the solution word that's 11 letters long and connected to those hints.

Starting with hint 1: "El 10 de novembre celebrem el dia mundial de la..." November 10th is World Day of something. In Catalan, "dia mundial de la..." probably refers to a common world day. Let me think. World Day of what? Maybe "Consciència" (Consciousness) or "Creativitat" (Creativity)? Wait, but the most common one I know is World Science Day on November 10th. So "Ciència" (Science) in Catalan. So hint\_word\_1 could be "Ciència".

Hint 2: "El paper de Harrison Ford amb un fuet" (The role of Harrison Ford with a whip). Harrison Ford played Indiana Jones, who uses a whip. So the character is Indiana Jones, but in Catalan, maybe "Indiana Jones" or "Indi Jones"? Or perhaps "Arqueòleg" (archaeologist), but the main role with a whip is definitely Indiana Jones. So hint\_word\_2 is "Indiana Jones".

Hint 3: "Estar ben mort i \_\_\_\_\_" (To be very dead and \_\_\_\_\_). This sounds like a phrase. Maybe "estar ben mort i enterrat" (be very dead and buried)? "Enterrat" is buried. But the blank might be a single word. Alternatively, maybe "estar ben mort i sencer" (dead and whole?), but that doesn't make sense. Wait, common expressions: "Ben mort i enterrat" makes sense. So the hint\_word\_3 could be "enterrat".

Hint 4: "L'explicació del passat" (The explanation of the past). That could be "història" (history) because history explains the past. So hint\_word\_4 is "Història".

Now the solution is an 11-letter word related to these hints: Ciència (Science), Indiana Jones, Enterrat (Buried), Història (History). Let me think of the connections. Indiana Jones is an archaeologist, which relates to history and exploring the past. "Enterrat" refers to being buried, which could relate to archaeological finds. Science and history are both involved in studying the past. The common theme might be something like "Arqueologia" (Archaeology). Let's check the letters: A-R-Q-U-E-O-L-O-G-I-A. That's 11 letters. Yes! So the solution is "Arqueologia".

Figure 7: Response provided by Qwen3-235B-A22B to the instance with ID 876.

# Conceptual Cultural Index: A Metric for Cultural Specificity via Relative Generality

**Takumi Ohashi**

Hosei University, Tokyo, Japan  
takumi.ohashi.4g@gmail.com

**Hitoshi Iyatomi**

Hosei University, Tokyo, Japan  
iyatomi@hosei.ac.jp

## Abstract

Large language models (LLMs) are increasingly deployed in multicultural settings; however, systematic evaluation of cultural specificity at the sentence level remains underexplored. We propose the Conceptual Cultural Index (CCI), which estimates cultural specificity at the sentence level. CCI is defined as the difference between the generality estimate within the target culture and the average generality estimate across other cultures. This formulation enables users to operationally control the scope of culture via comparison settings and provides interpretability, since the score derives from the underlying generality estimates. We validate CCI on 400 sentences (200 culture-specific and 200 general), and the resulting score distribution exhibits the anticipated pattern: higher for culture-specific sentences and lower for general ones. For binary separability, CCI outperforms direct LLM scoring, yielding more than a 10-point improvement in AUC for models specialized to the target culture. Our code is available at <https://github.com/IyatomiLab/CCI>.

## 1 Introduction

Large language models (LLMs) exhibit broad multilingual competence and are increasingly used for tasks such as search, summarization, and dialogue (Brown et al., 2020; Achiam et al., 2023; Grattafiori et al., 2024; Yang et al., 2024). However, as applications expand, a key challenge remains in determining whether it is possible to ensure consistent and culturally aware responses across different regions. Everyday knowledge, such as dietary practices, greeting conventions, linguistic expressions, and seasonal events, varies systematically across cultures and regions, and how models handle this knowledge has direct implications for fairness, safety, and reliability (Cao et al., 2023; Naous et al., 2024; Shen et al., 2024; Rao et al., 2025). Consequently, there is a need to develop models that can appropriately accommodate diverse cultural

characteristics and differences, as well as models specialized for a specific culture.

Benchmarks of LLM capabilities have progressed beyond general-knowledge evaluations toward frameworks that focus on cultural knowledge, which advance the visualization of regional differences and biases (Myung et al., 2024; Chiu et al., 2025). Many benchmarks use QA formats with overall accuracy as the main metric, consistently showing that culturally specific questions are harder than culture-agnostic ones (Shen et al., 2024; Arora et al., 2025). However, cultural knowledge spans a continuum—from phenomena shared across regions to those unique to a specific locale—and existing benchmarks fail to capture this aspect, making it difficult to conduct error analysis and formulate targeted improvement strategies.

On the data side, large-scale corpora of cultural knowledge have been proposed (Nguyen et al., 2023; Shi et al., 2024), but they lack annotations indicating the degree of cultural specificity of each sentence. Thus, both evaluation and data resource development require a framework for quantitatively assessing sentence-level cultural specificity, yet no such framework currently exists. Manual annotation of sentence-level cultural specificity is labor-intensive, requires domain expertise and contextual understanding, and often yields low inter-annotator agreement, highlighting the need for automation. However, culture is a multifaceted, high-level construct, and prior work has rarely provided an explicit definition (Adilazuarda et al., 2024).

In this paper, we propose the Conceptual Cultural Index (CCI), a sentence-level metric for quantifying cultural specificity, to address this challenge. CCI uses an LLM to estimate a sentence’s generality across multiple cultures and, based on these scores, quantifies the target culture’s specificity relative to others. This formulation allows users to control the scope of “culture” by adjusting the set of non-target cultures used for comparison.

The contributions of this study are as follows:

- We introduce CCI, a new sentence-level metric for quantifying cultural specificity.
- Compared with direct LLM-based scoring of cultural specificity, CCI yields clearer separability between culture-specific and general sentences and offers greater interpretability.
- We present a practical use case of CCI by assigning item-level CCI scores to existing benchmarks and showing that model performance varies with the level of cultural specificity, enabling culture-aware error analysis.

## 2 Related Work

Cultural evaluation benchmarks for language models include broad, multi-region datasets such as GeoMLAMA (Yin et al., 2022), BLEnD (Myung et al., 2024), CDEval (Wang et al., 2024), and CulturalBench (Chiu et al., 2025), as well as country- or region-specific benchmarks such as CLiCK (Kim et al., 2024), IndoCulture (Koto et al., 2024), and CHARM (Sun et al., 2024). These resources support comparisons of cultural knowledge across models, but most adopt a QA format and rely on overall accuracy, which conflates culture-specific difficulty with general knowledge errors.

Text-based resources for collecting cultural knowledge, including StereoKG (Deshpande et al., 2022), CANDLE (Nguyen et al., 2023), CultureAtlas (Fung et al., 2024), CultureBank (Shi et al., 2024), and MANGO (Nguyen et al., 2024), extract and organize cultural assertions at scale. However, they do not provide sentence-level annotations with continuous scores of cultural specificity, leaving a gap that our sentence-level metric CCI aims to fill.

## 3 CCI

We propose the Conceptual Cultural Index (CCI), a sentence-level index of cultural specificity. As illustrated in Figure 1, given a target culture and a set of comparison cultures, we use an LLM to estimate how common a sentence is in each culture and derive a specificity score for the target culture.

### 3.1 Obtaining Generality Scores

Given an input sentence  $x$ , a set of cultures  $C$ , and a target culture  $t \in C$ , we use an LLM<sup>1</sup> to estimate, for each  $c \in C$ , how common  $x$  is in the

<sup>1</sup>Model selection is discussed in the experiments section.

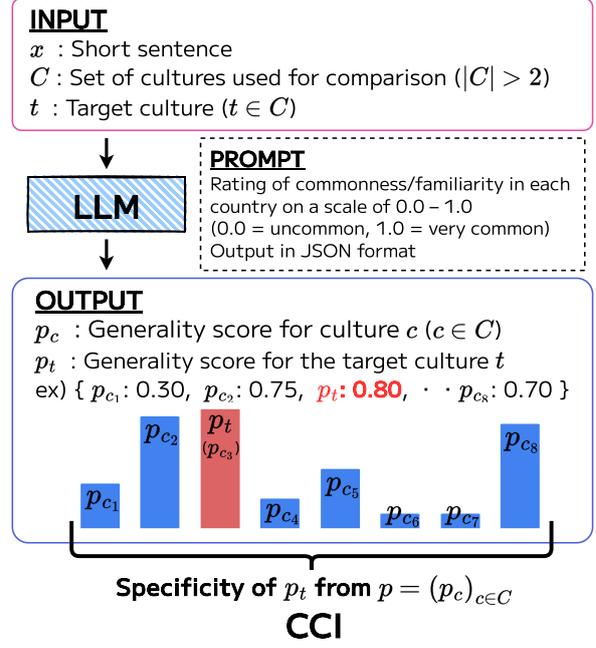


Figure 1: Overview of CCI.

culture  $c$ , yielding a continuous generality score  $p_c(x) \in [0, 1]$ . In practice, we query all cultures in  $C$  within a single prompt and parse the scores from a JSON-formatted response. The prompt used for the generality score is provided in Appendix A.

To mitigate run-to-run variability in LLM outputs, we average results over  $N$  independent runs (in this paper,  $N = 3$ ):

$$\bar{p}_c(x) = \frac{1}{N} \sum_{n=1}^N f_{\text{LLM}}^{(n)}(x; C)[c], \quad c \in C. \quad (1)$$

Here,  $f_{\text{LLM}}^{(n)}(x; C)[c]$  denotes the score for culture  $c$  returned by the  $n$ -th run.

### 3.2 Definition of CCI

For a target culture  $t \in C$ , we define CCI as the difference between the generality score in the target culture and the average generality score across the other cultures:

$$CCI(x; t, C) = \bar{p}_t(x) - \frac{1}{|C| - 1} \sum_{c \in C \setminus \{t\}} \bar{p}_c(x). \quad (2)$$

CCI takes values in  $[-1, 1]$ : values near 0 indicate that  $x$  is cross-culturally general, values near 1 indicate that  $x$  is specific to the target culture, and values near  $-1$  indicate that  $x$  is specific to non-target cultures.

For Eq. (2), we also examined a sharpness-based formulation that weights the target culture

by its log-softmax-normalized generality<sup>2</sup>, but it produced nearly constant scores across input sentences and compressed the score range, especially for larger  $|C|$ . In contrast, the simple difference is less sensitive to  $|C|$  and directly measures the gap on the original  $[0, 1]$  scale, so we adopt this definition.

## 4 Experiments

### 4.1 Experimental Setup

To assess whether CCI reflects cultural specificity, we use Japan as the target culture  $t$  and compute CCI for two classes: culture-specific sentences (positive) and general sentences (negative). We plot ROC curves for detecting Japanese cultural sentences and evaluate separability using the area under the ROC curve (AUC) and the difference in class medians. We also conduct a qualitative analysis of representative examples.

As a baseline, we use an LLM that directly outputs a  $[0, 1]$  specificity score and compute AUC with the same protocol to compare its discriminative performance with CCI. As with CCI, 0 denotes cross-cultural generality and 1 denotes specificity to the target culture. The prompt used for the baseline is provided in Appendix A.

**Data** We first used GPT-5 (model as of August 7, 2025) to generate 300 short Japanese cultural sentences and 300 general sentences. All generated sentences were manually reviewed and filtered, with duplicates and clear misclassifications removed, although some borderline cases may remain because cultural boundaries are inherently ambiguous. The final evaluation set consists of 200 Japanese cultural sentences and 200 general sentences. The prompts used for data generation are provided in Appendix A.

**Models** To identify suitable LLMs for computing CCI, we compared CCI and the baseline under a common protocol across five LLMs: multilingual models (Llama 3.1 (Grattafiori et al., 2024), Qwen 2.5 (Yang et al., 2024), gpt-oss (Agarwal et al., 2025)) and Japanese-specialized models (Llama 3.1 Swallow (Fujii et al., 2024), llm-jp 3.1 (Aizawa et al., 2024)). The exact model identifiers and links are listed in Appendix B.

<sup>2</sup> $q_t = \frac{\exp(\bar{p}_t)}{\sum_{c \in C} \exp(\bar{p}_c)}, CCI_{\log} = \left(1 + \frac{\log q_t}{\log(|C|)}\right) \bar{p}_t.$

### 4.2 Varying the Set of Cultures $C$

CCI allows the set  $C$  of cultures to vary. To assess how controllable the cultural scope is, we conduct experiments under two modes.

**Global mode**  $C$  is fixed to the 19 G20 member countries, excluding the European Union and the African Union, as we restrict  $C$  to country names.

**Custom mode**  $C$  is configured to match the task objective. In this experiment, to test whether the inclusion of neighboring cultures can be controlled, we defined two conditions across four countries:

1. **+Neighbor Culture**, which includes neighboring countries in  $C$ : [“China”, “Republic of Korea”, “United States of America”, “Japan”];
2. **-Neighbor Culture**, which excludes neighboring countries from  $C$ : [“Brazil”, “France”, “United States of America”, “Japan”];

For the baseline, we also evaluate two prompting conditions: one prompt that explicitly includes the instruction “*If the practice is also common in neighboring or culturally adjacent countries, do not consider it specific to the target.*” (**+Neighbor Culture**), and another that omits this instruction.

### 4.3 CCI-Based Benchmark Stratification

As a use case of CCI, we perform benchmark stratification by assigning CCI scores to each item and analyzing how task accuracy varies across CCI levels. We use two datasets that capture Japanese commonsense: JCommonsenseQA (JCQA) (Kurihara et al., 2022)<sup>3</sup> and JCommonsenseMorality (JCM) (Takeshita and Rzepka, 2025)<sup>4</sup>, both of which include items that may reflect phenomena specific to the Japanese cultural sphere. JCQA is a five-way multiple-choice commonsense question answering task; we compute CCI using as input  $x$  the concatenation of the question text and the gold option text. JCM is a binary classification task that judges whether an action is morally acceptable; we compute CCI using as input  $x$  the target sentence together with the gold label. For scoring, we use CCI in Global mode computed with gpt-oss. We evaluate on the JCQA dev set (1,119 items) and the JCM test set (3,992 items) using predictions from Qwen 2.5, Llama 3.1, and llm-jp 3.1.

<sup>3</sup><https://github.com/yahoojapan/JGLUE>

<sup>4</sup><https://github.com/Language-Media-Lab/jethics>

Models	Baseline		CCI (Global)	CCI (Custom)	
	+Neighbor			+Neighbor	-Neighbor
	$(C_{\text{median}} \uparrow, G_{\text{median}} \downarrow)$				
Qwen2.5-7B	(0.815, 0.800)	(0.980, 0.800)	(0.800, 0.505)	(0.633, 0.267)	(0.833, 0.467)
Llama-3.1-8B	(0.870, 0.800)	(0.870, 0.800)	(0.778, 0.648)	(0.664, 0.283)	(0.980, 0.711)
Llama-3.1-Swallow-8B	(0.950, 0.850)	(0.950, 0.850)	(0.761, 0.324)	(0.331, 0.117)	(0.933, 0.300)
llm-jp-3.1-13b	(0.800, 0.785)	(0.800, 0.700)	(0.869, 0.568)	(0.792, 0.467)	(0.897, 0.593)
gpt-oss-20b	(0.880, 0.100)	(0.775, 0.100)	(0.836, 0.063)	(0.697, 0.111)	(0.817, 0.104)

Table 1: Class-wise medians of specificity scores for CCI and the baseline.  $C_{\text{median}}$  denotes the median score for culture-specific sentences, and  $G_{\text{median}}$  denotes the median score for general sentences. While  $C_{\text{median}}$  should ideally be close to 1 and  $G_{\text{median}}$  close to 0, it is not necessary for every instance to reach these extremes; it suffices that the overall trend is observed. The cultural specificity score aims to assign context-appropriate values to each sentence.

Models	Baseline AUC / $\Delta$	CCI (Global) AUC / $\Delta$
Qwen2.5-7B	0.816 / 0.015	<b>0.884 / 0.295</b>
Llama-3.1-8B	<b>0.803</b> / 0.070	0.796 / <b>0.130</b>
Llama-3.1-Swallow-8B	0.842 / 0.100	<b>0.945 / 0.437</b>
llm-jp-3.1-13b	0.768 / 0.015	<b>0.908 / 0.301</b>
gpt-oss-20b	<b>0.963 / 0.780</b>	0.956 / 0.773

Table 2: Separability between culture-specific and general sentences, reported in terms of the AUC and the median gap  $\Delta = C_{\text{median}} - G_{\text{median}}$ .

## 5 Results and Discussion

### 5.1 Separability between Culture-Specific and General Sentences

Table 1 shows, for each LLM, the class-wise median specificity scores for CCI and the direct-estimation baseline:  $C_{\text{median}}$  for culture-specific sentences and  $G_{\text{median}}$  for general sentences. Table 2 shows separability, measured by AUC, and the gap between the medians ( $\Delta = C_{\text{median}} - G_{\text{median}}$ ). We report median gaps rather than mean gaps because the baseline scores are bounded in  $[0, 1]$ , whereas CCI scores lie in  $[-1, 1]$ . Given these different ranges, mean differences could inadvertently favor CCI; using medians avoids this issue in our experiments.

**CCI vs. Baseline** CCI achieves AUC comparable to or higher than the baseline and yields clearer separation, with higher scores for culture-specific sentences and lower scores for general ones. The baseline, in contrast, tends to assign relatively high scores to many sentences, and for some models the class medians are nearly identical. This may be because directly quantifying “culture” as a single scalar is inherently difficult, whereas CCI decomposes the task into per-culture generality estimates, thereby stabilizing the inference process.

**Model suitability** Regarding which LLMs

are suitable for computing cultural specificity scores, gpt-oss achieves near-ideal separation under both CCI and the baseline. This appears to reflect not only model size but also its reasoning-oriented architecture, which captures cultural differences through step-by-step reasoning. In addition, Japanese-specialized models show better separation than multilingual models. Overall, models that combine strong reasoning capabilities and a deep understanding of the target culture, while also possessing knowledge of other cultures, are most suitable for computing cultural specificity scores.

### 5.2 Controllability of Cultural Scope

**CCI vs. Baseline** From Table 1, we observe that under CCI’s Custom mode (+*Neighbor*), the median score for the culture-specific class is lower than in the Global mode. This suggests that the cultural scope can be adjusted to avoid overestimating practices common in neighboring cultures. By contrast, the baseline appears to be sensitive to prompt wording and input language, indicating that cultural scope is difficult to control solely through textual instructions. Additionally, CCI provides a numerical assessment of cultural specificity together with per-culture generality scores; even when baseline accuracy is high, CCI offers greater interpretability by indicating in which cultures a sentence is considered common or uncommon.

**Case analysis** Figure 2 shows a subset of the evaluation instances along with the CCI scores produced by gpt-oss. In individual cases, actions that may be taboo in neighboring cultures (e.g., “Pick up the small bowl and bring it to your mouth.”) should not receive excessively low scores even under +*Neighbor*. Conversely, practices widely observed across regions (e.g., “Taking milk out of the refrigerator.”) should not receive high scores even under -*Neighbor*. Consistent with these ex-

Label	Sentence $x$	Mode	Generality by country $p_c$	CCI
General	冷蔵庫から牛乳を取り出す。 Taking milk out of the refrigerator.	Custom (-Neighbor)	🇺🇸 : 0.88, 🇫🇷 : 0.90, 🇺🇸 : 0.92, 🇯🇵 : 0.93	0.033
		Custom (+Neighbor)	🇯🇵 : 0.90, 🇺🇸 : 0.91, 🇺🇸 : 0.92, 🇯🇵 : 0.95	0.039
Cultural	玄関で靴を脱ぐ。 Take off your shoes at the entrance.	Custom (-Neighbor)	🇺🇸 : 0.50, 🇫🇷 : 0.52, 🇺🇸 : 0.33, 🇯🇵 : 0.98	0.533
		Custom (+Neighbor)	🇯🇵 : 0.50, 🇺🇸 : 0.80, 🇺🇸 : 0.23, 🇯🇵 : 0.97	0.456
	小鉢を手に持って口に運ぶ。 Pick up the small bowl and bring it to your mouth.	Custom (-Neighbor)	🇺🇸 : 0.20, 🇫🇷 : 0.25, 🇺🇸 : 0.25, 🇯🇵 : 0.93	0.700
		Custom (+Neighbor)	🇯🇵 : 0.43, 🇺🇸 : 0.45, 🇺🇸 : 0.08, 🇯🇵 : 0.93	0.611
	エスカレーターで片側を空ける。 Leave one side open on the escalator.	Custom (-Neighbor)	🇺🇸 : 0.42, 🇫🇷 : 0.38, 🇺🇸 : 0.30, 🇯🇵 : 0.93	0.567
		Custom (+Neighbor)	🇯🇵 : 0.43, 🇺🇸 : 0.80, 🇺🇸 : 0.23, 🇯🇵 : 0.92	0.428
	節分に豆を撒く。 Scatter beans on Setsubun.	Custom (-Neighbor)	🇺🇸 : 0.04, 🇫🇷 : 0.04, 🇺🇸 : 0.04, 🇯🇵 : 0.95	0.910
		Custom (+Neighbor)	🇯🇵 : 0.07, 🇺🇸 : 0.12, 🇺🇸 : 0.03, 🇯🇵 : 0.98	0.912

Figure 2: Example sentences and their corresponding CCI scores computed by gpt-oss.

Bin statistics		Accuracy		
Range	#Items	Qwen 2.5	Llama 3.1	llm-jp 3.1
$CCI \leq 0.1$	583	0.940	0.899	0.967
$0.1 < CCI \leq 0.2$	75	0.893	0.880	0.987
$0.2 < CCI \leq 0.3$	69	0.826	0.870	0.971
$0.3 < CCI \leq 0.4$	40	0.900	0.775	0.925
$0.4 < CCI \leq 0.5$	19	0.789	0.789	0.842
$0.5 < CCI \leq 0.6$	63	0.873	0.825	0.905
$0.6 < CCI \leq 0.7$	36	0.889	0.806	0.972
$0.7 < CCI \leq 0.8$	45	0.867	0.800	0.978
$0.8 < CCI \leq 0.9$	173	0.873	0.821	0.942
$0.9 < CCI \leq 1.0$	16	0.750	0.750	0.938
Overall Accuracy		0.904	0.864	0.958

Table 3: Number of JCQA items and model-wise accuracy when CCI is binned in increments of 0.1.

Bin statistics		Accuracy		
Range	#Items	Qwen 2.5	Llama 3.1	llm-jp 3.1
$CCI \leq 0.1$	3217	0.846	0.836	0.924
$0.1 < CCI \leq 0.2$	117	0.778	0.795	0.906
$0.2 < CCI \leq 0.3$	116	0.741	0.724	0.897
$0.3 < CCI \leq 0.4$	72	0.708	0.694	0.931
$0.4 < CCI \leq 0.5$	60	0.733	0.767	0.883
$0.5 < CCI \leq 0.6$	75	0.680	0.773	0.880
$0.6 < CCI \leq 0.7$	94	0.755	0.691	0.894
$0.7 < CCI \leq 0.8$	138	0.761	0.681	0.891
$0.8 < CCI \leq 0.9$	102	0.745	0.657	0.892
$0.9 < CCI \leq 1.0$	1	1.000	1.000	1.000
Overall Accuracy		0.826	0.814	0.918

Table 4: Number of JCM items and model-wise accuracy when CCI is binned in increments of 0.1.

expectations, the presence or absence of similar practices in neighboring cultures yields systematic differences between the *+Neighbor* and *-Neighbor* conditions, indicating that CCI effectively operationalizes cultural scope control as intended.

### 5.3 Task Accuracy Shifts across CCI Levels

Tables 3 and 4 show the results for JCQA and JCM, respectively, where items are binned by CCI in increments of 0.1. Across both datasets, the CCI distribution is skewed toward lower values, suggesting that JCQA and JCM contain many culturally non-specific commonsense questions.

Accuracy tends to decrease as CCI increases, and higher-CCI bins often fall below the overall dataset accuracy. This indicates that items with higher cultural specificity tend to be more challenging, and that current models may not sufficiently acquire culture-specific knowledge. In contrast, llm-jp maintains higher overall accuracy and exhibits a comparatively smaller drop in high-CCI bins. This suggests that models trained on Japanese data may have an advantage on items that can reflect

commonsense knowledge in the Japanese cultural sphere. Overall, CCI-based stratification makes clear how performance varies with cultural specificity. This variation is difficult to observe from overall accuracy alone, and the results suggest that model gaps tend to widen in higher-CCI bins.

## 6 Conclusion

We introduced the Conceptual Cultural Index (CCI), a metric that quantifies sentence-level cultural specificity as the relative difference in generality between a target culture and other cultures. CCI remains effective even when the baseline struggles to assign stable scores and provides interpretable estimates grounded in an explicit definition. CCI supports practical culture-related workflows, including annotating and stratifying benchmarks for model evaluation, as well as filtering culture-specific knowledge data. Our method is presented as an evaluation framework that can be applied consistently as models evolve.

## Limitations

This study has three limitations. First, because cultures were approximated primarily at the country level, intra-country heterogeneity, such as regional or generational differences, may not be fully captured. Second, our experiments focused on Japan as the target culture, and thus the generalizability to other languages and regions remains to be established. Third, CCI relies on LLM-generated generality scores and thus inherits the biases and calibration issues of the underlying models. With these points in mind, we plan to develop CCI into a more robust and general-purpose evaluation framework by refining the granularity of cultural groupings and conducting broader multilingual and cross-regional evaluations.

## References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, and 261 others. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*.
- Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Shivdutt Singh, Alham Fikri Aji, Jacki O’Neill, Ashutosh Modi, and Monojit Choudhury. 2024. Towards Measuring and Modeling “Culture” in LLMs: A Survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15763–15784.
- Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, Che Chang, Kai Chen, and 106 others. 2025. gpt-oss-120b & gpt-oss-20b Model Card. *arXiv preprint arXiv:2508.10925*.
- Akiko Aizawa, Eiji Aramaki, Bowen Chen, Fei Cheng, Hiroyuki Deguchi, Rintaro Enomoto, Kazuki Fujii, Kensuke Fukumoto, Takuya Fukushima, Namgi Han, Yuto Harada, Chikara Hashimoto, Tatsuya Hiraoka, Shohei Hisada, Sosuke Hosokawa, Lu Jie, Keisuke Kamata, Teruhito Kanazawa, Hiroki Kanezashi, and 62 others. 2024. LLM-jp: A Cross-organizational Project for the Research and Development of Fully Open Japanese LLMs. *arXiv preprint arXiv:2407.03963*.
- Shane Arora, Marzena Karpinska, Hung-Ting Chen, Ipsita Bhattacharjee, Mohit Iyyer, and Eunsol Choi. 2025. CaLMQA: Exploring culturally specific long-form question answering across 23 languages. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11772–11817.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language Models are Few-Shot Learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, volume 33, pages 1877–1901.
- Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. 2023. Assessing Cross-Cultural Alignment between ChatGPT and Human Societies: An Empirical Study. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP*, pages 53–67.
- Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, and Yejin Choi. 2025. CulturalBench: A Robust, Diverse and Challenging Benchmark for Measuring LMs’ Cultural Knowledge Through Human-AI Red-Teaming. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25663–25701.
- Awantee Deshpande, Dana Ruitter, Marius Mosbach, and Dietrich Klakow. 2022. StereoKG: Data-Driven Knowledge Graph Construction For Cultural Knowledge and Stereotypes. In *Proceedings of the Sixth Workshop on Online Abuse and Harms*, pages 67–78.
- Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. 2024. Continual Pre-Training for Cross-Lingual LLM Adaptation: Enhancing Japanese Language Capabilities. In *Proceedings of the First Conference on Language Modeling*.
- Yi Fung, Ruining Zhao, Jae Doo, Chenkai Sun, and Heng Ji. 2024. Massively Multi-Cultural Knowledge Acquisition & LM Benchmarking. *arXiv preprint arXiv:2402.09369*.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*.
- Eunsu Kim, Juyoung Suk, Philhoon Oh, Haneul Yoo, James Thorne, and Alice Oh. 2024. CLiCK: A Benchmark Dataset of Cultural and Linguistic Intelligence in Korean. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics*,

- Language Resources and Evaluation*, pages 3335–3346.
- Fajri Koto, Rahmad Mahendra, Nurul Aisyah, and Timothy Baldwin. 2024. IndoCulture: Exploring Geographically Influenced Cultural Commonsense Reasoning Across Eleven Indonesian Provinces. *Transactions of the Association for Computational Linguistics*, 12:1703–1719.
- Kentaro Kurihara, Daisuke Kawahara, and Tomohide Shibata. 2022. JGLUE: Japanese General Language Understanding Evaluation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2957–2966.
- Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Afina Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, Víctor Gutiérrez-Basulto, Yazmín Ibáñez García, Hwaran Lee, Shamsuddeen Hassan Muhammad, Kiwoong Park, Anar Sabuhi Rzayev, Nina White, Seid Muhie Yimam, Mohammad Taher Pilehvar, and 3 others. 2024. BLEnD: A Benchmark for LLMs on Everyday Knowledge in Diverse Cultures and Languages. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, volume 37, pages 78104–78146.
- Tarek Naous, Michael J Ryan, Alan Ritter, and Wei Xu. 2024. Having Beer after Prayer? Measuring Cultural Bias in Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16366–16393.
- Tuan-Phong Nguyen, Simon Razniewski, Aparna Varde, and Gerhard Weikum. 2023. Extracting Cultural Commonsense Knowledge at Scale. In *Proceedings of the ACM Web Conference 2023*, page 1907–1917.
- Tuan-Phong Nguyen, Simon Razniewski, and Gerhard Weikum. 2024. Cultural Commonsense Knowledge for Intercultural Dialogues. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, page 1774–1784.
- Abhinav Sukumar Rao, Akhila Yerukola, Vishwa Shah, Katharina Reinecke, and Maarten Sap. 2025. NormAd: A Framework for Measuring the Cultural Adaptability of Large Language Models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2373–2403.
- Siqi Shen, Lajanugen Logeswaran, Moontae Lee, Honglak Lee, Soujanya Poria, and Rada Mihalcea. 2024. Understanding the Capabilities and Limitations of Large Language Models for Cultural Commonsense. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5668–5680.
- Weiyan Shi, Ryan Li, Yutong Zhang, Caleb Ziems, Sunny Yu, Raya Horesh, Rogério Abreu De Paula, and Diyi Yang. 2024. CultureBank: An Online Community-Driven Knowledge Base Towards Culturally Aware Language Technologies. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4996–5025.
- Jiaxing Sun, Weiquan Huang, Jiang Wu, Chenya Gu, Wei Li, Songyang Zhang, Hang Yan, and Conghui He. 2024. Benchmarking Chinese Commonsense Reasoning of LLMs: From Chinese-Specifics to Reasoning-Memorization Correlations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11205–11228.
- Masashi Takeshita and Rafal Rzepka. 2025. JETHICS: Japanese Ethics Understanding Evaluation Dataset. *arXiv preprint arXiv:2506.16187*.
- Yuhang Wang, Yanxu Zhu, Chao Kong, Shuyu Wei, Xiaoyuan Yi, Xing Xie, and Jitao Sang. 2024. CDEval: A Benchmark for Measuring the Cultural Dimensions of Large Language Models. In *Proceedings of the 2nd Workshop on Cross-Cultural Considerations in NLP*, pages 1–16.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 23 others. 2024. Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115*.
- Da Yin, Hritik Bansal, Masoud Monajatipoor, Lillian Harold Li, and Kai-Wei Chang. 2022. GeoMLAMA: Geo-Diverse Commonsense Probing on Multilingual Pre-Trained Language Models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2039–2055.

## A Prompts

This section provides the prompts used in our experiments. Table 5 presents the prompt for obtaining per-culture generality scores used to compute CCI; Table 6 shows the prompt for direct cultural specificity scoring used in the baseline comparison; and Table 7 and 8 provides the prompts used to generate the evaluation data (culture-specific and general sentences).

## B Model Details

We use the following LLMs in our experiments:

- Llama 3.1 8B Instruct:  
<https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct>

---

```

1 Task:
2 Rate how COMMON/FAMILIAR the following item is in each country (0.00 = not
  common, 1.00 = very common). Treat countries independently. Be language-
  agnostic: interpret the statement regardless of its language.
3
4 Statement: {sentence}
5 Countries: {cultures}
6
7 Rules:
8 - Use general knowledge; avoid stereotypes.
9 - If similarly common across many countries, use similar (even identical) scores
10 - If unsure, use mid values (e.g., 0.50).
11 - Do NOT normalize across countries.
12
13 Output JSON ONLY (no prose):
14 Schema: {"scores": {"<country>": <float>}}
15 Constraints: use the country names exactly as provided; floats in [0.00, 1.00],
  rounded to two decimals.

```

---

Table 5: Prompt for obtaining per-culture generality scores used in computing CCI. The input consists of a sentence  $x$  and a set of cultures  $\mathcal{C}$ .

---

```

1 Task:
2 Decide whether the following item is culturally specific to the TARGET country,
  and output a numeric specificity score only
3 (0.00 = globally common; 1.00 = unique to the target).
4
5 Statement: {sentence}
6 Target country: {target_culture}
7
8 Rules:
9 - Use general knowledge; avoid stereotypes.
10 - Be language-agnostic; interpret the statement regardless of its language.
11 - If unsure, use mid values (e.g., 0.50).
12
13 Output JSON ONLY (no prose):
14 {"score": <float in [0.00,1.00] rounded to two decimals>}

```

---

Table 6: Prompt for directly predicting cultural specificity scores in the target culture by an LLM (Baseline). The input consists of a sentence  $x$  and a target culture  $t$ .

- Qwen 2.5 7B Instruct:  
<https://huggingface.co/Qwen/Qwen2.5-7B-Instruct>
- gpt-oss-20B:  
<https://huggingface.co/openai/gpt-oss-20b>
- Llama 3.1 Swallow 8B Instruct v0.5:  
<https://huggingface.co/tokyotech-llm/Llama-3.1-Swallow-8B-Instruct-v0.5>
- LLM-jp 3.1 13B instruct4:  
<https://huggingface.co/llm-jp/llm-jp-3.1-13b-instruct4>

---

```

1 You are an assistant for creating a short-sentence corpus.
2 Strictly satisfy the requirements below.
3
4 Goal:
5 - Collect very ordinary events that could occur in any region of the world.
6
7 Output:
8 - Exactly one natural Japanese sentence, short in length (about 10-20 characters
9   ).
10 - Describe facts plainly without evaluations, impressions, or subjectivity.
11 - Avoid place names, specific store names, and personal names. Avoid excessive
12   stereotypes.
13 - Vary expressions, vocabulary, and scenes so that the same sentence endings and
14   the same constructions do not appear consecutively.
15
16 Strict requirements:
17 - The output must be a JSON array. Each element must have the form { "text": "<
18   one sentence>" }.
19 - The number of items must be exactly 300.
20 - Do not include any additional explanations, labels, or numbering (do not
21   output any strings other than JSON).
22
23 Example (format only):
24 [
25   { "text": "Turn off the alarm in the morning." },
26   { "text": "Wait for the train at the station." }
27 ]

```

---

Table 7: Prompt for generating general sentences. We used a Japanese prompt in our experiments; the version shown here is the English translation.

---

```

1 You are an assistant for creating a short-sentence corpus.
2 Strictly satisfy the requirements below.
3
4 Goal:
5 - Broadly collect, in short sentences, Japan-specific customs, daily life
6   culture, annual events, food culture, public manners, etc.
7
8 Output:
9 - Exactly one natural Japanese sentence, short in length (about 10-20 characters
10  ).
11 - Describe events plainly without evaluations or impressions.
12 - Avoid place names, specific store names, and personal names. Avoid excessive
13   stereotypes.
14 - Vary expressions, vocabulary, and scenes so that the same sentence endings and
15   the same constructions do not appear consecutively.
16
17 Strict requirements:
18 - The output must be a JSON array. Each element must have the form { "text": "<
19   one sentence>" }.
20 - The number of items must be exactly 300.
21 - Do not include any additional explanations, labels, or numbering (do not
22   output any strings other than JSON).
23
24 Example (format only):
25 [
26   { "text": "Take off your shoes at the entrance." },
27   { "text": "Use the purification basin at a shrine." }
28 ]

```

---

Table 8: Prompt for generating Japan-specific (culture-specific) sentences. We used a Japanese prompt in our experiments; the version shown here is the English translation.

# The Anthropology of Food: How NLP can Help us Unravel the Food Cultures of the World

Arij Riabi<sup>1\*</sup>, Sougata Saha<sup>2</sup>, Monojit Choudhury<sup>2</sup>

<sup>1</sup>INRIA Paris, <sup>2</sup>Mohamed bin Zayed University of Artificial Intelligence

<sup>1</sup>arij.riabi@devoteam.com, <sup>2</sup>{sougata.saha, monojit.choudhury}@mbzuai.ac.ae

## Abstract

Food carries cultural meaning beyond nutrition. It shapes identity, memory, and social norms, which makes it a central concern in anthropology. Given the diversity of food practices across cultures, analyzing them at scale while preserving their depth (“thick” descriptions) remains difficult for ethnographic methods, where Natural Language Processing (NLP) methods can help. Earlier NLP tools often captured only surface-level “thin” descriptions. Recent methods, especially Large Language Models (LLMs), create openings to recover cultural nuance. In this position paper, we outline research questions at the intersection of food anthropology and NLP, and discuss how LLMs can enable a scalable and culturally grounded anthropology of food. We present a case study examining what LLMs represent about global eating habits, which are often shaped by colonial histories and globalization. Our findings suggest that LLMs’ internal representations recognize cultural clusters, such as shared food habits among formerly colonized regions, but fail to grasp the pragmatic and experiential aspects of food, like the worldwide spread of dishes like pizza or biryani. We conclude by highlighting some of the potential risks and gaps of using NLP for cultural analysis.

## 1 Introduction

*“Dis-moi ce que tu manges, je te dirai ce que tu es”* – Brillat-Savarin (1842)

The famous aphorism by Jean Anthelme Brillat-Savarin links food and identity by suggesting that what we eat reflects who we are, our culture, values, and way of life. It frames eating not simply as a biological act but as an expression of the self. Anthropology has long embraced this view. Food is not merely sustenance but a symbolic system encoding identity, hierarchy, emotion, and collective

memory (Fischler, 1988; Douglas, 1972; Sutton, 2010). It is studied both *as an artifact* - an object worthy of analysis in its own right - and *as a lens* to understand broader social, historical, and political processes (Mintz, 1985; Poulain, 2012). From the structure of meals that signal intimacy or exclusion (Douglas, 1972) to the ways sugar production reveals colonial power (Mintz, 1985), food functions as a cultural archive that is both deeply embodied and symbolically charged, including through the socially learned cultivation of tastes (Bourdieu, 1984). Such work exemplifies Geertz’s notion of “thick description” (Geertz, 1973) of culture, which seeks to interpret meaning, context, and symbolism beyond observable behavior.

However, food, as an integral part of culture, shares many of culture’s defining characteristics: it is ever-evolving, long-tailed (composed of diverse practices rather than a few dominant ones), and experiential (it is a human experience rather than mere factual knowledge) (Saha et al., 2025). Classically, anthropology relies on ethnography, a method grounded in deep, qualitative immersion within specific communities to uncover “thick” descriptions of meaning. While this approach provides rich, contextualized insights, it is inherently labor-intensive and difficult to scale or compare systematically across global populations. This makes it difficult for ethnographic approaches alone to capture the diversity and scale of food’s cultural complexity, an area where computational methods, particularly NLP, can help. So far, most NLP studies of food have largely focused on recipes, sentiment, or trivia datasets (Hu et al., 2024; Palta and Rudinger, 2023; Zhou et al., 2025b), operating within what anthropology might call a “thin” descriptive paradigm. With the advent of LLMs, however, NLP systems can now engage with richer contextual and cultural dimensions of language, opening the possibility of scaling “thick” interpretation (Kommers et al., 2025). This shift presents a grow-

\*Now at Devoteam.

ing opportunity for NLP to move beyond shallow text processing and toward enabling a deeper, large-scale understanding of how food encodes identity, power, and collective memory across societies.

In this paper, we argue for re-conceptualizing food in NLP along two complementary dimensions: (i) **Food as Artifact** - the subject of analysis, and (ii) **Food as Lens** - a proxy for studying identity, class, and sociohistorical change. We contrast how food has been studied in anthropology and NLP, and propose a set of research questions that emerging NLP methods, especially LLMs, can help explore more deeply. As an initial demonstration, we present a case study analyzing what LLMs implicitly represent about global daily eating habits and how these representations align with anthropological evidence on colonial histories and globalization. Overall, our contributions are:

- A systematic survey of recent NLP research on food, contrasted with anthropological accounts of food as symbolic and contextual.
- A two-dimensional conceptual framework to reorient NLP toward more culturally meaningful representations.
- A set of interdisciplinary research questions where NLP methods can help study societal concepts such as identity, behavior, status, and cultural change, using food as the lens.
- A case study demonstrating how LLMs can help frame novel anthropological hypotheses to study food and society.

## 2 A Primer on Food in Anthropology

*To incorporate a food is, in both real and imaginary terms, to incorporate all or some of its properties: we become what we eat. Incorporation is a foundation for identity - Fischler (1988)*

Anthropology has long treated food as a central object of inquiry (as an artifact), not only for what it reveals about eating practices themselves but also for how it reflects broader cultural, social, and political processes (as a lens). In practice, most anthropological works encompass both perspectives - analyzing food as a tangible object of study and as a symbolic medium of meaning, as illustrated in Figure 1. Here, we highlight some existing works that capture these dual roles of food.

**Food, Identity, and Incorporation.** Food is viewed as a symbolic system key to sociality and

identity (Mintz and Du Bois, 2002; Douglas, 1972), rather than mere sustenance. Fischler (1988)’s *principle of incorporation* frames eating as physical and symbolic assimilation: food internalizes cultural norms, affirms belonging, and marks boundaries of otherness (Grignon and Grignon, 1980).

**Food as Embodied and Everyday Practice.** Unlike music or dance, food is ubiquitous, multi-sensory, and both private and communal (Lupton, 1996; Fischler, 1988). Its taste, smell, and ritual regularity make it a vehicle for expressing identity, belonging, and exclusion (Caplan, 2013; Bourdieu, 1984). Food also encodes class, religion, gender, and power (Heldke, 1992), while carrying histories of migration, colonialism, and resistance (Pilcher, 2023; Wilk, 2006).

**Food as Social Code and Political Economy.** Douglas conceptualized meals as “codes” that signal social order through structure and participation (Douglas, 1972). Similarly, food classification systems reinforce group cohesion. At the political level, Mintz’s history of sugar (Mintz, 1985) illustrates how commodities embody colonial power and economic dependency. Culinary categories (e.g., “cuisine du terroir”) and gastronomic discourse also encode class, ideology, and resistance to homogenization (Nonini, 2013; Sutton, 2018).

**Food, Emotion, and Memory.** Food is deeply affective, eliciting disgust, nostalgia, comfort, and pleasure (Fischler, 1988; Lupton, 1996). Such emotions link eating to collective memory and personal identity. Proust’s “Madeleine” (Proust, 1913) remains a canonical symbol of food’s capacity to evoke involuntary memory. This affective power underscores food’s role in cultural continuity across time and place (Holtzman, 2006).

**Beyond Symbolism: Sensory Anthropology and Embodiment.** Recent frameworks emphasize food as a lived, sensory practice rather than just symbolic code (Fox, 2003; Counihan, 2018). The TEP10 model (Sproesser et al., 2025) distinguishes what people eat (ingredients, origins) from how they eat (rituals, mealtime structures). Religious traditions, such as Catholic culinary discourse, frame pleasure and embodiment as integral to communal and spiritual life (Poulain, 2012). Sensory anthropology thus foregrounds food as a site where culture is enacted, felt, and moralized (Howes, 2010; Korsmeyer and Sutton, 2011).

### Food as a Cultural Lens for Computation.

Food, in Newmark’s typology (Newmark, 2003), cuts across material culture, social organization, norms, and habits. It can be studied as a material object (tools, recipes) (Gremillion, 2011; Dalby, 2003; Alcock, 2005; Sutton, 2010; Staller, 2003) and as a lens onto ritual practice, hierarchy, and everyday behavior. Traditional anthropology offers thick, small-scale insights but struggles to compare cultures systematically. NLP, by contrast, can scale analyses of cookbooks, menus, and food discourse, revealing diachronic patterns, identity markers, and implicit norms. Embedding-based methods can uncover associations between food and identity, while LLMs offer a way to complement thick description with cross-cultural breadth (Kommers et al., 2025).

Collectively, such applications point to a growing space where NLP can enhance anthropology’s capacity to analyze culture at both depth and scale.

## 3 Food for Cultural Analytics in NLP

To understand how food has been treated across NLP studies, we compiled a curated corpus by systematically gathering papers from multiple sources.

### 3.1 Survey Methodology

We focused on major NLP venues: ACL, EMNLP, EACL, and NAACL, conducting a systematic search across titles, abstracts, and full texts. We used a Boolean combination of the following keywords “cultural adaptation,” or “culture” and “food,” “cuisine,” “recipe,” “gastronomy,” and “identity.” We deliberately prioritize \*CL venues to ensure methodological rigor, reproducibility, and alignment with the NLP community’s standards for peer-reviewed research. To broaden our scope, we also searched Google Scholar<sup>1</sup> and Semantic Scholar<sup>2</sup> for interdisciplinary work that integrates perspectives from anthropology, food studies, and cultural sociology, using the keywords “food and identity,” “food culture,” and “food as social practice.” We excluded works that focus only on nutrition or agriculture, resulting in 21 studies (Figure 1).

Although we did not restrict our search by publication date, we observed that most of the 21 publications were within the last three years (2023-2025), highlighting the recent emergence and growing momentum of interdisciplinary research at the intersection of NLP and culture. In a related sur-

vey on culture and NLP, Adilazuarda et al. (2024) analyzed 90 papers published between 2020 and 2024 and noted a clear shift in NLP research toward cultural awareness, particularly following the rise of LLMs. Liu et al. (2025) similarly highlighted this trend, showing an increasing emphasis on cultural adaptation and sensitivity across various NLP tasks. We observe this shift mirrored in the domain of food and NLP as well.

### 3.2 Survey Findings

**Food as Artifact, Not Lens.** Despite food’s deep cultural resonance, NLP research has often reduced it to a static artifact. Many studies use food data to test factual recall, translation accuracy, or retrieval performance without engaging with its social meaning. For example, Hu et al. (2024) developed a cross-cultural recipe retrieval dataset, while Cao et al. (2024) focused on Chinese-English recipe translation. Both highlight technical challenges but remain centered on correctness, ingredient substitution, or retrieval accuracy. Similarly, multimodal work (Fu et al., 2020; Bagler and Singh, 2018) shows that while models recognize food semantics across text and images, they fail to account for cultural context. For instance, two dishes may look visually similar but carry entirely distinct ritual or regional significance.

**Food as Diagnostic Probe.** Several works use food as a diagnostic probe to reveal cultural bias or commonsense gaps in models. FORK (Palta and Rudinger, 2023) surfaces US-centric assumptions in commonsense QA through culinary questions. Zhou et al. (2025b) introduce MAPO, a multilingual dataset for ingredient prediction, which demonstrates how models stereotype or overgeneralize cultural food knowledge. While valuable, these approaches primarily use food as factual tests, rather than as a lens into lived cultural practice.

### Food in Cultural Evaluation Frameworks.

Several benchmarks incorporate food into large-scale cultural evaluation, but often remain at a surface level. Winata et al. (2025a) developed visual QA tasks focused on dish recognition and cuisine origin, while Pellegrini et al. (2021) highlights the subjective and context-dependent challenges of modeling culinary substitutions. Counterfactual recipe generation by Liu et al. (2022) advances this line by probing compositional generalization, showing how models struggle to simulate stylistic or culturally specific preparation methods. To-

<sup>1</sup><https://scholar.google.com/>

<sup>2</sup><https://www.semanticscholar.org/>

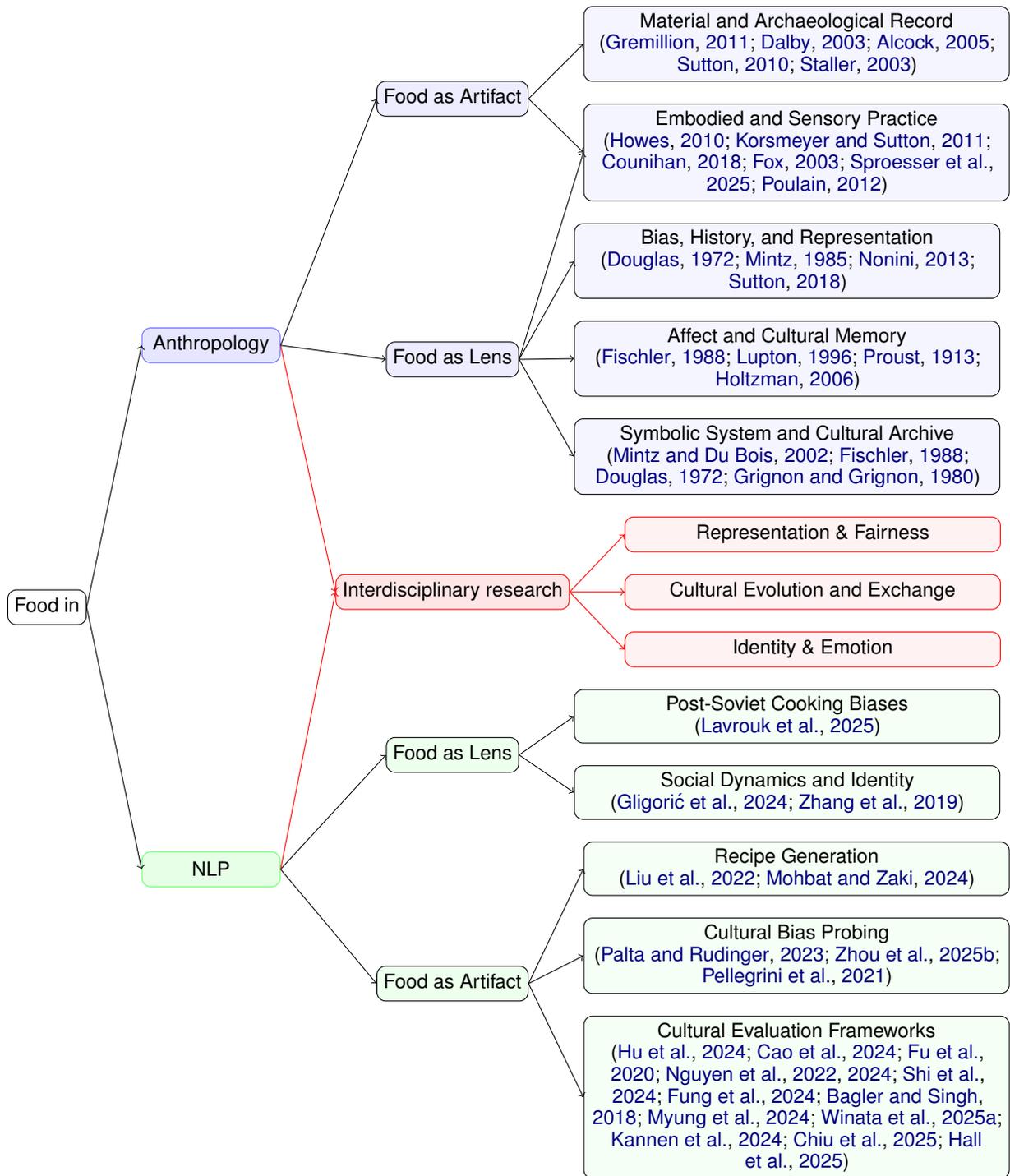


Figure 1: Taxonomy of the survey mapping research on food in NLP (highlighted in green) and anthropology (in violet), contrasting artifact-oriented with lens-based studies. The taxonomy illustrates our call for an interdisciplinary framework (in red) bridging computational and anthropological approaches.

gether, these studies highlight food’s potential as a domain for evaluating procedural knowledge and cultural embodiment, yet they stop short of engaging deeper cultural theorization.

Food also consistently appears as an important category in cultural evaluation benchmarks (Nguyen et al., 2022, 2024; Shi et al., 2024; Fung

et al., 2024; Myung et al., 2024; Chiu et al., 2025). Strikingly, Fung et al. (2024) shows that LLMs perform worse on cuisine-related reasoning than on domains such as education or holidays, since food often requires fine-grained, localized knowledge that is less documented, particularly for underrepresented communities. This illustrates the gap be-

tween models' generalized representations and the situated specificity of food practices.

**Food as Social and Behavioral Practice.** Few works treat food as a social practice. Gligorić et al. (2024) studies food purchase mimicry on a university campus, showing how choices reflect social alignment and symbolic inclusion. Their findings resonate with Fischler (1988)'s claim that eating is not just ingestion but assimilation of cultural values. Such perspectives position food as a *behavioral proxy* for social belonging, suggesting opportunities for NLP to model culture as practice, performance, and interaction.

**Bias, History, and Representation.** Recent work also highlights how LLMs reproduce cultural biases in food knowledge. Lavrouk et al. (2025) demonstrate that models misattribute dish origins, particularly in multilingual contexts. For instance, Russian prompts lead to systematic over-association with Russian cuisine, reflecting how linguistic dominance shapes cultural representation in LLMs. Zhang et al. (2019), focusing on food images, caution against assuming national datasets as mono-cultural. Their study links visual aesthetics to user engagement (likes, bookmarks, ratings) and calls for interdisciplinary collaboration with anthropologists to interpret cultural meaning. Together, these works underscore the socio-historical and representational dimensions of food in NLP.

**Parallel Streams in Nutrition and Health.** We also acknowledge parallel NLP research that studies food's nutrition and health aspects. Such studies use NLP to improve dietary outcomes and environmental sustainability (McCarthy, 2025), classify foods, recipes and assess diets (Zhou et al., 2025a; Rezayi et al., 2025), and extract dietary patterns (Choi et al., 2022). While these works are valuable for health and industry, since they treat food as a technical object and not as a cultural or social practice, they fall outside the scope of this paper.

### 3.3 Research on Culture and LLMs

Adilazuarda et al. (2024) survey how "culture" is represented and measured in LLMs, showing that most studies rely on abstract or *demographic proxies* and primarily test for factual correctness - the "thin" description of culture - leaving its experiential and contextual dimensions underexplored. We extend this discussion by focusing on food, an integral and embodied aspect of culture. Many

existing approaches, often using prompting techniques such as *socio-demographic prompting* (Li et al., 2024; AlKhamissi et al., 2024; Wan et al., 2023), instruct LLMs to assume diverse cultural identities and exhibit behaviors representative of an average person from the assumed culture. A model is then considered culturally aware if its generated behavior aligns with expected responses (Kharchenko et al., 2024; LI et al., 2024; Dawson et al., 2024). However, several works emphasize that culture should not be reduced to static facts or trivia but understood as a dynamic, situated, and context-dependent practice, which LLMs can help understand at scale (Zhou et al., 2025c; Kommers et al., 2025). Similarly, Saha et al. (2025) advocates for developing *meta-cultural competent* models that adaptively engage with unfamiliar cultural contexts instead of relying solely on static data.

Overall, NLP research has mostly examined food as an *artifact*, a *diagnostic probe*, or a *benchmark category*. Building on insights from anthropology and the social sciences, we propose that NLP can also treat food as a *lens*, to study broader cultural patterns, meanings, and social relations.

## 4 Towards Cross-Disciplinary Research

NLP methods, particularly LLMs, enable cultural analysis at a scale that traditional anthropological and social science approaches cannot easily achieve (Kommers et al., 2025). Bridging these fields, we outline ten research questions along three broad dimensions (illustrated in Figure 1) through which LLMs can uncover global food patterns, empirically revisit theories in structural anthropology (Strauss, 1974; Hénaff, 1998) and computational social science (Lazer et al., 2009; Goel and Bagler, 2022; Ahnert, 2013; Edelman et al., 2020).

### 4.1 Identity and Emotion

**How does food-related discourse reflect emerging identities & groups?** Food practices signal inclusion or exclusion, shaping groups and collective identities (e.g., veganism, nationalism). NLP can track how communities construct belonging through hashtags, reviews, or recipe-sharing forums, revealing negotiated identity markers.

**How do food preferences reflect lifestyle, value systems, or behavioral patterns?** Shifts from traditional to modern food practices mirror broader societal change (Sproesser et al., 2022). NLP can analyze online posts and blogs to capture how food-

related language reflects urbanization, global influences, individual lifestyle choices (e.g., clean eating, comfort food), and evolving cultural identity.

**How are emotions and affect encoded in food-related language across cultures?** Food evokes strong emotions such as nostalgia, pride, disgust, etc, which shape identity and memory (Lupton, 1996; Holtzman, 2006). Sentiment analysis, emotion tagging, and metaphor mining can uncover cross-cultural variation and support more culturally aware emotion detection systems.

## 4.2 Cultural Evolution and Exchange

**How do food narratives reflect postcolonial influences & cultural exchange?** Food discourse often carries traces of colonial histories through ingredients, preparation methods, and naming conventions. Analyzing large-scale recipe corpora or food blogs can uncover how colonial legacies persist or evolve, revealing implicit power asymmetries and cultural adaptations. For example, words like “tacos,” “kimchi,” or “couscous” may appear in different cultural embeddings, traditional or modern eating, depending on regional context and temporal framing. This variation can be captured through dynamic embedding models or temporal word sense induction, offering insights into how culinary terms shift meaning and connotation over time and across communities (Sproesser et al., 2025). Our case study further illustrates how constructing country-country similarity networks from LLM probabilities enables systematic testing of anthropological hypotheses, such as the role of colonial histories or globalization in shaping food practices, on a scale far beyond traditional ethnographies.

**How have recipes evolved over time under cultural, economic, and environmental pressures?** Recipes are living cultural artifacts, continuously adapted across time and space. Diachronic NLP analysis of historical cookbooks and recipe websites can detect ingredient substitutions, preparation shifts, and linguistic trends that reflect socioeconomic and ecological change.

**How are food rituals and communal practices reflected in text, and what can they reveal about social cohesion?** Food rituals reinforce collective norms. NLP can analyze ceremonial menus, festival descriptions, or religious texts to uncover recurring structures of ritualistic eating (e.g., Passover, Eid, Onam feasts, etc), which reflect temporality

and shared group values.

**How does food-related discourse reflect emerging cultural shifts around health, sustainability, and environmental consciousness?** Language around plant-based diets, organic food, zero-waste cooking, etc, reflects evolving cultural values tied to global health and environmental movements. NLP methods can analyze trends in social media, blogs, or menus to trace how these discourses spread, who participates in them, and how they intersect with local traditions or resistance. This offers insight into how cultural norms around food are negotiated and transformed in response to global challenges, which can in turn inform policy or behavioral interventions.

**How can LLMs simulate social dynamics in food discourse (e.g., norm shifts, dietary trends)?** Recent LLM advancements have demonstrated their potential as *world models* (Shavit et al., 2023; Acharya et al., 2025), indicating the possibility of simulating human behavior at scale. Studies have shown that LLMs can approximate social behavior in multi-agent environments and generative simulations (Park et al., 2023; Zhou et al., 2023; Wang et al., 2024; Zhou et al., 2024). Although these models are limited in scope and lack deep generalization (Saha and Choudhury, 2025), they point towards a promising direction: the potential to simulate cultural dynamics, including food-related behaviors. LLMs can be fine-tuned or prompted to model hypothetical cultural changes (e.g., how food taboos emerge/fade). Although these simulations cannot replace anthropological expertise, they can support hypothesis testing by letting researchers study sociolinguistic drift, normative change, and symbolic meaning at scale.

## 4.3 Linguistic Variation and Representation

Food names reflect hierarchies and regional identities. For example, the multiple ways of expressing “breakfast” in Malay, Mandarin, and Tamil, such as “morning rice” or “layering the belly”, illustrate how meals are culturally framed rather than just names, highlighting challenges of translation ambiguity in NLP (Poulain et al., 2023).

**How do dialects and local food terms affect model understanding and cultural representation?** Ingredient names and local terminology often encode social or regional identity as culture plays a role in shaping food semantics (Mazzuca

and Majid, 2023). Studying how LLMs interpret or generate these can reveal representational biases and inform improvements in cultural adaptation. For example, ingredient names (“bhindi” vs “okra”, “aubergine” vs “eggplant”, etc) often vary by region and carry cultural significance. NLP can surface biases, mismatches, or translation failures in handling this diversity. This could also enable studying the dimensions along which cultures vary: aboutness, common ground, and linguistic form and style (Hershcovich et al., 2022). For example, analyzing Google reviews across different price ranges and cuisine categories - from fine dining to fast food - could help uncover variations in language registers, dialect usage, and cultural framing.

**How can NLP mitigate representational disparities for underrepresented cuisines and communities?** Food practices often reflect regional and cultural identities that do not neatly align with nation-states. For example, while countries like China, India, Mexico, and Brazil show strong regional-national alignment (Sproesser et al., 2022), the U.S. exhibits distinct ethnic culinary clusters (e.g., African American, Latin American) that transcend a single national food identity and may even align with transnational groups. These patterns highlight the need for NLP models to capture intranational diversity rather than reducing multiethnic cuisines to monolithic national categories.

## 5 Case Study

To illustrate how LLMs can enrich our understanding of food as a cultural lens, we present a case study examining how textual representations of food reflect patterns of identity, belonging, and social change. Anthropological work shows that both colonization and geographic proximity shape eating habits (Douglas, 1972; Mintz, 1985; Fischler, 1988). Yet in today’s globalized world, dishes often transcend their place of origin. For instance, in 2024, DoorDash, a widely used food-delivery app in the USA and Canada, reported Indian dishes such as naan, butter chicken, and biryani among the most ordered items in the region (sources in Table 1). Food ordering platforms such as Talabat and Zomato reported pizza, an Italian dish, as a staple in Oman, the UAE, and India. We ask two questions inspired by anthropology: (Q1) **What country-level clusters of food practices do LLMs produce, and do these clusters reflect colonial histories?** (Q2) How connected are these clusters

in light of contemporary globalization - **do LLMs mirror lived cultural practices, or do they reproduce more static, fact-based representations?**

### 5.1 Setup

Several studies show that online food-ordering behavior reflects a combination of familiarity and convenience, both of which are closely tied to cultural exposure and dietary norms (Rozin, 2005; Warde, 2016; Zhang et al., 2025; Keeble et al., 2020). While convenience influences ordering frequency, familiarity largely determines what individuals feel comfortable consuming or associating with their identity. Thus, what people choose to order online serves as a meaningful behavioral proxy for cultural familiarity, particularly at the dish level, where preferences reflect social and cultural embeddedness rather than mere accessibility.

To answer the two questions, we adapted the Food Choice Questionnaire (FCQ) (Steptoe et al., 1995), focusing on two dimensions: **convenience** (foods easily available nearby) and **familiarity** (familiar foods). We prompted two instruction-tuned LLMs - Llama-3.1-8B-Instruct (Dubey et al., 2024) and Gemma-2-9B-it (Team et al., 2024) - across 170 countries, and observed their log probabilities for tokens corresponding to a curated list of approximately 3,500 food items. For each model and dimension, we constructed a country  $\times$  food matrix  $M$  by collating the token log-probabilities (converted to probabilities using softmax), and computed a country-country similarity matrix  $S_{cc} = MM^T$ . We detected communities by constructing a weighted mutual  $k$ -NN graph on  $S_{cc}$  with  $k = 10$ , and applied the Leiden community detection algorithm (Traag et al., 2018), which optimizes modularity  $Q$  (Newman and Girvan, 2004), a measure that compares intra-cluster connectivity with a random baseline. Further details in Section A.1.

### 5.2 Results and Discussion

To answer Q1, we compared the clusters against a curated mapping between countries and their historical colonizers (details in Section A.1) and used Fisher’s exact test (Fisher, 1922) to evaluate whether LLM-derived food-based communities aligned with colonial histories within continents. Across both models and questions, the country-country graphs have high modularity ( $Q \in [0.79, 0.83]$ ) with 10-11 communities and far more intra than inter-community edges (detailed statistics in Table 2). **The clusters significantly align**

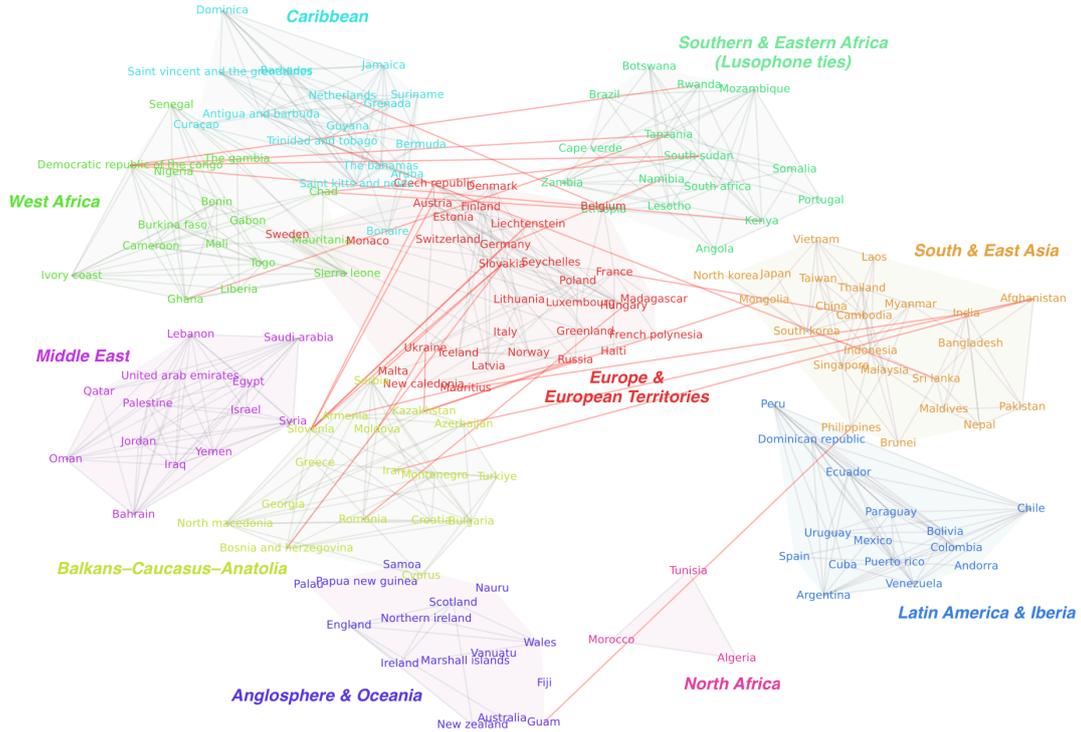


Figure 2: **Country communities from Llama (convenience).** Weighted mutual- $k$ NN graph ( $k = 10$ ) on  $S_{cc} = MM^T$ ; Leiden partition on modularity. High modularity ( $Q = 0.829$ ) with coherent regional clusters. Intra-community edges shown in gray, inter-community edges in red. Community details in Table 3. **Labels of the communities are heuristic; they reflect the majority of countries in each cluster.** Node colors are used solely for visual separation of clusters and do not carry additional semantic meaning.

**with colonial histories:** Fisher tests show higher odds ratios (OR) that countries sharing a colonizer fall in the same community (Llama: OR = 3.4 for convenience, 2.6 for familiarity; Gemma: 3.7 and 2.7). Map-based visualisations confirm tight regional blocks and few cross-block ties (Fig 6 - 9).

Figure 2 illustrates Llama’s convenience-based clusters, where countries were grouped into 10 communities. Interestingly, **we observe some countries occupying brokerage positions, acting as inter-community bridges.** From a structural perspective, such bridging roles are crucial for the diffusion of ideas, practices, and goods across cultural boundaries, as noted in classical network studies of innovation and knowledge transfer (Gould, 1993; Burt, 2004). While this remains a hypothesis that requires validation with real-world food exchange and migration patterns, our results suggest that LLMs capture relational structures that resonate with known mechanisms of cultural flow.

Beyond colonial clustering, Figure 2 also reveals **striking disconnections and over-connections.** For example, Morocco, Tunisia, and Algeria, commonly grouped as the Maghreb, appear fragmented,

while Arab countries are split into disconnected sub-clusters. **Europe shows dense interconnections,** plausibly reflecting its overrepresentation in textual corpora, while **Latin America forms a cluster with few external ties,** despite strong historical and migratory links to Europe. These anomalies suggest hypotheses: that LLM-derived clusters overemphasize colonial and linguistic segmentation, underrepresent globalization and migration, and reproduce biases in training data.

Some observed patterns may reflect artifacts of the training corpora rather than real-world dynamics. For instance, brokerage roles could shift if we examined an Arabic LLM trained on more region-specific data. This underscores both the promise and the limits of LLMs: while they encode patterns of cultural knowledge, these patterns may be shaped as much by training data distributions as by real-world dynamics. Nonetheless, analyzing such structures can generate testable hypotheses for validation against migration records, trade flows, or ethnographic accounts, thereby opening pathways for computational methods to complement anthropological research. We observe similar insights for

both dimensions and models (Figures 4, 3, 5).

To address Q2, we examined food ordering data (2021-2024) from DoorDash, Talabat, and Zomato for Australia, Canada, India, New Zealand, Oman, the UAE, and the USA. Table 1 shows that **pizza is a global staple** (e.g., Australia, India, UAE), while **Indian dishes such as naan, butter chicken, and biryani are frequently ordered** in Canada, New Zealand, India, and the USA, alongside local favorites like Wagyu beef dishes and California rolls. These trends highlight how globalization reshapes lived food practices beyond colonial and historical divides. The persistence of colonial clustering in LLM-derived communities suggests that models draw on historical categories embedded in training data as the basis for cultural organization. However, **sparse inter-cluster connectivity shows that models underrepresent transnational food flows** (migration-driven exchanges, globalized staples, etc) despite their prominence in the real world, reflecting the difference between knowledge in training data and the fluid practices of everyday eating.

Visualizing these clusters geographically reveals that countries and their food communities appear largely disconnected, highlighting the need for anthropologically grounded LLM evaluations that capture culture as lived and dynamic rather than merely factual artifacts (see Figure 6). Additional visualizations across models and dimensions further confirm these patterns of regional cohesion and sparse global connectivity (Figures 8,7,9).

## 6 Discussion and Conclusion

Although our proposed re-framing opens interdisciplinary possibilities, it also raises practical and conceptual challenges, a few of which are below:

**1. Do NLP systems need to engage with symbolic and experiential aspects of food?** One might argue that tasks aligned with factual correctness, such as recognizing dishes, generating recipes, or identifying cuisine origins, are sufficient for models. However, if NLP aspires to develop models that truly “understand” culture, it must capture the symbolic and experiential dimensions of food. Current benchmarks that treat cultural knowledge as static trivia fail to evaluate these deeper capabilities; moving from food as artifact to lens is thus a necessary step for robust cultural evaluation of LLMs. Computational approaches simulating societies (Epstein, 2012; Shavit et al., 2023) further emphasize the need to model the interconnectedness of food and

society, beyond mere factual recall.

**2. Are computational methods a replacement for cultural expertise?** Culturally adept NLP models are not substitutes for anthropologists, sociologists, or cultural historians. Instead, they serve as complementary tools, enabling empirical, scalable analysis (Newell and Simon, 2007; Polak, 2016), uncovering implicit cultural patterns, and testing hypotheses across large corpora (Kozłowski et al., 2019; Arseniev-Koehler and Foster, 2022), as further demonstrated by our case study. Nonetheless, careful reflection on methodological and epistemological limits is essential to avoid over-claiming what models can “know” about food and culture.

**3. Safety, stereotyping, and cultural bias.** Treating food as a cultural lens introduces ethical concerns. NLP models can perpetuate stereotypes, essentialize cuisines, or reinforce colonial hierarchies, reflecting biases in training data (Sheng et al., 2019; Gehman et al., 2020; Nadeem et al., 2021; Bender et al., 2021). In applications like virtual assistants, recipe recommenders, or educational tools, these biases can shape user perceptions, for example, by associating certain cuisines with poverty, hygiene, or spiciness. Western-centric framing in prior studies also marginalizes non-Western culinary traditions (Palta and Rudinger, 2023; Cao et al., 2024; Zhou et al., 2025b). Addressing these issues requires culturally diverse datasets, transparent documentation, and participatory design.

**4. Integrating cultural reflexivity and relativism.** Anthropology stresses reflecting on one’s own cultural assumptions (“cultural reflexivity”) and understanding practices in their own contexts (“cultural relativism”) (Davies, 2012; Bourdieu, 2004; Deer, 2014; Boas, 1941). For example, assumptions linking political hierarchy to elaborate cuisines are not universal, as shown by Gellner’s ethnography in Africa (Gellner, 1985). Similarly, NLP can benefit from reflexive and relativist principles, promoting models that account for meaning as situated and context-dependent, leading to more culturally competent and inclusive systems.

In conclusion, food is a rich cultural artifact that encodes values, identities, and social practices, offering NLP a unique opportunity to contribute to a deeper, anthropological understanding of culture through advanced computational methods such as LLMs. To advance this vision, we outline research directions at the intersection of food studies and NLP and illustrate their potential through an exploratory case study, and discuss their challenges.

## Limitations

While this paper highlights new interdisciplinary directions, several limitations must be acknowledged. **First**, our survey of NLP and mainly anthropology literature is necessarily selective and focuses primarily on English-language sources and major conferences. There is extensive anthropological work on food practices in non-Western contexts and diverse languages that we do not fully capture here. **Second**, many of the NLP studies we reviewed rely on digital corpora (e.g., online reviews, social media, cookbooks), which may over-represent urban, affluent, or globally connected communities while under-representing rural or marginalized voices. Food and its encompassing rituals and social dynamics are deeper, which digital sources might fail to capture. **Third**, our focus on symbolic and social dimensions of food intentionally excludes nutritional, agricultural, and product-level applications, which are important but outside our scope. Future work might consider how these perspectives intersect, particularly when addressing health disparities, fitness, and food security. **Fourth**, our case study, limited to two models and a small set of questions, is not intended as an exhaustive account of food-related patterns. Rather, it illustrates the feasibility of using LLMs for knowledge and pattern discovery through a network analysis approach, while acknowledging that other methods could equally be applied. Moreover, our analysis primarily used English food item names; incorporating local terms may reveal different cultural patterns, an important direction for future validation. **Finally**, the three dimensions proposed in Section 4 - “identity and emotion”, “cultural evolution and exchange”, and “linguistic variation and representation” - broadly capture most of the pertinent research directions at the intersection of food anthropology and NLP. However, the ten research questions themselves are not exhaustive. More questions are possible that belong to the three dimensions.

## Acknowledgements

This research was supported by Microsoft Accelerate Foundation Models Research (AFMR) Grant. Arij Riabi was partly funded by the BPI “Scribe” project.

## References

- Deepak Bhaskar Acharya, Karthigeyan Kuppan, and B Divya. 2025. Agentic ai: Autonomous intelligence for complex goals—a comprehensive survey. *IEEE Access*.
- Muhammad Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Singh, Alham Aji, Jacki O’Neill, Ashutosh Modi, and Monojit Choudhury. 2024. Towards measuring and modeling “culture” in llms: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15763–15784.
- Sebastian E Ahnert. 2013. Network analysis and data mining in food science: the emergence of computational gastronomy. *Flavour*, 2:1–3.
- Joan P Alcock. 2005. *Food in the ancient world*. Bloomsbury Publishing USA.
- Badr AlKhamissi, Muhammad ElNokrashy, Mai Alkhamissi, and Mona Diab. 2024. Investigating cultural alignment of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12404–12422, Bangkok, Thailand. Association for Computational Linguistics.
- Alina Arseniev-Koehler and Jacob G Foster. 2022. Machine learning as a model for cultural learning: Teaching an algorithm what it means to be fat. *Sociological Methods & Research*, 51(4):1484–1539.
- Ganesh Bagler and Navjot Singh. 2018. Data-driven investigations of culinary patterns in traditional recipes across the world. In *2018 IEEE 34th International Conference on Data Engineering Workshops (ICDEW)*, pages 157–162.
- Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.
- Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008.
- Franz Boas. 1941. Race, language and culture. *The Journal of Nervous and Mental Disease*, 94(4):513–514.
- Pierre Bourdieu. 1984. *Distinction: A Social Critique of the Judgement of Taste*. Harvard University Press.
- Pierre Bourdieu. 2004. *Science of science and reflexivity*. Polity.
- Jean Anthelme Brillat-Savarin. 1842. *Physiologie du Gout, ou, Méditations de gastronomie transcendante...* Charpentier.

- Ronald S Burt. 2004. Structural holes and good ideas. *American journal of sociology*, 110(2):349–399.
- Yong Cao, Yova Kementchedjheva, Ruixiang Cui, Antonia Karamolegkou, Li Zhou, Megan Dare, Lucia Donatelli, and Daniel Hershcovich. 2024. [Cultural adaptation of recipes](#). *Transactions of the Association for Computational Linguistics*, 12:80–99.
- Pat Caplan. 2013. *Food, health and identity*. Routledge.
- Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, and Yejin Choi. 2025. [CulturalBench: A robust, diverse and challenging benchmark for measuring LMs’ cultural knowledge through human-AI red-teaming](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25663–25701, Vienna, Austria. Association for Computational Linguistics.
- Insu Choi, Jihye Kim, and Woo Chang Kim. 2022. [Dietary pattern extraction using natural language processing techniques](#). *Frontiers in Nutrition*, Volume 9 - 2022.
- Carole M Counihan. 2018. *The anthropology of food and body: Gender, meaning and power*. Routledge.
- Andrew Dalby. 2003. *Food in the Ancient World from A to Z*. Psychology Press.
- Charlotte Aull Davies. 2012. *Reflexive ethnography: A guide to researching selves and others*. Routledge.
- Fiifi Dawson, Zainab Mosunmola, Sahil Pocker, Raj Abhijit Dandekar, Rajat Dandekar, and Sreedath Panat. 2024. [Evaluating cultural awareness of llms for yoruba, malayalam, and english](#). *Preprint*, arXiv:2410.01811.
- Cécile Deer. 2014. Reflexivity. In *Pierre Bourdieu*, pages 195–208. Routledge.
- Mary Douglas. 1972. Deciphering a meal. In *Implicit meanings: Essays in anthropology*, pages 249–275. Routledge.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Achim Edelmann, Tom Wolff, Danielle Montagne, and Christopher A Bail. 2020. Computational social science and sociology. *Annual review of sociology*, 46(1):61–81.
- Joshua M Epstein. 2012. *Generative social science: Studies in agent-based computational modeling*. Princeton University Press.
- Claude Fischler. 1988. Food, self and identity. *Social Science Information*, 27(2):275–292.
- Ronald A Fisher. 1922. On the interpretation of  $\chi^2$  from contingency tables, and the calculation of p. *Journal of the royal statistical society*, 85(1):87–94.
- Robin Fox. 2003. Food and eating: an anthropological perspective. *Social Issues Research Centre*, 2003:1–21.
- Han Fu, Rui jin Wu, Chenghao Liu, and Jianling Sun. 2020. [Mcen: Bridging cross-modal gap between cooking recipes and dish images with latent variable model](#). *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14558–14568.
- Yi Fung, Ruining Zhao, Jae Doo, Chenkai Sun, and Heng Ji. 2024. [Massively multi-cultural knowledge acquisition & lm benchmarking](#). *Preprint*, arXiv:2402.09369.
- Clifford Geertz. 1973. *The Interpretation of Cultures*. Basic Books, New York.
- Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. [RealToxicityPrompts: Evaluating neural toxic degeneration in language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, Online. Association for Computational Linguistics.
- Ernest Gellner. 1985. No haute cuisine in africa. *Relativism and the Social Sciences*.
- Kristina Gligorić, Arnaud Chiolero, Emre Kıcıman, Ryen W White, Eric Horvitz, and Robert West. 2024. [Food choice mimicry on a large university campus](#). *PNAS Nexus*, 3(12):pgae517.
- Mansi Goel and Ganesh Bagler. 2022. Computational gastronomy: A data science approach to food. *Journal of Biosciences*, 47(1):12.
- Roger V Gould. 1993. Collective action and network structure. *American sociological review*, pages 182–196.
- Kristen J Gremillion. 2011. *Ancestral appetites: food in prehistory*. Cambridge University Press.
- Claude Grignon and Christiane Grignon. 1980. Styles d’alimentation et goûts populaires. *Revue française de sociologie*, pages 531–569.
- Siobhan Mackenzie Hall, Samantha Dalal, Raesetje Se-fala, Foutse Yuehgoh, Aisha Alaagib, Imane Hamzaoui, Shu Ishida, Jabez Magomere, Lauren Crais, Aya Salama, and Tejumade Afonja. 2025. [The human labour of data work: Capturing cultural diversity through world wide dishes](#). *Proc. ACM Hum.-Comput. Interact.*, 9(7).
- Lisa M Heldke. 1992. Food politics, political food. *Cooking, eating, thinking: transformative philosophies of food*, pages 301–327.

- Marcel Hénaff. 1998. *Claude Levi-Strauss and the making of structural anthropology*. U of Minnesota Press.
- Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. [Challenges and strategies in cross-cultural NLP](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.
- Jon D. Holtzman. 2006. [Food and memory](#). *Annual Review of Anthropology*, 35(Volume 35, 2006):361–378.
- David Howes. 2010. *Sensual relations: Engaging the senses in culture and social theory*. University of Michigan Press.
- Tianyi Hu, Maria Maistro, and Daniel Hershcovich. 2024. [Bridging cultures in the kitchen: A framework and benchmark for cross-cultural recipe retrieval](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1068–1080, Miami, Florida, USA. Association for Computational Linguistics.
- Nithish Kannen, Arif Ahmad, Marco Andreetto, Vinodkumar Prabhakaran, Utsav Prabhu, Adji Bousso Dieng, Pushpak Bhattacharyya, and Shachi Dave. 2024. [Beyond aesthetics: Cultural competence in text-to-image models](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 13716–13747. Curran Associates, Inc.
- Matthew Keeble, Jean Adams, Gary Sacks, Lana Vanderlee, Christine M White, David Hammond, and Thomas Burgoine. 2020. Use of online food delivery services to order food prepared away-from-home and associated sociodemographic characteristics: a cross-sectional, multi-country analysis. *International journal of environmental research and public health*, 17(14):5190.
- Julia Kharchenko, Tanya Roosta, Aman Chadha, and Chirag Shah. 2024. [How well do llms represent values across cultures? empirical analysis of llm responses based on hofstede cultural dimensions](#). Preprint, arXiv:2406.14805.
- Cody Kommers, Drew Hemment, Maria Antoniak, Joel Z. Leibo, Hoyt Long, Emily Robinson, and Adam Sobey. 2025. [Meaning is not a metric: Using llms to make cultural context legible at scale](#). Preprint, arXiv:2505.23785.
- Carolyn Korsmeyer and David Sutton. 2011. The sensory experience of food. *Food, Culture & Society*, 14(4):461–475.
- Austin C Kozlowski, Matt Taddy, and James A Evans. 2019. The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review*, 84(5):905–949.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626.
- Anton Lavrouk, Tarek Naous, Alan Ritter, and Wei Xu. 2025. [What are foundation models cooking in the post-soviet world?](#) In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20698–20720, Suzhou, China. Association for Computational Linguistics.
- David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, and 1 others. 2009. Computational social science. *Science*, 323(5915):721–723.
- CHENG LI, Mengzhuo Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024. [CultureLLM: Incorporating cultural differences into large language models](#). In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- Huihan Li, Liwei Jiang, Jena D Hwang, Hyunwoo Kim, Sebastin Santy, Taylor Sorensen, Bill Yuchen Lin, Nouha Dziri, Xiang Ren, and Yejin Choi. 2024. Culture-gen: Revealing global cultural perception in language models through natural language prompting. *arXiv preprint arXiv:2404.10199*.
- Chen Cecilia Liu, Iryna Gurevych, and Anna Korhonen. 2025. [Culturally aware and adapted NLP: A taxonomy and a survey of the state of the art](#). *Transactions of the Association for Computational Linguistics*, 13:652–689.
- Xiao Liu, Yansong Feng, Jizhi Tang, Chengang Hu, and Dongyan Zhao. 2022. [Counterfactual recipe generation: Exploring compositional generalization in a realistic scenario](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7354–7370, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Deborah Lupton. 1996. Food, the body and the self.
- Claudia Mazzuca and Asifa Majid. 2023. [The semantic representation of food is shaped by cultural experience](#). *Language and Cognition*, 15(4):651–669.
- Danielle I. McCarthy. 2025. [Nutritional intelligence in the food system: Combining food, health, data and ai expertise](#). *Nutrition Bulletin*, 50(1):142–150.
- Sidney W Mintz. 1985. *Sweetness and power: The place of sugar in modern history*. Penguin.

- Sidney W. Mintz and Christine M. Du Bois. 2002. [The anthropology of food and eating](#). *Annual Review of Anthropology*, 31(Volume 31, 2002):99–119.
- Fnu Mohbat and Mohammed J. Zaki. 2024. [Llava-chef: A multi-modal generative model for food recipes](#). In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, CIKM '24*, page 1711–1721, New York, NY, USA. Association for Computing Machinery.
- Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Afina Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, Víctor Gutiérrez-Basulto, Yazmín Ibáñez García, Hwaran Lee, Shamsuddeen Hassan Muhammad, Kiwoong Park, Anar Sabuhi Rzayev, Nina White, Seid Muhie Yimam, Mohammad Taher Pilehvar, and 3 others. 2024. [Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 78104–78146. Curran Associates, Inc.
- Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. [StereoSet: Measuring stereotypical bias in pretrained language models](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online. Association for Computational Linguistics.
- Allen Newell and Herbert A Simon. 2007. Computer science as empirical inquiry: Symbols and search. In *ACM Turing award lectures*, page 1975.
- M. E. J. Newman and M. Girvan. 2004. [Finding and evaluating community structure in networks](#). *Phys. Rev. E*, 69:026113.
- Peter Newmark. 2003. A textbook of translation.
- Tuan-Phong Nguyen, Simon Razniewski, Aparna S. Varde, and Gerhard Weikum. 2022. [Extracting cultural commonsense knowledge at scale](#). *Proceedings of the ACM Web Conference 2023*.
- Tuan-Phong Nguyen, Simon Razniewski, and Gerhard Weikum. 2024. [Cultural commonsense knowledge for intercultural dialogues](#). In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, CIKM '24*, page 1774–1784, New York, NY, USA. Association for Computing Machinery.
- Donald M Nonini. 2013. The local-food movement and the anthropology of global systems. *American Ethnologist*, 40(2):267–275.
- Shramay Palta and Rachel Rudinger. 2023. [FORK: A bite-sized test set for probing culinary cultural biases in commonsense reasoning models](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9952–9962, Toronto, Canada. Association for Computational Linguistics.
- Joon Sung Park, Joseph O’Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22.
- Chantal Pellegrini, Ege Özsoy, Monika Wintergerst, and Georg Groh. 2021. [Exploiting food embeddings for ingredient substitution](#). In *International Conference on Health Informatics*.
- Jeffrey M Pilcher. 2023. *Food in world history*. Routledge.
- Paweł Polak. 2016. Computing as empirical science-evolution as a concept. *Studies in Logic, Grammar and Rhetoric*, 48(1).
- Jean-Pierre Poulain. 2012. The sociology of gastronomic decolonization. *The gaze of the west and framings of the east. Frontiers of Globalization Series. Palgrave Macmillan, London*, pages 218–232.
- Jean-Pierre Poulain, Elise Mognard, Jacqui Kong, Jan Li Yuen, Laurence Tibère, Cyrille Laporte, Fong-Ming Yang, Anindita Dasgupta, Pradeep Kumar Nair, Neethiahnanthan Ari Ragavan, and Ismail Mohd Noor. 2023. [Much more than food: The malaysian breakfast, a socio-cultural perspective](#). *Sustainability*, 15(3).
- Marcel Proust. 1913. *À la recherche du temps perdu*. Grasset.
- Saed Rezayi, Zhengliang Liu, Zihao Wu, Chandra Dhakal, Bao Ge, Haixing Dai, Gengchen Mai, Ninghao Liu, Chen Zhen, Tianming Liu, and Sheng Li. 2025. [Exploring new frontiers in agricultural nlp: Investigating the potential of large language models for food applications](#). *IEEE Transactions on Big Data*, 11(3):1235–1246.
- Paul Rozin. 2005. The meaning of food in our lives: a cross-cultural perspective on eating and well-being. *Journal of nutrition education and behavior*, 37:S107–S112.
- Sougata Saha and Monojit Choudhury. 2025. [User behavior prediction as a generic, robust, scalable, and low-cost evaluation strategy for estimating generalization in LLMs](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 11047–11065, Vienna, Austria. Association for Computational Linguistics.
- Sougata Saha, Saurabh Kumar Pandey, and Monojit Choudhury. 2025. [Meta-cultural competence: Climbing the right hill of cultural awareness](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8025–8042, Albuquerque, New Mexico. Association for Computational Linguistics.

- Yonadav Shavit, Sandhini Agarwal, Miles Brundage, Steven Adler, Cullen O’Keefe, Rosie Campbell, Teddy Lee, Pamela Mishkin, Tyna Eloundou, Alan Hickey, and 1 others. 2023. Practices for governing agentic ai systems. *Research Paper, OpenAI*.
- Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. [The woman worked as a babysitter: On biases in language generation](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3407–3412, Hong Kong, China. Association for Computational Linguistics.
- Weiyang Shi, Ryan Li, Yutong Zhang, Caleb Ziems, Sunny Yu, Raya Horesh, Rogério Abreu De Paula, and Diyi Yang. 2024. [CultureBank: An online community-driven knowledge base towards culturally aware language technologies](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4996–5025, Miami, Florida, USA. Association for Computational Linguistics.
- G. Sproesser, M. B. Ruby, C. S. Akotia, M. dos Santos Alvarenga, R. Bhangaonkar, I. Furumitsu, S. Imada, G. Kaptan, M. Kaufer-Horwitz, C. Fischler, P. Rozin, H. T. Schupp, and B. Renner. 2025. [Food cultures: Understanding traditional and modern eating](#). In H. Meiselman, editor, *Handbook of Eating and Drinking: Interdisciplinary Perspectives*. Springer. In Press.
- Gudrun Sproesser, Matthew B. Ruby, Naomi Arbit, Charity S. Akotia, Marle dos Santos Alvarenga, Rachana Bhangaokar, Isato Furumitsu, Xiaomeng Hu, Sumio Imada, Gülbanu Kaptan, Martha Kaufer-Horwitz, Usha Menon, Claude Fischler, Paul Rozin, Harald T. Schupp, and Britta Renner. 2022. [Similar or different? comparing food cultures with regard to traditional and modern eating across ten countries](#). *Food Research International*, 157:111106.
- John E Staller. 2003. An examination of the palaeobotanical and chronological evidence for an early introduction of maize (*zea mays* l.) into south america: a response to pearsall. *Journal of Archaeological Science*, 30(3):373–380.
- Andrew Steptoe, Tessa M Pollard, and Jane Wardle. 1995. Development of a measure of the motives underlying the selection of food: the food choice questionnaire. *Appetite*, 25(3):267–284.
- C Levi Strauss. 1974. Structural anthropology. *Persona & Derecho*, 1:571.
- David Sutton. 2018. [Cooking in theory: Risky events in the structure of the conjuncture](#). *Anthropological Theory*, 18(1):81–105.
- David E. Sutton. 2010. [Food and the senses](#). *Annual Review of Anthropology*, 39(Volume 39, 2010):209–223.
- Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, and 1 others. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.
- Vincent Antonio Traag, Ludo Waltman, and Nees Jan van Eck. 2018. [From louvain to leiden: guaranteeing well-connected communities](#). *Scientific Reports*, 9.
- Yixin Wan, Jieyu Zhao, Aman Chadha, Nanyun Peng, and Kai-Wei Chang. 2023. [Are personalized stochastic parrots more dangerous? evaluating persona biases in dialogue systems](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9677–9705, Singapore. Association for Computational Linguistics.
- Ruiyi Wang, Haofei Yu, Wenxin Zhang, Zhengyang Qi, Maarten Sap, Yonatan Bisk, Graham Neubig, and Hao Zhu. 2024. [SOTOPIA- \$\pi\$ : Interactive learning of socially intelligent language agents](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12912–12940, Bangkok, Thailand. Association for Computational Linguistics.
- Alan Warde. 2016. *The practice of eating*. John Wiley & Sons.
- Richard Wilk. 2006. *Home cooking in the global village: Caribbean food from buccaneers to ecotourists*. Berg.
- Genta Indra Winata, Frederikus Hudi, Patrick Amadeus Irawan, David Anugraha, Rifki Afina Putri, Wang Yutong, Adam Nohejl, Ubaidillah Ariq Prathama, Nedjma Ousidhoum, Afifa Amriani, Anar Rzayev, Anirban Das, Ashmari Pramodya, Aulia Adila, Bryan Wilie, Candy Olivia Mawalim, Cheng Ching Lam, Daud Abolade, Emmanuele Chersoni, and 32 others. 2025a. [WorldCuisines: A massive-scale benchmark for multilingual and multicultural visual question answering on global cuisines](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3242–3264, Albuquerque, New Mexico. Association for Computational Linguistics.
- Genta Indra Winata, Frederikus Hudi, Patrick Amadeus Irawan, David Anugraha, Rifki Afina Putri, Wang Yutong, Adam Nohejl, Ubaidillah Ariq Prathama, Nedjma Ousidhoum, Afifa Amriani, Anar Rzayev, Anirban Das, Ashmari Pramodya, Aulia Adila, Bryan Wilie, Candy Olivia Mawalim, Cheng Ching Lam, Daud Abolade, Emmanuele Chersoni, and 32 others. 2025b. [WorldCuisines: A massive-scale benchmark for multilingual and multicultural visual question answering on global cuisines](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3242–3264, Albuquerque, New Mexico. Association for Computational Linguistics.

Qing Zhang, Christoph Trattner, Bernd Ludwig, and David Elsweiler. 2019. [Understanding cross-cultural visual food tastes with online recipe platforms](#). *Proceedings of the International AAAI Conference on Web and Social Media*, 13(01):671–674.

Yunke Zhang, Yiran Fan, Peijie Liu, Fengli Xu, and Yong Li. 2025. Cyber food swamps: Investigating the impacts of online-to-offline food delivery platforms on healthy food choices. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, pages 2260–2272.

Huixue Zhou, Lisa S. Chow, Lisa Harnack, Satchidananda Panda, Emily N.C. Manoogian, Minchen Li, Yongkang Xiao, and Rui Zhang. 2025a. [Nutrirag: Unleashing the power of large language models for food identification and classification through retrieval methods](#). *medRxiv*.

Li Zhou, Taelin Karidi, Wanlong Liu, Nicolas Garneau, Yong Cao, Wenyu Chen, Haizhou Li, and Daniel Herscovich. 2025b. [Does mapo tofu contain coffee? probing LLMs for food-related cultural knowledge](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9840–9867, Albuquerque, New Mexico. Association for Computational Linguistics.

Naitian Zhou, David Bamman, and Isaac L. Bleaman. 2025c. [Culture is not trivia: Sociocultural theory for cultural NLP](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25869–25886, Vienna, Austria. Association for Computational Linguistics.

Xuhui Zhou, Zhe Su, Tiwalayo Eisape, Hyunwoo Kim, and Maarten Sap. 2024. [Is this the real life? is this just fantasy? the misleading success of simulating social interactions with LLMs](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21692–21714, Miami, Florida, USA. Association for Computational Linguistics.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and 1 others. 2023. [Sotopia: Interactive evaluation for social intelligence in language agents](#). *arXiv preprint arXiv:2310.11667*.

## A Appendix

### A.1 Case Study Details

The Food Choice Questionnaire (FCQ) (Step-toe et al., 1995) captures food-related behaviors across nine dimensions: familiarity, convenience, health, mood, sensory appeal, natural content, price, weight control, and ethical concern. For

our study, we selected two dimensions: (i) **Convenience**: Represented by the question “*The food in country that can be bought in shops close to where someone lives or works is?*” (ii) **Familiarity**: represented by the question “*The food in country that people usually eat is?*”

We prompted two LLMs - Llama-3.1-8B-Instruct (Dubey et al., 2024) and Gemma-2-9B-it (Team et al., 2024) - for 170 countries ( $C$ ), using a curated list of approximately 3,500 ( $N$ ) food items collected from resources such as TasteAtlas<sup>3</sup> and cross-country cuisine knowledge bases (Winata et al., 2025b). For each prompt, we recorded the log probabilities of tokens corresponding to the  $N$  food items. If a food item spanned multiple tokens, we averaged across them to obtain its final probability. This produced, for each model, an  $C \times N$  matrix ( $M$ ), where each row represents the probability distribution of  $N$  food items for a given country. We derived a country-country weighted adjacency matrix ( $S_{cc}$ ) by computing  $MM^T$ , which encodes the relatedness of countries based on food probabilities. To retain meaningful structure, we build a *mutual*  $k$ -NN graph with  $k = 10$ , and applied a community detection algorithm to identify clusters of countries.

Finally, we curated a mapping between countries and their historical colonizers and used Fisher’s exact test to evaluate whether the LLM-derived food-based communities significantly aligned with colonial histories within continents.

All experiments were run using vLLM (Kwon et al., 2023) with a batch size of 8 and temperature set at 1 on a 48 GB NVIDIA RTX 6000 Ada Generation GPU with FP16 quantization, and took approximately 12 hours to execute. Below is a sample prompt for Llama.

```
Prompt for logprobs prediction for a Question
<|begin_of_text|><|start_header_id|>system<|end_header_id|>

Cutting Knowledge Date: December 2023
Today Date: 26 Jul 2024

<|eot_id|><|start_header_id|>user<|end_header_id|>

Answer the given question related to food items in a country. Generate only one food item. Do not output any extra text. [Question]
<|eot_id|><|start_header_id|>assistant<|end_header_id|>
```

<sup>3</sup><https://www.tasteatlas.com/>

Country	Year	Source	Top Ordered Dishes	URL
Australia	2023	Doordash	Chicken dishes, Pizza, Burgers, Wagyu beef dishes, Chips	<a href="https://downloads.ctfassets.net/trvmqu12jq2l5fTVhyjeP5pNS29PrXNaA5/5c859c8d0567564f7e881692afd0a22b/2024-Restaurant-Alcohol-Online-Ordering-Trends-Report-AU-NZ.pdf">https://downloads.ctfassets.net/trvmqu12jq2l5fTVhyjeP5pNS29PrXNaA5/5c859c8d0567564f7e881692afd0a22b/2024-Restaurant-Alcohol-Online-Ordering-Trends-Report-AU-NZ.pdf</a>
New Zealand	2023	Doordash	Chicken dishes, Chips, Burgers, Naan, Noodle dishes	<a href="https://downloads.ctfassets.net/trvmqu12jq2l5fTVhyjeP5pNS29PrXNaA5/5c859c8d0567564f7e881692afd0a22b/2024-Restaurant-Alcohol-Online-Ordering-Trends-Report-AU-NZ.pdf">https://downloads.ctfassets.net/trvmqu12jq2l5fTVhyjeP5pNS29PrXNaA5/5c859c8d0567564f7e881692afd0a22b/2024-Restaurant-Alcohol-Online-Ordering-Trends-Report-AU-NZ.pdf</a>
USA	2024	Doordash	French fries, Garlic naan, Tacos, Pad thai, Miso soup, California roll, Chicken tikka masala, Edamame, Chips, Burrito	<a href="https://assets.ctfassets.net/trvmqu12jq2l5fTVhyjeP5pNS29PrXNaA5/2c94dfa2b45b5ba60cf844d4dd598617/2025-DoorDash-Delivery-Trends-Report.pdf">https://assets.ctfassets.net/trvmqu12jq2l5fTVhyjeP5pNS29PrXNaA5/2c94dfa2b45b5ba60cf844d4dd598617/2025-DoorDash-Delivery-Trends-Report.pdf</a>
Canada	2024	Doordash	Naan, Butter chicken, Miso soup, California roll, Tandoori roti, Burrito, Dynamite roll, Poutine, Chicken biryani, Steamed rice	<a href="https://downloads.ctfassets.net/trvmqu12jq2l5fTVhyjeP5pNS29PrXNaA5/9117580f0d1b1ef9e49aa0a19ce4eec6/2025-DoorDash-Canada-Delivery-Trends-Report.pdf">https://downloads.ctfassets.net/trvmqu12jq2l5fTVhyjeP5pNS29PrXNaA5/9117580f0d1b1ef9e49aa0a19ce4eec6/2025-DoorDash-Canada-Delivery-Trends-Report.pdf</a>
UAE	2021	Talabat	French fries, chicken sandwich, Burgers, potato wedges, Pizza	<a href="https://hotelandcatering.com/foodservice/talabat-unveils-uaes-most-ordered-food-and-grocery-items-in-2021-and-youll-be-surprised">https://hotelandcatering.com/foodservice/talabat-unveils-uaes-most-ordered-food-and-grocery-items-in-2021-and-youll-be-surprised</a>
Oman	2023	Talabat	Fried Chicken, Burgers, sandwiches	<a href="https://www.zawya.com/en/press-release/companies-news/talabat-unveils-customer-ordering-trends-across-food-groceries-and-donations-throughout-2023-lhwjfcrb">https://www.zawya.com/en/press-release/companies-news/talabat-unveils-customer-ordering-trends-across-food-groceries-and-donations-throughout-2023-lhwjfcrb</a>
India	2024	Zomato	Biryani, Pizza	<a href="https://world.hey.com/bhari/india-food-trends-2024-from-zomato-34ea104e">https://world.hey.com/bhari/india-food-trends-2024-from-zomato-34ea104e</a>

Table 1: Country-wise top-ordered dishes between 2021 and 2024. DoorDash is a U.S.-based food delivery platform operating primarily in North America, Australia, and New Zealand, which publishes annual regional trends reports. Talabat is one of the largest food delivery services in the Middle East and North Africa (UAE, Kuwait, Saudi Arabia, Oman, etc.). Zomato, headquartered in India, is a major platform across South Asia offering food delivery and restaurant discovery. For Talabat and Zomato, statistics were drawn from trusted media reports (e.g., NDTV, Gulf News, Zawya) that cite the platforms’ annual “year-in-review” data, as these platforms do not publish standalone public reports.

Model	Question	Nodes	Edges	# Communities	$Q$	Drawn intra / inter <sup>†</sup>
Llama	Convenience	165	567	10	0.829	482 / 28
	Familiarity	169	568	11	0.831	488 / 23
Gemma	Convenience	166	554	11	0.821	460 / 38
	Familiarity	168	523	10	0.796	426 / 44

Table 2: **Country–country networks.** Mutual  $k$ -NN ( $k=10$ ) on  $S_{cc}=MM^T$ ; communities of size  $< 3$  removed;  $Q$  from Leiden. <sup>†</sup>Edges drawn after the 90% weight quantile filter used *only* for visualization.

**Leiden algorithm** We used the Leiden algorithm (Traag et al., 2018) for community detection. Leiden is an iterative graph partitioning method that optimizes a quality function such as modularity. It improves upon the Louvain algorithm (Blondel et al., 2008) by ensuring that all detected communities are well connected and by modifying the communities to avoid fragmentation.

The algorithm proceeds in three main phases:

1. **Local moving of nodes:** Nodes are moved to neighboring communities if this improves the quality function (e.g., modularity).
2. **Refinement:** Each community is further refined into subcommunities to ensure internal connectivity and prevent the formation of disconnected clusters.
3. **Aggregation:** A new aggregate network is cre-

ated where each refined community becomes a node. The process is then repeated on this coarser network.

These steps are iterated until no further improvements are possible. Unlike Louvain, Leiden ensures that all resulting communities are internally connected, and when run iteratively, it converges to partitions that are locally optimal and free of disconnected components.

**Country-colonizer mapping.** We curate colonizer labels from CEPII’s Geo dataset<sup>4</sup> (geo\_ceprii.xls) and map coordinates from Google’s canonical countries file<sup>5</sup>. For each country, we compute a *colonizer set* as the union

<sup>4</sup>[https://www.cepii.fr/CEPII/en/bdd\\_modele/bdd\\_modele\\_item.asp?id=6](https://www.cepii.fr/CEPII/en/bdd_modele/bdd_modele_item.asp?id=6)

<sup>5</sup>[https://developers.google.com/public-data/docs/canonical/countries\\_csv](https://developers.google.com/public-data/docs/canonical/countries_csv)

of all nonempty entries in `colonizer1..4` and `short_colonizer1..3`. We restrict Fisher tests to within-continent pairs, testing the association between *same colonizer* (nonempty intersection of colonizer sets) and *same community*. Restricting to the same continent controls for geography. Countries on one continent often share ingredients, trade links, and migration routes, so we compare like with like. This checks whether a shared colonizer still predicts being in the same community after accounting for regional similarity. It prevents inflated odds from obvious continent clustering (for example, Europe with Europe).

Community label	Countries	Pairs sharing colonizers count	Colonizers
Europe & European Territories	Austria, Belgium, Czech republic, Denmark, Estonia, Finland, France, French polynesia, Germany, Greenland, Haiti, Hungary, Iceland, Italy, Latvia, Liechtenstein, Lithuania, Luxembourg, Madagascar, Malta, Mauritius, Monaco, New caledonia, Norway, Poland, Russia, Seychelles, Slovakia, Sweden, Switzerland, Ukraine	40	Austria, Denmark, France, Germany, Hungary, Netherlands, Russian federation, Spain, Sweden, United kingdom
South & East Asia	Afghanistan, Bangladesh, Brunei, Cambodia, China, India, Indonesia, Japan, Laos, Malaysia, Maldives, Mongolia, Myanmar, Nepal, North korea, Pakistan, Philippines, Singapore, South korea, Sri lanka, Taiwan, Thailand, Vietnam	53	China, France, India, Japan, Netherlands, Spain, United kingdom, United states of america
Balkans–Caucasus–Anatolia	Armenia, Azerbaijan, Bosnia and herzegovina, Bulgaria, Croatia, Cyprus, Georgia, Greece, Iran, Kazakhstan, Moldova, Montenegro, North macedonia, Romania, Serbia, Slovenia, Türkiye	69	Austria, Greece, Hungary, Russian federation, Serbia and montenegro, Türkiye, United kingdom
West Africa	Benin, Burkina faso, Cameroon, Chad, Democratic republic of the congo, Gabon, Ghana, Ivory coast, Liberia, Mali, Mauritania, Nigeria, Senegal, Sierra leone, The gambia, Togo	61	Belgium and luxembourg, France, United kingdom
Southern & Eastern Africa (Lusophone ties)	Angola, Botswana, Brazil, Cape verde, Ethiopia, Kenya, Lesotho, Mozambique, Namibia, Portugal, Rwanda, Somalia, South africa, South sudan, Tanzania, Zambia	36	Belgium and luxembourg, Egypt, Germany, Italy, Netherlands, Portugal, South africa, United kingdom
Caribbean	Antigua and barbuda, Aruba, Barbados, Bermuda, Bonaire, Curaçao, Dominica, Grenada, Guyana, Jamaica, Netherlands, Saint kitts and nevis, Saint vincent and the grenadines, Suriname, The bahamas, Trinidad and tobago	65	France, Netherlands, Spain, United kingdom
Latin America & Iberia	Andorra, Argentina, Bolivia, Chile, Colombia, Cuba, Dominican republic, Ecuador, Mexico, Paraguay, Peru, Puerto rico, Spain, Uruguay, Venezuela	78	Brazil, France, Haiti, Spain, United states of america
Anglosphere & Oceania	Australia, England, Fiji, Guam, Ireland, Marshall islands, Nauru, New zealand, Northern ireland, Palau, Papua new guinea, Samoa, Scotland, Vanuatu, Wales	31	Australia, France, Germany, Japan, New zealand, United kingdom, United states of america
Middle East	Bahrain, Egypt, Iraq, Israel, Jordan, Lebanon, Oman, Palestine, Qatar, Saudi arabia, Syria, United arab emirates, Yemen	42	France, Poland, Spain, Türkiye, United kingdom
North Africa	Algeria, Morocco, Tunisia	3	France, Türkiye

Table 3: Llama’s convenience communities: countries, shared-colonizer pairs, and colonizers. **Labels are heuristic; they reflect the majority of countries in each cluster.**

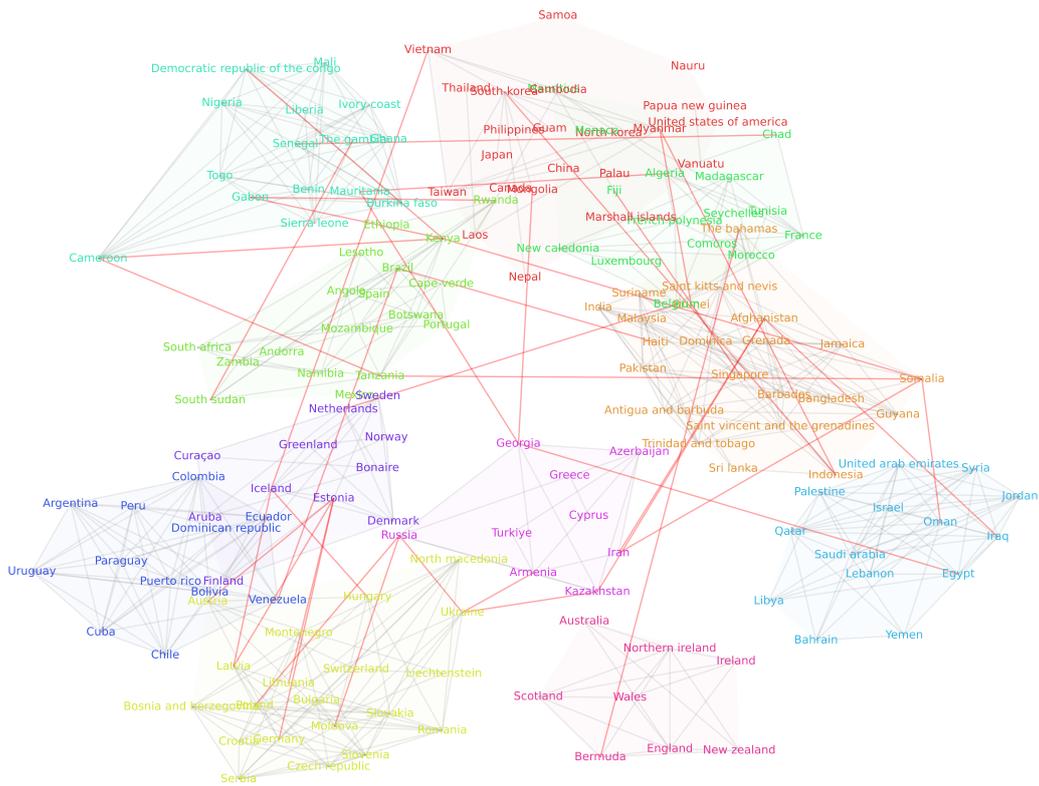


Figure 3: **Country communities from Gemma (convenience).** Weighted mutual- $k$ NN graph ( $k = 10$ ) on  $S_{cc} = MM^T$ ; Leiden partition on modularity. High modularity ( $Q = 0.821$ ) with coherent regional clusters. Intra-community edges shown in gray, inter-community edges in red. Node colors are used solely for visual separation of clusters and do not carry additional semantic meaning.

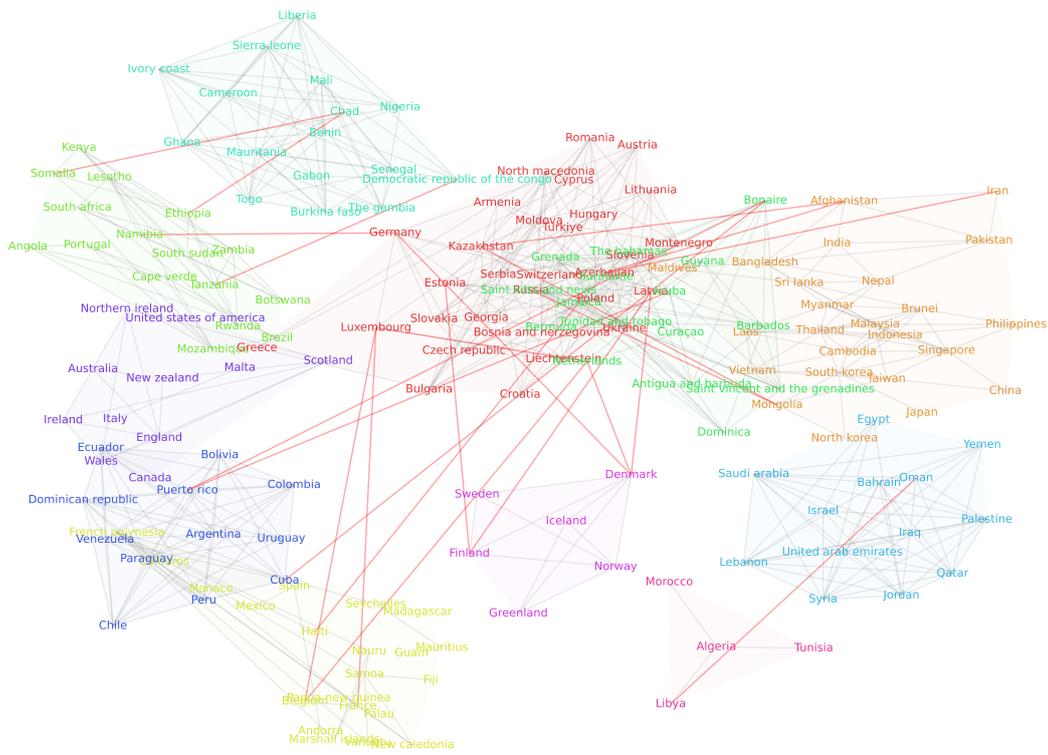


Figure 4: **Country communities from Llama (familiarity).** Weighted mutual- $k$ NN graph ( $k = 10$ ) on  $S_{cc} = MM^T$ ; Leiden partition on modularity. High modularity ( $Q = 0.831$ ) with coherent regional clusters. Intra-community edges shown in gray, inter-community edges in red; light hulls mark communities. Node colors are used solely for visual separation of clusters and do not carry additional semantic meaning.

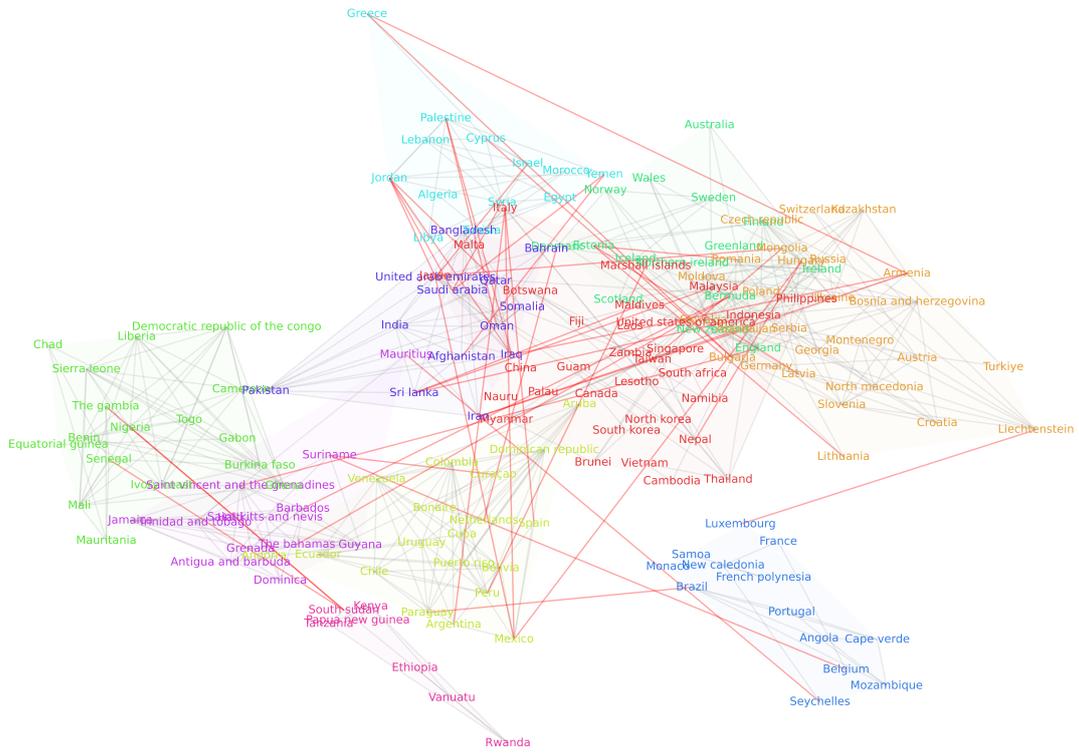


Figure 5: **Country communities from Gemma (familiarity)**. Weighted mutual- $k$ NN graph ( $k = 10$ ) on  $S_{cc} = MM^T$ ; Leiden partition on modularity. High modularity ( $Q = 0.796$ ) with coherent regional clusters. Intra-community edges shown in gray, inter-community edges in red. Node colors are used solely for visual separation of clusters and do not carry additional semantic meaning.

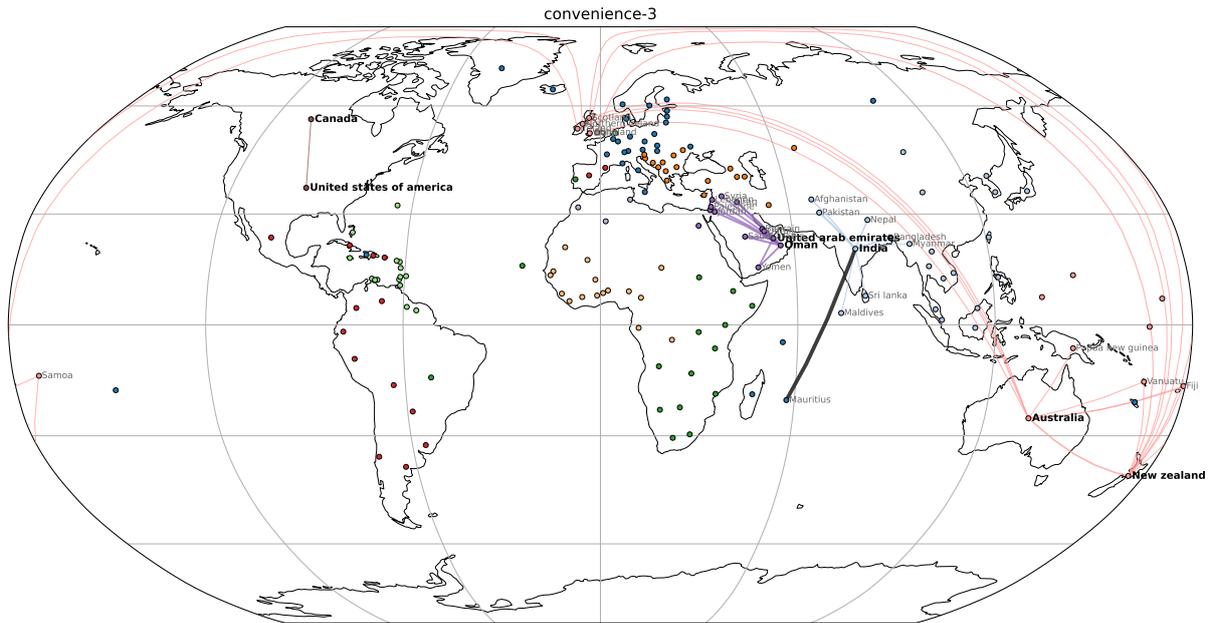


Figure 6: **Country food-similarity graph on world map (Llama, convenience)**. Intra-community edges use the community color with weight-scaled thickness; inter-community edges are dark gray. Labels are shown only for focal countries and their neighbors. This map projection highlights how model-inferred communities align with geographic and colonial patterns, while inter-community ties remain sparse. Node colors are used solely for visual separation of clusters and do not carry additional semantic meaning.

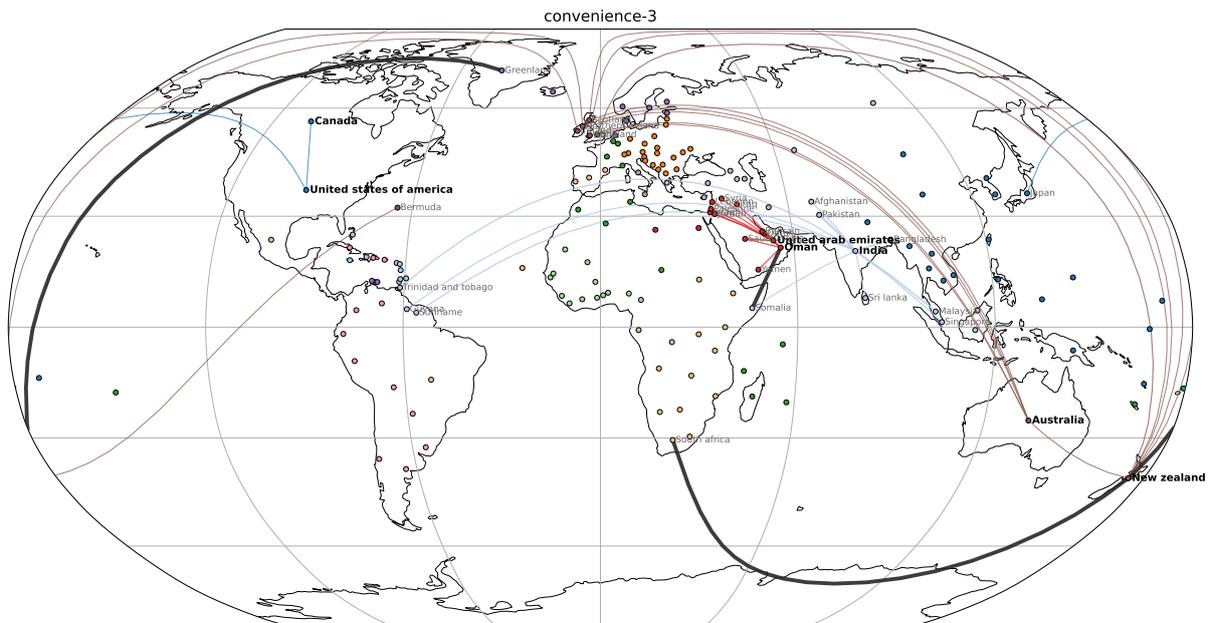


Figure 7: **Country food-similarity graph on world map (Gemma, convenience)**. Node colors are used solely for visual separation of clusters and do not carry additional semantic meaning.

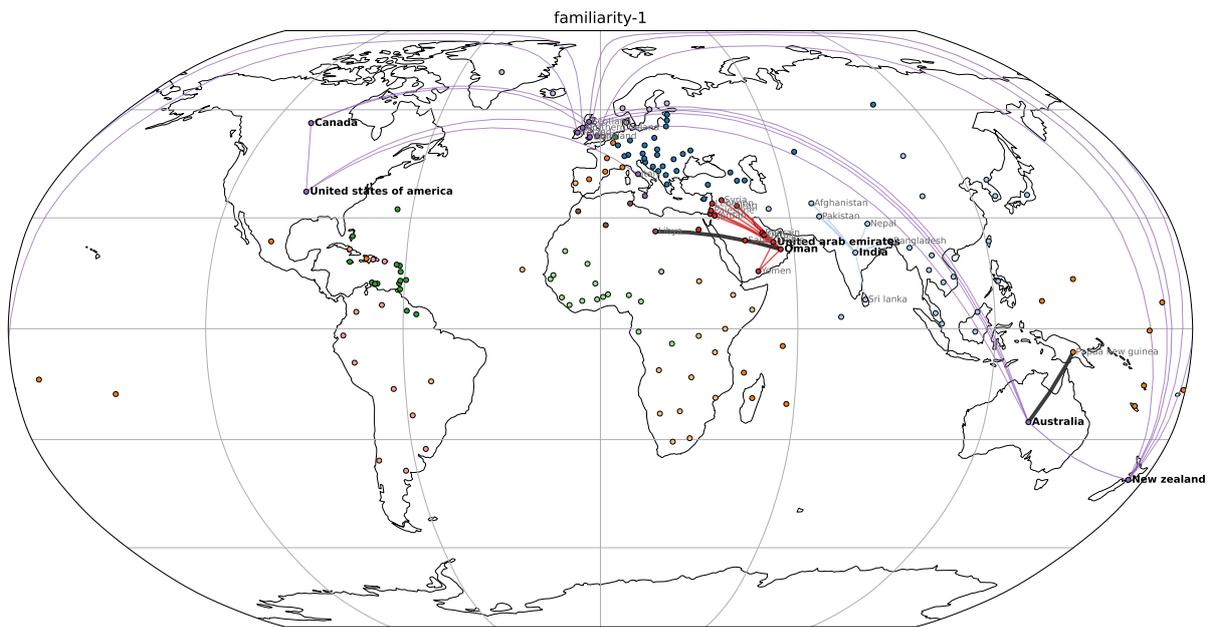


Figure 8: **Country food-similarity graph on world map (Llama, familiarity).** Node colors are used solely for visual separation of clusters and do not carry additional semantic meaning.

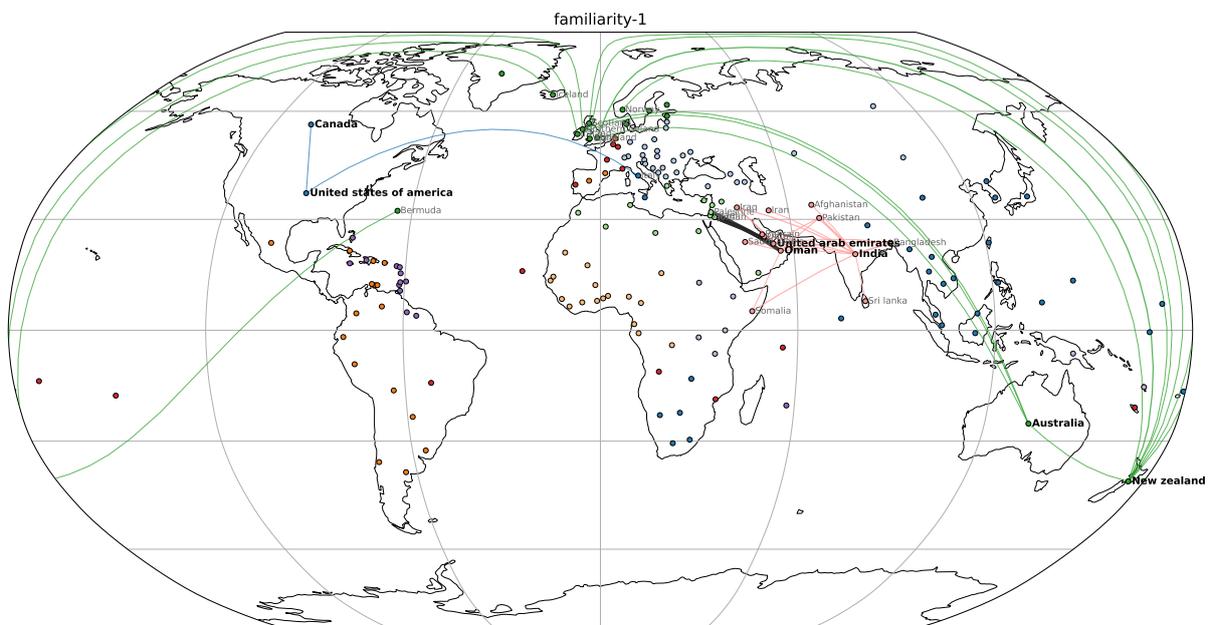


Figure 9: **Country food-similarity graph on world map (Gemma, familiarity).** Node colors are used solely for visual separation of clusters and do not carry additional semantic meaning.

# LLM-as-a-qualitative-judge: automating error analysis in natural language generation

Nadezhda Chirkova<sup>1</sup> Tunde Oluwaseyi Ajayi<sup>2</sup> Seth Aycok<sup>3</sup> Zain Muhammad Mujahid<sup>4</sup>  
Vladana Perlić<sup>5,6</sup> Ekaterina Borisova<sup>7,8</sup> Markarit Vartampetian<sup>9</sup>

<sup>1</sup>Naver Labs Europe

<sup>2</sup>Insight Research Ireland Centre for Data Analytics, Data Science Institute, University of Galway

<sup>3</sup>University of Amsterdam <sup>4</sup>University of Copenhagen <sup>5</sup>STMicroelectronics <sup>6</sup>Télécom Paris

<sup>7</sup>Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI)

<sup>8</sup>Technische Universität Berlin

<sup>9</sup>Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France

**Correspondence:** [nadia.chirkova@naverlabs.com](mailto:nadia.chirkova@naverlabs.com)

## Abstract

Prompting large language models (LLMs) to evaluate generated text, known as *LLM-as-a-judge*, has become a standard evaluation approach in natural language generation (NLG), but is primarily used as a *quantitative* tool, i.e. with numerical scores as main outputs. In this work, we propose *LLM-as-a-qualitative-judge*, an LLM-based evaluation approach with the main output being a *structured report* of *common issue types* in the NLG system outputs. Our approach is targeted at providing developers with meaningful insights on what improvements can be done to a given NLG system and consists of two main steps, namely open-ended per-instance issue analysis and clustering of the discovered issues using an intuitive cumulative algorithm. We also introduce a strategy for evaluating the proposed approach, coupled with ~300 annotations of issues in instances from 12 NLG datasets. Our results show that instance-specific issues output by *LLM-as-a-qualitative-judge* match those annotated by humans in 2/3 cases, and that *LLM-as-a-qualitative-judge* is capable of producing error type reports resembling the reports composed by human annotators. We also demonstrate in a case study how the use of *LLM-as-a-qualitative-judge* can substantially improve NLG systems performance. Our code and data are publicly available<sup>1</sup>.

## 1 Introduction

Prompting large language models (LLMs) to output evaluation scores, known as *LLM-as-a-judge* (Chiang et al., 2023; Zheng et al., 2023a), has become a standard approach for evaluating performance in natural language generation (NLG) tasks. In contrast to classic statistical measures such as BLEU

<sup>1</sup>Code & data: <https://github.com/tunde-ajayi/llm-as-a-qualitative-judge>

Sentence-level detoxification			
GPT4o report		Human annotator's report	
Issue type	Cnt	Issue type	Cnt
Inadequate Toxicity Mitigation	12	Sentence not detoxified	7
Language Preservation Failure	6	Generation in En instead of Ru	6
Contextual Meaning Alteration	5	Changed meaning of a sentence	5
Extraneous Commentary Inclusion	4	Extra comment	4
Inconsistent Task Refusal	2	LLM refuses to answer	6
Tone Modification Through Omission	1	A part of a sentence is omitted	3
Non-Toxicity Commentary Deviation	1	Reply instead of detoxification	2

Biomedical question answering			
GPT4o report		Human annotator's report	
Issue type	Cnt	Issue type	Cnt
Format Discrepancy Mismatch	13	No verbatim match vs ground truth	12
Incorrect Entity Substitution	5	Different list vs ground truth	3
Semantic Variance Mismatch	4	Different formulation of an answer vs ground truth	3
Incomplete Response Composition	3	Misinterpretation of the retrieved document	2
Irrelevant Retrieval Error	1	Irrelevant retrieval	2

Figure 1: Issue types reports for two datasets composed by the proposed *LLM-as-a-qualitative-judge* (GPT-4o) and by a human annotator. All steps of analysis performed by GPT-4o, including error types formulation and error grouping. The full generated report also includes comprehensive error type descriptions, omitted here due to the space limit. Cnt represents issue type counts.

(Papineni et al., 2002), ROUGE (Lin, 2004), or METEOR (Banerjee and Lavie, 2005), which primarily rely on surface-level lexical overlap, LLM-as-a-judge evaluates text based on deep semantic understanding, allowing it to better handle diverse phrasings that convey equivalent meanings. Consequently, it shows stronger alignment with human judgment in various tasks, including machine translation (Kocmi and Federmann, 2023), summarization (Clark et al., 2023), or open-ended instruction following (Ye et al., 2024), especially with strong recent LLMs.

Recent works propose various extensions of the LLM-as-a-judge approach, including pairwise model comparison (Zheng et al., 2023b), finetuning

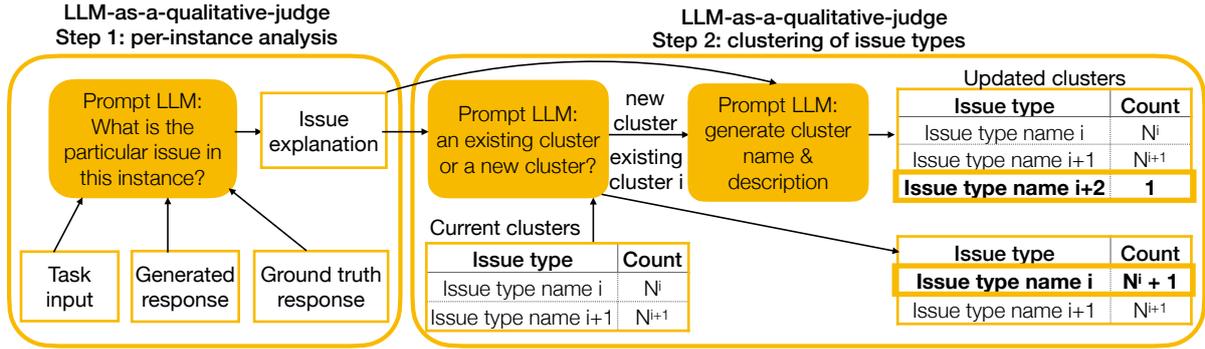


Figure 2: Illustration of the proposed LLM-as-a-qualitative-judge approach.

LLMs for evaluation (Kim et al., 2024b,c), multi-criteria evaluation (Liang et al., 2023; Fu et al., 2024), the dynamic selection of evaluation criteria (Ye et al., 2024), or even using per-instance evaluation checklists (Cook et al., 2024; Kim et al., 2025). A common technique to improve LLM-as-a-judge is to ask a model to output an explanation for the predicted score(s).

However, even with the extensions listed above, LLM-as-a-judge remains primarily a *quantitative* evaluation tool, i.e., the final result used by researchers and practitioners is *quantitative evaluation scores*. At the same time, language generation is a complex multi-faceted task with a vast space of potential issues, including in various aspects of generated texts (grammaticality, factuality, logical coherence, etc.), in preprocessing of the input data and postprocessing of the NLG outputs, or even with user requests. An effective and commonly used strategy for spotting such issues is a manual *qualitative error analysis* of a subset of predictions, which allows developers to identify artifacts, fix system issues, and detect flaws in quantitative evaluation. Yet this analysis is often skipped in practice (van Miltenburg et al., 2023, 2021), due to overreliance of developers on quantitative metrics, as well as high demand in terms of time and effort needed to conduct such analysis.

In this work, we introduce *LLM-as-a-qualitative-judge*, a novel approach which automates error analysis, with the main output being a *structured report* aggregating the common *qualitative* error types in the NLG outputs for a given dataset. The two key steps of LLM-as-a-qualitative-judge are (1) open-ended per-instance error analysis and (2) clustering of the discovered error types. Per-instance analysis implies prompting an LLM to detect an issue in the given NLG system output, where an issue may be arbitrary, i.e. we do not provide any prede-

efined set of possible issues. For error clustering, we propose an intuitive and effective algorithm which resembles how humans solve the corresponding task. Examples of the generated report are presented in Figure 2 (left) and Appendix H, and a high-level illustration of the proposed approach is presented in Figure 2 (right).

To summarize, our contributions are three-fold:

- We introduce an *LLM-as-a-qualitative-judge*, an approach for LLM-based evaluation, outputting a structured report of common error types in a dataset;
- In a case study on BigBenchHard tasks, we demonstrate that LLM-as-a-qualitative-judge can substantially improve the performance of NLG systems;
- We collect  $\sim 300$  manual annotations of open-ended issues in the instances coming from 12 diverse NLG datasets, as well as the manual annotations of their per-dataset clustering;
- We introduce a strategy for meta-evaluating *LLM-as-a-qualitative-judge* and show that *LLM-as-a-qualitative-judge* is capable of producing error type reports which resemble the reports produced by humans.

We hope that the proposed *LLM-as-a-qualitative-judge* approach will reduce the time and effort required for issue analysis and will help practitioners to more easily improve their NLG pipelines. Our code and data are available as <https://github.com/tunde-ajayi/llm-as-a-qualitative-judge>.

## 2 Proposed approach

The main goal of our proposed approach, *LLM-as-a-qualitative-judge*, is to provide a developer with

a *structured report* of the main *types of issues* (and their counts) in the outputs of a given NLG system for a given dataset. In the rest of the work, we use terms *issues*, *errors*, or *failures* interchangeably to denote any problems which may occur in the NLG outputs. Examples of such problems include (but are not limited to) unfinished generation due to reaching the maximum new tokens limit, an error in one of the reasoning steps, a problem with the retrieved documents in retrieval-augmented generation, or an error in evaluation due to the use of an inappropriate metric. We do not employ any predefined set of possible issues, and use the term *open-ended issue analysis* to refer to the problem of detecting arbitrary issues in NLG outputs.

For the purposes of our algorithm, the dataset consists of *instances*, each represented by a *task input* (a string containing a task instruction and the input data), a *ground truth response* (a string defining a correct answer), and a *generated response* (a final output of the NLG system). Each instance can be optionally augmented with other fields, e.g., the intermediate outputs of an NLG system such as retrieved documents in retrieval-augmented generation, or additional information on the NLG system, e.g., a definition of a task metric. The *LLM-as-a-qualitative-judge* algorithm is summarized in Algorithm 1, illustrated in Figure 2 and described step-by-step below.

**Preliminary step: detecting examples with errors.** Our algorithm focuses only on the instances from the dataset with any sort of issues. We rely on the task-specific quantitative metric to select such instances, i.e., instances which did not get high scores in the quantitative evaluation.

**Step 1: per-instance analysis.** For each instance, we prompt an LLM to identify “*one, most important, specific, clearly visible issue*”, provided with a task input, a ground truth response, a generated response, and optionally other fields as described above. We prompt an LLM to output a detailed analysis of a given instance, followed by a special separator and a final 1–2 sentence description of an identified issue, which is referred to as a *per-instance issue explanation* in the following steps of the approach. The particular prompt used for per-instance analysis is presented in App. Figure 7.

**Step 2: issue clustering.** The second step in *LLM-as-a-qualitative-judge* consists of clustering issues discovered in the first step and forming a

---

### Algorithm 1 LLM-as-a-qualitative-judge

---

**Input:** a list of task inputs  $U$ , a list of ground truth responses  $R^{\text{gt}}$ , a list of generated responses  $R$  — all of length  $N$ ;  
**Output:** a report  $C$  listing issue types and their counts;  
 a list  $A$  of per-instance issue explanations

```

1:  $A \leftarrow \square$  // empty initialization for per-instance analysis
2:  $C \leftarrow \square$  // empty initialization for a report
3: for  $i = 1, \dots, N$  do
4:   // per-instance analysis
5:    $A[i] \leftarrow \text{LLM}(\text{prompt}_{\text{analysis}}; U[i], R^{\text{gt}}[i], R[i])$ 
6:   //  $A[i]$  is a string containing issue explanation
7:   // report generation
8:   if  $i > 1$  then
9:     // an existing issue type or a new one?
10:     $K \leftarrow \text{LLM}(\text{prompt}_{\text{decision}}; A[i], C)$ 
11:    //  $K \in \{1, \dots, |C|, \text{None}\}$ 
12:   else
13:     $K \leftarrow \text{None}$  // first step is always new issue type
14:   if  $K$  is None then
15:     // create a new issue type
16:     $E \leftarrow \text{LLM}(\text{prompt}_{\text{new\_type}}; A[i])$ 
17:    //  $E$  is a dictionary containing a short issue name
    // and an issue description
18:     $E[\text{"count"}] \leftarrow 1$ 
19:     $C.\text{append}(E)$ 
20:   else
21:     // augment an existing issue type
22:     $C[K][\text{"count"}] \leftarrow C[K][\text{"count"}] + 1$ 
23: return  $C, A$ 

```

---

final report of main issue types based on the clustering results. This can be, in principle, done with any clustering algorithm, e.g., k-means with BERT-based embeddings (Devlin et al., 2019) or *directly prompting* a strong LLM to output clustering (Viswanathan et al., 2024), provided with clustering inputs in a single prompt. In the experiments, we demonstrate the downsides of these approaches, e.g., classic approaches perform poorly on our data, and clustering with direct prompting fails for larger datasets, weaker LLMs, and does not ensure the structural correctness of the generated report.

Inspired by how humans would cluster issues, we propose an intuitive *cumulative* issue clustering algorithm. Our clustering algorithm goes through instances one-by-one and gradually builds the issue types report. For each instance, we provide an LLM with the current report and the per-instance issue explanation, and prompt the LLM to *decide* whether this issue explanation can be attributed to one of the already discovered issue types (clusters) or it should form a new issue type (cluster). In the former case, we augment the counter of the corresponding issue type by one. In the latter case, we also prompt an LLM to formulate a short name and a 1–2 sentence description of a new error type, based on the per-instance issue description. In particular, we instruct an LLM to formulate “*a fine-*

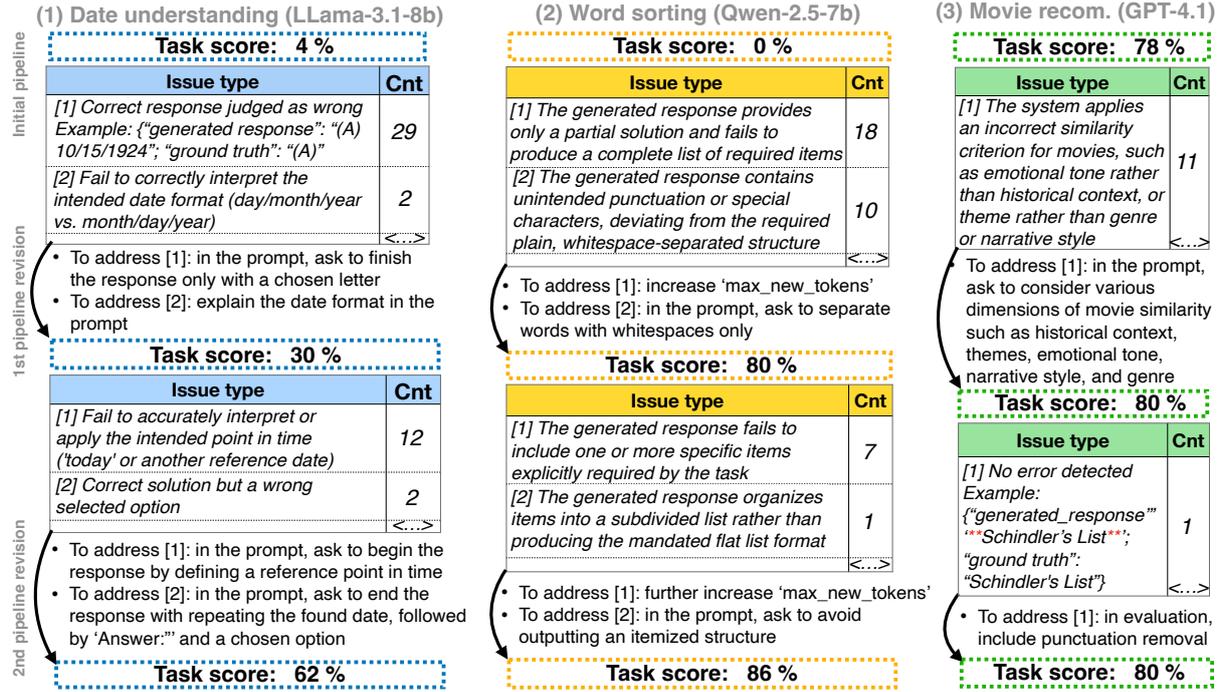


Figure 3: Case study on three BigBenchHard tasks: after building a simple pipeline for a task, we perform two rounds of generating issue reports with LLM-as-a-qualitative-judge (a table with issue types and their counts) and manually revising the pipeline based solely on the generated reports. Task performance is improved in all cases.

grained issue type that can be generalized to other instances”. The new issue type is then added to the report, represented by the generated issue type name, description, and the counter set to one. The first instance in the dataset is always a new issue type. Appendix Figures 9 and 10 show the prompts used for the two described clustering steps.

The final issue types report is composed of issue type names and descriptions, paired with the counts of how many instances were attributed to the corresponding issue type.

### 3 Case study

Our first set of experiments is targeted at demonstrating a practical utility of the proposed *LLM-as-a-qualitative-judge*.

**Experimental setup.** We pick three tasks from a BigBenchHard collection (Suzgun et al., 2022), namely Date understanding, Word sorting, and Movie recommendation. For each dataset, we build a simplest pipeline consisting of prepending a simple system prompt “You are a helpful assistant. Output your answer after a final separator ‘Answer:’”, LLM generation with default hyperparameters, and a string match-based evaluation function. We then perform two rounds of generating an issue report with *LLM-as-a-qualitative-judge* (GPT-4.1)

and manually revising the pipeline solely based on the generated report (issue types, their counts, and possibly 1 example of each issue type). More details are given in Appendix B.

**Results.** As shown in Figure 3, the task performance is improved in all three cases. For example, in the Date understanding task, revisions inspired by the generated issue reports include explaining the date format in the prompt, suggesting to begin the response with determining a reference point in time, and providing a specific template for the output. These revisions improved performance from 4% to 62%.

### 4 Meta-evaluation methodology

This section described a methodology that we propose to meta-evaluate *LLM-as-a-qualitative-judge*. The corresponding set of experiments aims both to assess the effectiveness of two steps and to identify the optimal configurations for *LLM-as-a-qualitative-judge*.

**Real-world data.** We manually annotate per-instance issues and their per-dataset clustering for a diverse pool of 12 datasets, with various open-source LLMs as generators. We consider 7 generative tasks, and for one of the tasks, namely retrieval-

Task	Dataset reference	# ex.
<b>Natural Language Generation</b>		
Instruction following	FLASK (Ye et al., 2024)	34
Translation en-ru	WMT'22 (Kocmi et al., 2022)	38
Long context QA	Elitr-Bench (Thonet et al., 2025)	26
Semantic parsing	PIZZA (Arkoudas et al., 2022)	34
Grade school math	GSM8K (Cobbe et al., 2021)	17
Detoxification	ParaDetox (Dementieva et al., 2024)	36
<b>Retrieval-augmented QA</b>		
Factoid QA in Russian	MKQA (ru) (Longpre et al., 2021)	29
Biomedical QA	BioASQ (Krithara et al., 2023)	27
Lifestyle forum QA	RobustQA (Han et al., 2024; Santhanam et al., 2022)	21
Search engine queries	SearchQA (Dunn et al., 2017)	13
Educational QA	SyllabusQA (Fernandez et al., 2024)	9
Total		297

Table 1: The statistics of the annotated evaluation data.

augmented question answering (RA-QA), we consider 6 domains. All the labeling was performed by the authors of the paper. The final dataset comprises 297 instances. Table 1 provides the data summary. More details on data annotation are presented in Appendix A.

**Synthetic data.** We also consider synthetic data for evaluating clustering: we define a set of possible issue types  $e$  and their frequencies  $n_e$ , then prompt GPT-4o to reformulate each issue  $e$  in various ways  $n_e$  times, and then use this data as per-instance analysis for clustering. This allows us to evaluate clustering on larger datasets, i.e., 100-1000 instances.

**Metrics.** For per-instance analysis, we prompt an *evaluator LLM* to judge whether the issue explanation determined by the *LLM-as-a-qualitative-judge* for a particular instance matches the issue determined by the human annotator. The outputs from the *evaluator LLM* are binary and are accumulated into a *per-instance analysis accuracy* score.

We evaluate cluster agreement using a *Rand index adjusted for chance*, or Adjusted Rand Index (ARI, the higher the better)<sup>2</sup>. We also evaluate the agreement in error type descriptions, by finding the best possible mapping between clusters found by a human annotator and by *LLM-as-a-qualitative-judge*, and then prompting an *evaluator LLM* to judge the semantic equivalence of the corresponding issue type descriptions. This metric is denoted as Semantic Label Consistency (SLC).

<sup>2</sup>We use the scikit learn implementation: [https://scikit-learn.org/stable/modules/generated/sklearn.metrics.adjusted\\_rand\\_score.html](https://scikit-learn.org/stable/modules/generated/sklearn.metrics.adjusted_rand_score.html).

## 5 Meta-evaluation experiments

### 5.1 Experimental setup

We test per-instance analysis with a range of commercial and open-source LLMs, and issue clustering with three LLMs: GPT-4o, Gemini-2-Flash, and Qwen-2.5-7B. For issue clustering, we compare the proposed cumulative clustering approach to the direct LLM prompting and classic clustering approaches. All clustering runs operate on the per-instance analysis output by GPT-4o, and clustering results are averaged over 3 runs.

Tables 4 and 5 in Appendix list references and licenses of the used models and datasets, respectively. Prompts used for all the stages are presented in the Appendix G. Exact task formulations in prompts were adjusted by only using three RA-QA datasets (MKQA (ru), RobustQA Lifestyle and Writing).

For classic clustering approaches, we use the `scikit-learn` implementation with BERT embeddings and tune hyperparameters as described in App. C. For methods requiring the number of clusters, we set it the same as in the annotator’s data.

More details on experimental settings and as well as results on meta-evaluating the evaluator LLM are presented in Appendix B.

Appendix H presents the examples of generated error type reports, per-instance issue explanations, and confusion matrices for all considered datasets. We also provide an archive with experimental results in the project repository<sup>3</sup>.

### 5.2 Per-instance analysis

Table 2 reports the performance of various LLMs in per-instance analysis. Strongest LLMs, including commercial LLMs and a larger open-source Qwen-2.5-32B achieve an accuracy of 62–67%, i.e. about 2/3 of issues in our dataset were successfully correctly explained by strong models. The accuracy of open-source LLMs is substantially influenced by their size: Qwen-2.5 accuracy raises from 32% to 67% when increasing size from 1.5B to 32B. Various LLMs of 7–8B size demonstrate analysis accuracy of 42–60%.

We note that our results are consistent with previously reported findings in the literature regarding the typical level of agreement between LLM-based evaluations and human judgments. For example, FLASK reports the highest correlation between

<sup>3</sup>A zip archive with experimental results is available at <https://github.com/tunde-ajayi/llm-as-a-qualitative-judge/tree/main/data>

Model	Accuracy (%)
GPT-4o	66.3
Gemini-2.0-Flash	65.0
Qwen-2.5-32B	68.7
Qwen-2.5-7B	55.5
Qwen-2.5-1.5B	30.7
DeepSeek-R1-Distill-Llama-8B	56.1
Aya-Expansive-8B	42.1
Llama-3.1-8B-Instruct	55.4
Minstral-8B-Instruct-2410	58.1

Table 2: Performance of various LLMs in per-instance analysis. Evaluator LLM: Claude-3-7-Sonnet-20250219.

model-based evaluation and human labelers, of 68% (Table 1 in Ye et al., 2024), or METAL reports the highest agreement between the LLM evaluators and human scores of 59-82% for the first three criteria in (Hada et al., 2024, Table 3, English).

*To sum up, our results demonstrate the high effectiveness of strong LLMs in open-ended issue explanation for generative tasks. For practical applications, we recommend using recent models such as GPT-4o, Gemini-2.5-Flash or Qwen-2.5-32B.*

### 5.3 Examples of per-issue analysis

Figure 4 shows examples of issue explanations generated by GPT-4o and Qwen-2.5-7B. In the manual inspection of the generated issue explanations, we observed correct explanations for various kinds of issues, and rows 1–3 demonstrate such examples.

We also notice three groups of mistakes. The first occasional problem in per-instance analysis is logical issues. In the example in row 4, the issue is that the ground truth response is not contained as a verbatim substring in the model-generated response, which is a definition of a task metric. However, both GPT-4o and Qwen-2.5-7B claim that the failure in the substring match is caused by the model response containing extra generated information. Such an explanation logically contradicts the task metric, i.e., extra content can only increase chances of finding a given substring in the response, but cannot be a reason for its absence.

The second occasional problem in per-instance analysis is the oversimplification of an issue, especially for more unexpected issues, such as an ambiguous task input or an error in evaluation. In the example in row 5, the issue is an ambiguous user question, i.e., both the ground truth and the generated response are correct and provide two different interpretations of the user question. However, GPT-4o and Qwen-2.5-7B report the over-simplified is-

Approach	Cluster assignment		Cluster descriptions	
	ARI <sub>real</sub> ↑	ARI <sub>syn</sub> ↑	SLC <sub>real</sub> ↑	SLC <sub>syn</sub> ↑
GPT-4o				
Cumulative	0.14 $\pm$ .05	0.73 $\pm$ .05	0.33 $\pm$ .10	0.70
Direct	0.15 $\pm$ .05	0.63 $\pm$ .04	0.42 $\pm$ .12	0.62
Gemini				
Cumulative	0.13 $\pm$ .05	0.70 $\pm$ .02	0.32 $\pm$ .12	0.71
Direct	0.17 $\pm$ .04	0.83 $\pm$ .01	0.42 $\pm$ .11	0.68
Qwen-2.5-7B				
Cumulative	0.11 $\pm$ .04	0.50 $\pm$ .07	0.41 $\pm$ .16	0.44
Direct	0.07 $\pm$ .05	0.01 $\pm$ .02	0.32 $\pm$ .11	0.12
K-means	0.05 $\pm$ .05	0.44 $\pm$ .08	n/a	n/a
Agglomerative	0.05 $\pm$ .00	0.49 $\pm$ .00	n/a	n/a
GMM	0.04 $\pm$ .03	0.41 $\pm$ .05	n/a	n/a
HDBSCAN	0.01 $\pm$ .02	0.13 $\pm$ .03	n/a	n/a

Table 3: Performance of various approaches and LLMs in issue clustering. Results averaged over 3 runs from different random seeds. Agreement in cluster assignment measured using Adjusted Rank Index (ARI) and in cluster descriptions using LLM-judged Semantic Label Consistency (SLC), both metrics the higher the better. Subscripts<sub>real</sub> and<sub>syn</sub> indicate tests on the real and synthetic data respectively. “N/a” indicates the metric is not applicable since classic approaches do not generate cluster names.

sue of the generated response not providing the same answer as the ground truth response.

Finally, the third occasional reason for a per-instance issue explanation not being accepted by the evaluator LLM, is the subjectivity of some issues in the dataset. For example, a human-annotated issue in row 6, “The generation was stopped too early because of the reached maximum new tokens limit”, is evaluated to be not equivalent to the LLM-as-a-qualitative-judge-generated issue “The generated response provides an incomplete overview <...>”. While these two issue explanations are indeed different, they are both correct, and the LLM-generated explanation follows from the annotator’s explanation.

To alleviate potential negative effects from erroneous issue explanations, we recommend developers to check a couple of examples of each issue, which are output by LLM-as-a-qualitative-judge in addition to the issue names and descriptions.

### 5.4 Issue clustering

Table 3 reports performance in issue clustering, for three LLMs. As described in Section 4, we evaluate clustering on both real and synthetic data. We find that clustering via direct prompting performs well for small datasets and strong LLMs, but fails for weaker LLM, e.g. Qwen-2.5-7B, and/or larger

Annotator's issue explanation	GPT-4o issue explanation	Qwen-2.5-7B issue explanation
1: An LLM loops in repeating the same link multiple times <i>(FLASK)</i>	The LLM response failed due to incorrect and duplicate URLs for different languages, indicating inaccurate retrieval and differentiation of relevant StackOverflow questions <b>Correct</b>	The primary issue is that the generated answer does not provide the correct URLs as specified in the ground truth labels, instead offering descriptions of how to remove items from arrays in different programming languages. <b>Wrong</b>
2: An LLM replied to a given sentence instead of detoxifying it. <i>(MultiDetox)</i>	The failure stems from the LLM generating an admonishing response instead of a refined paraphrase of the given toxic sentence. <b>Correct</b>	The primary failure is that the LLM-generated response does not adhere to the task requirement of refining the input sentence into a new, similar-meaning sentence, instead offering advice on language use. <b>Correct</b>
3: A mistake in a retrieved document which is propagated to the LLM response <i>(SearchQA)</i>	The system failed due to misleading information in the retrieved documents, where a fragmented sentence implied a false connection between President Felipe Calderon and Nicaragua instead of Mexico <b>Correct</b>	The primary issue is the incorrect identification of the country associated with President Felipe Calderón, likely due to the LLM retrieving and misinterpreting information that linked him to Nicaragua instead of Mexico. <b>Correct</b>
4: The target name is written in English rather than in Russian, while the ground truth is written in Russian, resulting in the absence of substring match <i>(MKQA (ru))</i>	The generated response includes extra information about Instagram's own account, which prevents it from matching the ground truth label focused solely on Cristiano Ronaldo. <b>Wrong</b>	The generated response included unnecessary details about the Instagram account's follower count, which caused it to fail the substring match criterion despite being factually correct. <b>Wrong</b>
5: Ambiguous question: the generated response and the ground truth provide two ways to answer the question. <i>(BioASQ)</i>	The LLM-generated response failed to include the exact term "Leishmania species" required for a substring match with the ground truth. <b>Wrong</b>	The generated answer failed to include the term "Leishmania Species," which is required by the ground truth label for the substring match to succeed. <b>Wrong</b>
6: Generation was stopped too early because of the reached maximum new tokens limit <i>(RobustQA Lifestyle)</i>	The generated answer provides an incomplete and misaligned overview compared to the detailed options and context provided in the ground truth. <b>Wrong</b>	The generated response fails to include key information about Just Right Menus, MacGourmet, SousChef, Yum, and the use of simple text files with Dropbox synchronization, explicitly mentioned in the ground truth labels <b>Wrong</b>

Figure 4: Examples of per-instance analysis.

datasets. For example, ARI reached by GPT-4o on the synthetic data drops from 1 to 0.05 when increasing the dataset size from 100 to 1000 instances. The proposed cumulative clustering demonstrates greater robustness and reaches high ranges of ARI in all cases. In addition, the proposed cumulative algorithm outputs correctly structured summaries by design, while the structural correctness of clustering with direct prompting is not guaranteed.

Comparing LLMs, we find that Gemini and GPT-4o reach highest scores in both cluster assignment and cluster descriptions generation, followed Qwen-2.5-7B. Classic clustering approaches reach rather low values of ARI.

Figure 5 demonstrates examples of confusion matrices for several datasets. Pronounced diagonals and matching cluster names illustrate the strong capabilities of *LLM-as-a-qualitative-judge* to output issue types reports that resemble the issue reports produced by humans. Due to the inherent subjectivity of clustering task, we observe occasional merging or splitting of annotator's clusters, e.g. clusters "Wrong topping" and "Wrong variables" were merged by *LLM-as-a-qualitative-judge* into one cluster "Entity Mislabeling" for the Pizza ordering dataset.

*To sum up, our results demonstrate the effectiveness of the proposed cumulative clustering approach to produce issue reports that resemble the ones produced by humans. For practical applications, we recommend using recent models such as GPT-4o or Gemini-2.5-Flash.*

## 6 Discussion

In this section, we discuss potential extensions of the proposed approach.

**Issues prefiltering.** As discussed in Section 2, *LLM-as-a-qualitative-judge* operates only over instances which received low scores from a quantitative task metric. Such prefiltering could in principle be removed and incorporated directly into per-instance analysis by modifying its prompt, e.g. "Output what is an issue with this instance. If there is no issue, output 'No issue'". Instances with predictions "No issue" then would be discarded from issue clustering. However, in preliminary experiments we found that LLM are prone to making up issues for fully correct instances. Hence, we do not recommend removing the prefiltering step (at least with the current state of LLMs), which is a reasonable design since *LLM-as-a-qualitative-judge* is an *error analysis* method.

**Multiple issues per instance.** *LLM-as-a-qualitative-judge* could be easily extended to detect multiple issues per instance, by modifying the prompt used for per-instance analysis and going through the generated issues one-by-one in the issue clustering step. However, same as with a previous discussion point, in our preliminary experiments we found that LLMs are prone to generating non-existing issues in such a scenario. For example, GPT-4o tends to generate a constant number of issues for any instance (in particular, 3). In practice, we believe that our design with

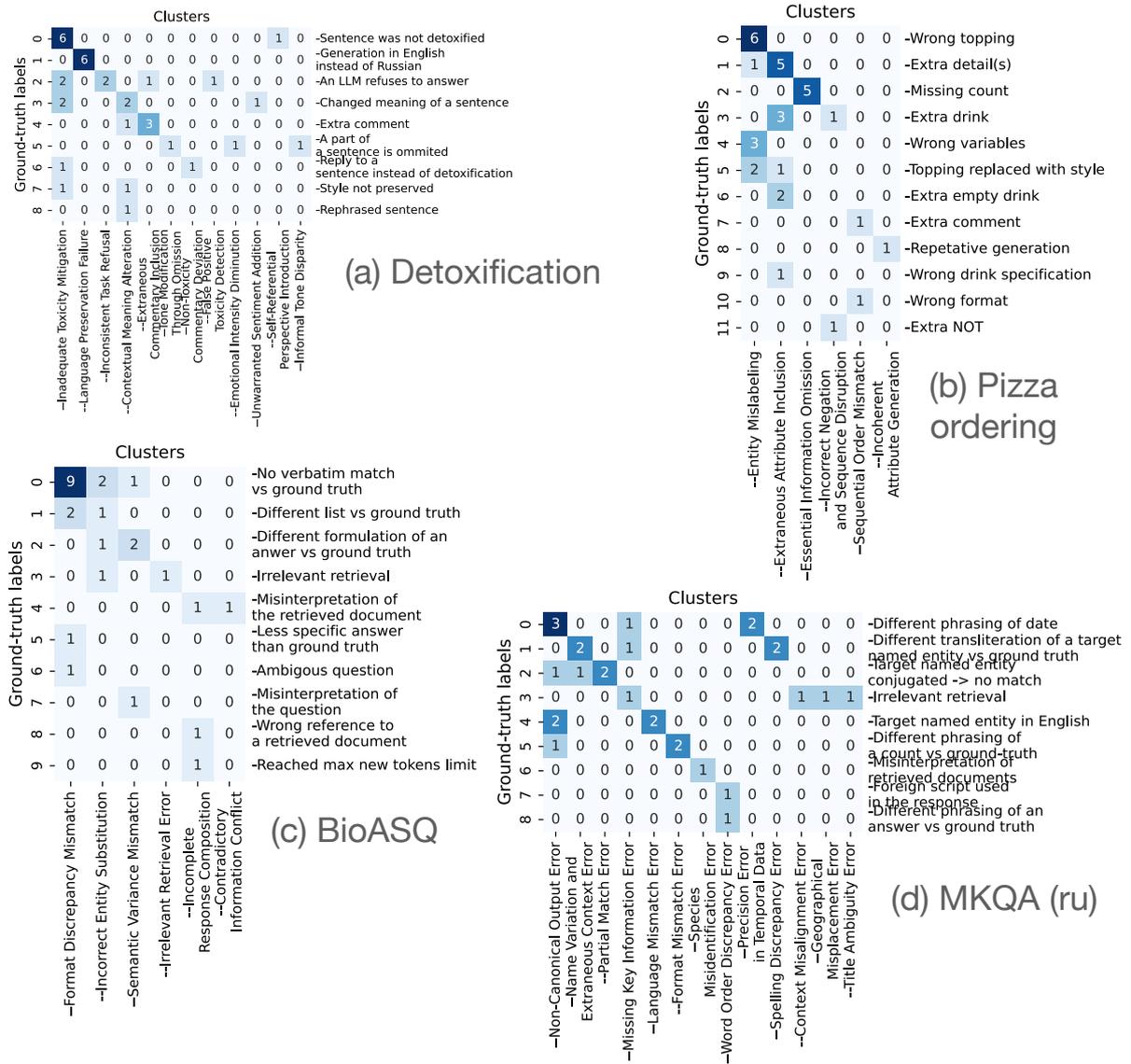


Figure 5: Examples of confusion matrices visualizing clustering agreement between *LLM-as-qualitative-judge*-generated and the annotator’s issue types reports. We find the optimal mapping between clusters found by a human annotator and by *LLM-as-a-qualitative-judge*, and then define a confusion matrix where each cell  $(i, j)$  denotes a number of dataset instances allocated into  $i$ -th annotator’s cluster and  $j$ -th *LLM-as-a-qualitative-judge*’s cluster.

one issue per instance is reasonable, since many of the erroneous instances have only one issue. Furthermore, even if some instances have repeating issues, our algorithm would still capture most of the issues in the dataset due to issue repetition.

**Pairwise comparison of models.** *LLM-as-a-qualitative-judge* can be straightforwardly extended to perform pairwise comparison of models, by running per-instance analysis for outputs of both models and using two counters (one per model) in the issue clustering step.

**Use without ground truth labels.** *LLM-as-a-qualitative-judge* can be straightforwardly used

without ground truth labels, i.e. as an unsupervised evaluation metric, if the LLM internal knowledge is sufficient to understand errors in a given task. *LLM-as-a-qualitative-judge* can also be provided with additional evaluation metadata, e.g. score rubrics used in quantitative evaluation.

## 7 Related work

**Quantitative LLM-based evaluation.** While using commercial LLMs for evaluation remains common practice, one line of work (Kim et al., 2024b,c; Pombal et al., 2025) focuses on tuning open-source LLMs on the synthetically generated evaluation data, to ensure reproducibility of evaluation. Other

works improve quantitative LLM-as-a-judge by conducting more fine-grained evaluation, e.g. using multiple evaluation criteria (Liang et al., 2023; Fu et al., 2024) or selecting evaluation criteria individually per instance (Ye et al., 2024; Cook et al., 2024; Kim et al., 2025). Composite evaluation approaches such as FactScore (Min et al., 2023) or RAGChecker (Ru et al., 2024) use LLMs in the intermediate evaluation steps.

**Qualitative LLM-based evaluation.** LLM-generated *qualitative* error explanations are often used to improve the precision of quantitative evaluation (Zeng et al., 2024; Ye et al., 2024) or to explain the assigned quantitative scores to a developer (Xu et al., 2023; Jiang et al., 2024). Such approaches only output *per-instance* explanations, and a *substantial human effort is still needed to read all of them*. Certain works (Jiang et al., 2024; Kasner et al., 2024; Perrella et al., 2022; Guerreiro et al., 2024) focus on outputting aggregated reports of frequent errors, but with a (limited) predefined error set, i.e. they solve the task of error classification. In contrast to these efforts, *LLM-as-a-qualitative-judge* outputs an *aggregated* report of issue types discovered in an *open-ended* manner, i.e. without any predefined issue set.

**Meta-evaluation.** A line of community efforts (Zeng et al., 2024; Lambert et al., 2025; Hada et al., 2024; Bavaresco et al., 2024) is devoted to an important task of meta-evaluating LLM-as-a-judge, i.e. collecting human annotations for various tasks, domains, or languages, and evaluating how closely LLMs mirror human judgments. Certain task-specific datasets (Freitag et al., 2021) can be used to meta-evaluate fine-grained issue detection. Our work further contributes to this direction by the release of a meta-evaluation dataset, containing *qualitative* issue explanations for 12 datasets from 7 tasks and their per-dataset *clustering*.

**Clustering with LLMs.** Earlier works (Petukhova et al., 2024; Miller and Alexander, 2024) demonstrate advantages of leveraging LLM-derived embeddings in place of traditional TF-IDF or BERT vectors in standard clustering algorithms. More recent works employ LLMs directly to cluster textual data. Viswanathan et al. (2024) instruct a GPT-3.5 model to cluster the provided data given few-shot demonstrations. Huang and He (2024) transform clustering into a two-stage classification task: first prompting an

LLM to infer a set of candidate clusters for the dataset, then prompting it to assign the best cluster to each instance. ClusterLLM (Zhang et al., 2023) uses an instruction-tuned LLM to guide clustering, i.e., to decide which clusters to merge. In our work, we propose an alternative intuitive approach for LLM-based clustering. Our approach can also be extended in the future with the listed strategies.

## 8 Conclusion

In this work, we present *LLM-as-a-qualitative-judge*, a novel approach for generating structured reports summarizing key types of issues in a given NLG system. We hope that this approach will help developers to spot more easily issues and artifacts in their NLG systems.

Future works could equip *LLM-as-a-qualitative-judge* with advanced reasoning or agentic pipelines, tune LLMs for issue report generation, and study the approach for a wider set of languages.

## Limitations

As any LLM-based system, *LLM-as-a-qualitative-judge* can make occasional mistakes in analysis or clustering. In Section 5, we discuss types of such mistakes and recommend checking a couple of examples of each issue, which are also output by *LLM-as-a-qualitative-judge*.

Regarding limitations of the evaluation methodology, despite our efforts in considering a diverse set of tasks, domains, and LLMs, we acknowledge the infeasibility of covering the entire breadth of NLG applications and models in our study. Another limitation is that we mainly focus on English. We believe our findings will transfer to other languages, with the use of strong recent multilingual LLMs, but acknowledge that the reliability of *LLM-as-a-qualitative-judge* in multilingual studies requires a separate study.

## Broader impact

We acknowledge that as any LLM-based system, *LLM-as-a-qualitative-judge* can make errors which could propagate to the downstream systems and decrease their performance. For example, if developers rely solely on the issue names formulated by Judge, this could occasionally lead to unnecessary or even harmful modifications of their NLG systems. This could also happen in case of misinterpretation of an issue by a developer due to issue subjectivity. To reduce such risks, we recommend

developers to check examples of issue types, which are also output by *LLM-as-a-qualitative-judge*, in addition to the issue names and description.

## Acknowledgments

We greatly appreciate the help of Alexandre Misrahi, Salah Ait-Mokhtar, and Maxime Louis. The project was initiated at the Advanced Language Processing School (ALPS 2025, <https://alps.imag.fr>).

A part of this work was carried out within the framework of the AugmentIA Chair, supported by the Fondation Grenoble INP thanks to the patronage of Artelia Group, and is affiliated with Laboratory of informatics in Grenoble (LIG). A part of this work received government funding managed by the French National Research Agency under France 2030, reference ANR-23-IACL-0006.

Tunde’s research is conducted with the financial support of Research Ireland under Grant Number 12/RC/2289\_P2 - Insight Research Ireland Centre for Data Analytics.

## References

- Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy P. Lillicrap, and 33 others. 2023. [Gemini: A family of highly capable multimodal models](#). *CoRR*, abs/2312.11805.
- Konstantine Arkoudas, Nicolas Guenon des Mesnards, Melanie Rubino, Sandesh Swamy, Saarthak Khanna, Weiqi Sun, and Haidar Khan. 2022. [PIZZA: A new benchmark for complex end-to-end task-oriented parsing](#). *CoRR*, abs/2212.00265.
- Satanjeev Banerjee and Alon Lavie. 2005. [METEOR: An automatic metric for MT evaluation with improved correlation with human judgments](#). In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.
- Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, André F. T. Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suggia, Aditya K Surikuchi, Ece Takmaz, and Alberto Testoni. 2024. [Llms instead of human judges? a large scale empirical study across 20 nlp evaluation tasks](#). *Preprint*, arXiv:2406.18403.
- Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. [Vicuna: An open-source chatbot impressing gpt-4 with 90%\\* chatgpt quality](#). <https://lmsys.org/blog/2023-03-30-vicuna/>.
- Elizabeth Clark, Shruti Rijhwani, Sebastian Gehrmann, Joshua Maynez, Roei Aharoni, Vitaly Nikolaev, Thibault Sellam, Aditya Siddhant, Dipanjan Das, and Ankur Parikh. 2023. [SEAHORSE: A multilingual, multifaceted dataset for summarization evaluation](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9397–9413, Singapore. Association for Computational Linguistics.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. [Training verifiers to solve math word problems](#). *CoRR*, abs/2110.14168.
- Jonathan Cook, Tim Rocktäschel, Jakob Foerster, Dennis Aumiller, and Alex Wang. 2024. [Ticking all the boxes: Generated checklists improve llm evaluation and generation](#). *Preprint*, arXiv:2410.03608.
- John Dang, Shivalika Singh, Daniel D’souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, and 26 others. 2024. [Aya expand: Combining research breakthroughs for a new multilingual frontier](#). *Preprint*, arXiv:2412.04261.
- DeepSeek-AI. 2025. [Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning](#). *Preprint*, arXiv:2501.12948.
- Daryna Dementieva, Daniil Moskovskiy, Nikolay Babakov, Abinew Ali Ayele, Naqee Rizwan, Florian Schneider, Xintong Wang, Seid Muhie Yimam, Dmitry Ustalov, Elisei Stakovskii, Alisa Smirnova, Ashraf Elnagar, Animesh Mukherjee, and Alexander Panchenko. 2024. [Overview of the multilingual text detoxification task at PAN 2024](#). In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), Grenoble, France, 9-12 September, 2024*, volume 3740 of *CEUR Workshop Proceedings*, pages 2432–2461. CEUR-WS.org.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

- Matthew Dunn, Levent Sagun, Mike Higgins, V. Ugur Güney, Volkan Cirik, and Kyunghyun Cho. 2017. [Searchqa: A new q&a dataset augmented with context from a search engine](#). *CoRR*, abs/1704.05179.
- Nigel Fernandez, Alexander Scarlato, and Andrew S. Lan. 2024. [Syllabusqa: A course logistics question answering dataset](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 10344–10369. Association for Computational Linguistics.
- Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. [Experts, errors, and context: A large-scale study of human evaluation for machine translation](#). *Transactions of the Association for Computational Linguistics*, 9:1460–1474.
- Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2024. [GPTScore: Evaluate as you desire](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6556–6576, Mexico City, Mexico. Association for Computational Linguistics.
- Xinyang Geng, Arnav Gudibande, Hao Liu, Eric Wallace, Pieter Abbeel, Sergey Levine, and Dawn Song. 2023. [Koala: A dialogue model for academic research](#). Blog post.
- Aidan Gomez and Cohere for AI. 2024. [Command r: Retrieval-augmented generation at production scale](#). <https://cohere.com/blog/command-r>.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. [xCOMET: Transparent machine translation evaluation through fine-grained error detection](#). *Transactions of the Association for Computational Linguistics*, 12:979–995.
- Rishav Hada, Varun Gumma, Mohamed Ahmed, Kallika Bali, and Sunayana Sitaram. 2024. [METAL: Towards multilingual meta-evaluation](#). In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2280–2298, Mexico City, Mexico. Association for Computational Linguistics.
- Rujun Han, Peng Qi, Yuhao Zhang, Lan Liu, Juliette Burger, William Wang, Zhiheng Huang, Bing Xiang, and Dan Roth. 2023. [Robustqa: Benchmarking the robustness of domain adaptation for open-domain question answering](#). In *ACL Findings 2023*.
- Rujun Han, Yuhao Zhang, Peng Qi, Yumo Xu, Jenyuan Wang, Lan Liu, William Yang Wang, Bonan Min, and Vittorio Castelli. 2024. [RAG-QA arena: Evaluating domain robustness for long-form retrieval augmented question answering](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 4354–4374. Association for Computational Linguistics.
- Chen Huang and Guoxiu He. 2024. [Text clustering as classification with LLMs](#). *CoRR*, abs/2410.00927.
- Dongfu Jiang, Yishan Li, Ge Zhang, Wenhao Huang, Bill Yuchen Lin, and Wenhui Chen. 2024. [TIGER-Score: Towards building explainable metric for all text generation tasks](#). *Transactions on Machine Learning Research*.
- Zdeněk Kasner, Ondrej Patek, Patricia Schmidtova, Simone Balloccu, and Ondrej Dusek. 2024. [factgenie: A framework for span-based evaluation of generated texts](#). In *Proceedings of the 17th International Natural Language Generation Conference: System Demonstrations*, pages 13–15, Tokyo, Japan. Association for Computational Linguistics.
- Sanghoon Kim, Dahyun Kim, Chanjun Park, Wonsung Lee, Wonho Song, Yunsu Kim, Hyeonwoo Kim, Yungi Kim, Hyeonju Lee, Jihoo Kim, Changbae Ahn, Seonghoon Yang, Sukyung Lee, Hyunbyung Park, Gyoungjin Gim, Mikyoung Cha, Hwalsuk Lee, and Sunghun Kim. 2024a. [SOLAR 10.7B: Scaling large language models with simple yet effective depth up-scaling](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, pages 23–35, Mexico City, Mexico. Association for Computational Linguistics.
- Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoon Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. 2024b. [Prometheus: Inducing fine-grained evaluation capability in language models](#). In *The Twelfth International Conference on Learning Representations*.
- Seungone Kim, Juyoung Suk, Ji Yong Cho, Shayne Longpre, Chaeun Kim, Dongkeun Yoon, Guijin Son, Yejin Cho, Sheikh Shafayat, Jinheon Baek, Sue Hyun Park, Hyeonbin Hwang, Jinkyung Jo, Hyowon Cho, Haebin Shin, Seongyun Lee, Hanseok Oh, Noah Lee, Namgyu Ho, and 13 others. 2025. [The BiGGen bench: A principled benchmark for fine-grained evaluation of language models with language models](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5877–5919, Albuquerque, New Mexico. Association for Computational Linguistics.
- Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham

- Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024c. [Prometheus 2: An open source language model specialized in evaluating other language models](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4334–4353, Miami, Florida, USA. Association for Computational Linguistics.
- Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. [Findings of the 2022 conference on machine translation \(WMT22\)](#). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Tom Kocmi and Christian Federmann. 2023. [Large language models are state-of-the-art evaluators of translation quality](#). In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.
- Anastasia Krithara, Anastasios Nentidis, Konstantinos Bougiatiotis, and Georgios Paliouras. 2023. [BioASQQA: A manually curated corpus for biomedical question answering](#). *Scientific Data*, 10(1):170.
- Nathan Lambert, Valentina Pyatkin, Jacob Morrison, Lester James Validad Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. 2025. [RewardBench: Evaluating reward models for language modeling](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1755–1797, Albuquerque, New Mexico. Association for Computational Linguistics.
- Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D Manning, Christopher Re, Diana Acosta-Navas, Drew A. Hudson, and 31 others. 2023. [Holistic evaluation of language models](#). *Transactions on Machine Learning Research*. Featured Certification, Expert Certification, Outstanding Certification.
- Chin-Yew Lin. 2004. [ROUGE: A package for automatic evaluation of summaries](#). In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Shayne Longpre, Yi Lu, and Joachim Daiber. 2021. [MKQA: A linguistically diverse benchmark for multilingual open domain question answering](#). *Transactions of the Association for Computational Linguistics*, 9:1389–1406.
- Justin K. Miller and Tristram J. Alexander. 2024. [Human-interpretable clustering of short-text using large language models](#). *CoRR*, abs/2405.07278.
- Sewon Min, Kalpesh Krishna, Xinxu Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. [FActScore: Fine-grained atomic evaluation of factual precision in long form text generation](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.
- Alexandre Misrahi, Nadezhda Chirkova, Maxime Louis, and Vassilina Nikoulina. 2025. [Adapting large language models for multi-domain retrieval-augmented-generation](#). *Preprint*, arXiv:2504.02411.
- Mistral AI. 2024. [Un ministral, des ministraux](#). Accessed May 19, 2025.
- OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. [Gpt-4 technical report](#). *Preprint*, arXiv:2303.08774.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [Bleu: a method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Stefano Perrella, Lorenzo Proietti, Alessandro Scirè, Niccolò Campolungo, and Roberto Navigli. 2022. [MaTese: Machine translation evaluation as a sequence tagging problem](#). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 569–577, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Alina Petukhova, João Pedro Matos-Carvalho, and Nuno Fachada. 2024. [Text clustering with LLM embeddings](#). *CoRR*, abs/2403.15112.
- José Pombal, Dongkeun Yoon, Patrick Fernandes, Ian Wu, Seungone Kim, Ricardo Rei, Graham Neubig, and André F. T. Martins. 2025. [M-prometheus: A suite of open multilingual LLM judges](#). *CoRR*, abs/2504.04953.
- Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. [Qwen2.5 technical report](#). *Preprint*, arXiv:2412.15115.

- Dongyu Ru, Lin Qiu, Xiangkun Hu, Tianhang Zhang, Peng Shi, Shuaichen Chang, Cheng Jiayang, Cunxiang Wang, Shichao Sun, Huanyu Li, Zizhao Zhang, Binjie Wang, Jiarong Jiang, Tong He, Zhiguo Wang, Pengfei Liu, Yue Zhang, and Zheng Zhang. 2024. [Ragchecker: A fine-grained framework for diagnosing retrieval-augmented generation](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 21999–22027. Curran Associates, Inc.
- Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. 2022. [ColBERTv2: Effective and efficient retrieval via lightweight late interaction](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3715–3734, Seattle, United States. Association for Computational Linguistics.
- Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, , and Jason Wei. 2022. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*.
- Alon Talmor, Ori Yoran, Ronan Le Bras, Chandra Bhagavatula, Yoav Goldberg, Yejin Choi, and Jonathan Berant. 2021. [Commonsenseqa 2.0: Exposing the limits of AI through gamification](#). In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.
- Thibaut Thonet, Laurent Besacier, and Jos Rozen. 2025. [Elitr-bench: A meeting assistant benchmark for long-context language models](#). In *Proceedings of the 31st International Conference on Computational Linguistics, COLING 2025, Abu Dhabi, UAE, January 19-24, 2025*, pages 407–428. Association for Computational Linguistics.
- Emiel van Miltenburg, Miruna Clinciu, Ondřej Dušek, Dimitra Gkatzia, Stephanie Inglis, Leo Leppänen, Saad Mahamood, Emma Manning, Stephanie Schoch, Craig Thomson, and Luou Wen. 2021. [Underreporting of errors in NLG output, and what to do about it](#). In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 140–153, Aberdeen, Scotland, UK. Association for Computational Linguistics.
- Emiel van Miltenburg, Miruna Clinciu, Ondřej Dušek, Dimitra Gkatzia, Stephanie Inglis, Leo Leppänen, Saad Mahamood, Stephanie Schoch, Craig Thomson, and Luou Wen. 2023. [Barriers and enabling factors for error analysis in NLG research](#). *Northern European Journal of Language Technology*, 9.
- Vijay Viswanathan, Kiril Gashteovski, Carolin Lawrence, Tongshuang Wu, and Graham Neubig. 2024. [Large language models enable few-shot clustering](#). *Trans. Assoc. Comput. Linguistics*, 12:321–333.
- Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. [Self-instruct: Aligning language models with self-generated instructions](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.
- Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, Qingwei Lin, and Daxin Jiang. 2024. [WizardLM: Empowering large pre-trained language models to follow complex instructions](#). In *The Twelfth International Conference on Learning Representations*.
- Wenda Xu, Danqing Wang, Liangming Pan, Zhenqiao Song, Markus Freitag, William Wang, and Lei Li. 2023. [INSTRUCTSCORE: Towards explainable text generation evaluation with automatic feedback](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5967–5994, Singapore. Association for Computational Linguistics.
- Seonghyeon Ye, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, Seungone Kim, Yongrae Jo, James Thorne, Juho Kim, and Minjoon Seo. 2024. [FLASK: Fine-grained language model evaluation based on alignment skill sets](#). In *The Twelfth International Conference on Learning Representations*.
- Zhiyuan Zeng, Jiatong Yu, Tianyu Gao, Yu Meng, Tanya Goyal, and Danqi Chen. 2024. [Evaluating large language models at evaluating instruction following](#). In *The Twelfth International Conference on Learning Representations*.
- Yuwei Zhang, Zihan Wang, and Jingbo Shang. 2023. [ClusterLLM: Large language models as a guide for text clustering](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 13903–13920. Association for Computational Linguistics.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023a. [Judging llm-as-a-judge with mt-bench and chatbot arena](#). volume 36, pages 46595–46623.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023b. [Judging LLM-as-a-judge with MT-bench and chatbot arena](#). In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

## A Details on data annotation

**Per-instance analysis.** The core of our meta-evaluation strategy is to collect manual annotations of failure cases for a set of instances from various tasks and domains. For each instance, consisting of a task input, a ground truth response, a generated response, a description of a task metric, and optionally retrieved documents, an annotator’s task is to formulate what the particular issue is in this instance. We then prompt an *evaluator LLM* to judge whether the issue explanation determined by the *LLM-as-a-qualitative-judge* for a particular instance matches the issue determined by the human annotator. The outputs from the *evaluator LLM* are binary and can be accumulated into a *per-instance analysis accuracy* score.

The annotation instruction asks to ignore instances which have multiple issues (to avoid ambiguity in per-instance analysis), instances where ground truth labels appear to be wrong, and instances where the annotator’s expertise is insufficient to judge the correctness of the generated answer. We also limit the number of instances with the same issue to not exceed 8 examples per dataset, to ensure the diversity of the final dataset.

**Issue clustering.** Human annotation also includes a step of manually clustering issues discovered in per-instance analysis, i.e., specifying cluster indices and cluster names (generalized issue types) for labeled instances. This annotation is then used to compute clustering agreement between the clustering produced by *LLM-as-a-qualitative-judge* and by a human annotator.

**Dataset composition.** For each considered dataset, we manually label failures in up to 40 generations from one of the open-source LLMs (Qwen-2.5-1b, Llama-3.2-1b, Command-R-35b, Vicuna-1.5-13B).

**Annotation details.** All the labeling was performed by the authors of the paper in Google Spreadsheets<sup>4</sup>. Each instance was annotated by one author. Authors of the paper are PhD students in the NLP field or have already completed their PhD in NLP and are employed as NLP researchers.

Time needed for data annotation varies between tasks: it took us 1–6 hours per task. Tasks with longer inputs, e.g. RA-QA, and from more complex domains, e.g. biomedical, take more time to

annotate, e.g. they require reading the retrieved documents carefully.

Annotation instruction is provided in Figure 6.

**Inter-annotator agreement.** We measure the inter-annotator agreement on a subset of 100 instances, i.e. each of these instances was labeled by two annotators and then we computed their agreement using the same evaluator LLM as in other experiments, i.e. Claude-3.7-sonnet. The resulting inter-annotator agreement was 57% (percentage of cases when two annotators suggested the same issue, as judged by Claude-3.7-sonnet), i.e. the similar range as the scores we obtain in Table 2.

The main factor contributing to the moderate agreement is the subjectivity of issue analysis. For example, in a situation when generation was stopped due to reaching the maximum new tokens limit, one annotator said “The response is incomplete” and another annotator said “Generation was stopped too early”. Both denote the same root issue, but are formulated differently and Claude judges these comments as different.

## B Further details on the experimental setup

**Case study.** For each of the three considered Big-BenchHars tasks, we build a simple initial generative pipeline. This pipeline is then improved in two rounds by generating issue reports with *LLM-as-a-qualitative-judge*. Configurations of the initial pipeline are as follows. System prompt: “You are a helpful assistant. Output your answer after a final separator ‘Answer:’”. Generation hyperparameters: all hyperparameters set to default values from the HuggingFace or OpenAI API, plus setting maximum new tokens or 500 for HuggingFace models. The final answers are obtained by cropping the content after a final separator “Answer:” and applying a `.strip()` python function. Evaluation function: exact match with ground truth. *LLM-as-a-qualitative-judge* is run with GPT-4.1 and providing a one-sentence description of a task metric, i.e. “Evaluation is conducted using exact matching between the ground-truth label and the content of the generated response after the final separator ‘Answer:’”.

**Meta-evaluation experiments.** For each instance, *LLM-as-a-qualitative-judge* is provided with a task input, a ground truth response, a generated response, 5 retrieved documents (only for

<sup>4</sup><https://docs.google.com/spreadsheets>

For each example, consisting of a user prompt, a ground truth label, an LLM generation, and optionally retrieved documents, an annotator’s task is to formulate what is a particular failure case in this example. Identify only one, most important specific, clearly visible issue in each test case. Please formulate the detected issue as a clear, full sentence, e.g. “The generated response is in German instead of French which is the language of the user input” or “The retrieved documents are from a datastore which is irrelevant to the given user question”.

Please add your annotations in the following Google Spreadsheet: [link], column “Per-instance analysis”. You can skip instances (rows) for which you feel that you do not have enough expertise to detect an issue, which have multiple issues, or for which ground truth labels appear to be wrong.

After annotating per-instance analysis, suggest a clustering of the detected issues, i.e. how would you group them, and add the corresponding cluster indexes and names in columns “Cluster index” and “Cluster name”.

Figure 6: Annotation instruction for meta-evaluation data.

RA-QA), and a short task comment. The task comment describes the task metric (in one sentence), provides a comment on the nature of ground truth responses (either that it is the expected answer or that it is only one of the possible correct answers), and also contains a comment that retrieval-augmented generation (RAG) or Chain-of-Thought (COT) prompting was used, when applicable (6 datasets with RAG and 2 datasets with COT). The used task metrics are binary LLM-as-a-judge (the generated response is accepted or not) or binary Match (outputs True is one of the ground truth answers is contained a substring in the generated response, and False otherwise). Task comments for all datasets are also presented in Appendix H.

For *LLM-as-a-qualitative-judge*, open-source LLMs are run on a single V100 GPU with greedy decoding (~20 GPU-hours in total). Commercial LLMs are run via API with requesting json output format.

The time of running the *LLM-as-a-qualitative-judge* algorithm depends on the setting (commercial vs open-source LLMs, type of GPU etc) and in our experiments was taking 2–30 min, i.e. reasonably short on the scale of the time needed to develop an NLG system.

**Meta-evaluation of an evaluator LLM.** To ensure the reliability of the *evaluator LLM*, we collected a small meta-evaluation dataset of 50 instances from 4 datasets (MKQA (ru), RobustQA Writing, FLASK, MultiDetox), where the equiva-

lence of the *LLM-as-a-qualitative-judge*’s and *human annotator*’s per-instance analysis was judged by a human annotator and can be compared to the *evaluator LLM*’s verdict. Strong commercial LLMs, such as GPT-4o, Gemini-2.0-Flash, and Claude-3.7-Sonnet, achieved a meta-evaluation accuracy of 85-90% on this dataset, and an open-source Solar-10.7B (Kim et al., 2024a) achieved a meta-accuracy of 60%. In all the experiments, we use claude-3-7-sonnet-20250219 as the *evaluator LLM*, to avoid using the same LLM for analysis and for evaluation.

## C Clustering experiment setup

In this experiment, we perform a hyperparameter grid search for five clustering algorithms: KMeans, Agglomerative Clustering, Spectral Clustering, Gaussian Mixture Models (GMM), and HDBSCAN on a synthetic set. Each algorithm is evaluated across a range of hyperparameter combinations. For KMeans, we vary the distance\_metric (euclidean, cosine), kmeans\_init strategy (kmeans++, random), kmeans\_n\_init (10, 50), and kmeans\_max\_iter (300, 500). For Agglomerative Clustering, we test all combinations of distance\_metric (euclidean, cosine) and linkage\_type (ward, average, complete), while ensuring that ward is only paired with euclidean (as required by the algorithm). Spectral Clustering configurations include distance\_metric (euclidean, cosine), assign\_labels (kmeans, discretize), spectral\_gamma (0.1, 0.5, 1.0,

2.0), and `spectral_n_neighbors` (5, 10, 20). For GMM, we explore `covariance_type` (full, diag), `gmm_init_params` (kmeans, random), and `gmm_max_iter` (100, 300). Lastly, HDBSCAN is tested with `distance_metric` (euclidean, cosine), `min_cluster_size` (3, 5, 10, 15, 20), `hdbscan_min_samples` (None, 1, 5), and `hdbscan_cluster_selection_method` (eom, leaf). Each valid configuration is evaluated over three independent trials with different random seeds to ensure robustness. After collecting results based on Adjusted Rand Index (ARI), the best-performing configuration for each algorithm on the synthetic validation set is selected. These best configurations are then applied to the test set of synthetic data and to the real dataset.

## D Models

Model	BibTeX	License	Model Repository
GPT-4o	(OpenAI et al., 2024)	Proprietary	<a href="https://platform.openai.com/docs/models/gpt-4o">https://platform.openai.com/docs/models/gpt-4o</a>
Gemini-2.0-Flash	(Anil et al., 2023)	Proprietary	<a href="https://deepmind.google/technologies/gemini/flash/">https://deepmind.google/technologies/gemini/flash/</a>
Qwen-2.5	(Qwen et al., 2025)	Apache 2.0	<a href="https://huggingface.co/collections/Qwen/qwen25-66e81a666513e518adb90d9e">https://huggingface.co/collections/Qwen/qwen25-66e81a666513e518adb90d9e</a>
DeepSeek-R1-Distill-Llama-8B	(DeepSeek-AI, 2025)	Llama 3.1 Community License	<a href="https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Llama-8B">https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Llama-8B</a>
Aya-Expansive-8B	(Dang et al., 2024)	Creative Commons Attribution Non Commercial 4.0	<a href="https://huggingface.co/CohereLabs/aya-expansive-8b">https://huggingface.co/CohereLabs/aya-expansive-8b</a>
Llama-3.1-8B-Instruct	(Grattafiori et al., 2024)	Llama 3.1 Community License	<a href="https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct">https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct</a>
Llama-3.2-1B-Instruct	(Grattafiori et al., 2024)	Llama 3.2 Community License	<a href="https://huggingface.co/meta-llama/Llama-3.2-1B-Instruct">https://huggingface.co/meta-llama/Llama-3.2-1B-Instruct</a>
Minstral-8B-Instruct	(Mistral AI, 2024)	Mistral AI Research License	<a href="https://huggingface.co/mistralai/Minstral-8B-Instruct-2410">https://huggingface.co/mistralai/Minstral-8B-Instruct-2410</a>
Solar-10.7B	(Kim et al., 2024a)	Creative Commons Attribution Non Commercial 4.0	<a href="https://huggingface.co/upstage/SOLAR-10.7B-Instruct-v1.0">https://huggingface.co/upstage/SOLAR-10.7B-Instruct-v1.0</a>
Vicuna-1.5-13B	(Chiang et al., 2023)	Llama 2 Community License Agreement	<a href="https://huggingface.co/lmsys/vicuna-13b-v1.5">https://huggingface.co/lmsys/vicuna-13b-v1.5</a>
Command-R-35B	(Gomez and for AI, 2024)	Creative Commons Attribution Non Commercial 4.0	<a href="https://huggingface.co/CohereLabs/c4ai-command-r-v01">https://huggingface.co/CohereLabs/c4ai-command-r-v01</a>

Table 4: References to the used LLMs; all LLMs allow use for research.

## E Datasets

Dataset name	Dataset reference	License
<b>Natural Language Generation</b>		
FLASK data mix: Self-Instruct, WizardLM, Koala, CommonSense QA	(Wang et al., 2023; Xu et al., 2024; Geng et al., 2023; Talmor et al., 2021)	Apache 2.0, MIT, Apache 2.0, Creative Commons Attribution 4.0
WMT’22	(Kocmi et al., 2022)	Apache 2.0
Elitr-Bench	(Thonet et al., 2025)	Attribution 4.0 International
PIZZA	(Arkoudas et al., 2022)	Attribution-NonCommercial 4.0 International
GSM8K	(Cobbe et al., 2021)	MIT License
ParaDetox	(Dementieva et al., 2024)	OpenRAIL++
<b>Retrieval-augmented QA</b>		
MKQA (ru)	(Longpre et al., 2021)	Creative Commons Attribution-ShareAlike 3.0 Unported License
BioASQ	(Krithara et al., 2023)	Attribution 2.5 Generic
RobustQA	(Santhanam et al., 2022; Han et al., 2024, 2023)	Apache-2.0
SearchQA	(Dunn et al., 2017)	BSD 3-Clause
SyllabusQA	(Fernandez et al., 2024)	Attribution-NonCommercialShareAlike
<b>BigBenchHard</b>		
Date understanding; Word sorting; Movie recommendation	(Suzgun et al., 2022)	MIT

Table 5: References to the used datasets; all datasets allow use for research. We select instances from test splits.

## F Per-dataset results

Table 6 presents per-dataset results for for GPT-4o as *LLM-as-a-qualitative-judge*.

## G Prompts

Figures 8–11 present prompts used for per-instance analysis, issue clustering, and evaluation.

Dataset	Per-inst. an. acc. (%)	Issue clust. ARI	Issue clust. SLC
Semantic parsing	94.1	0.41	0.29
Grade school math	88.2	0.04	0.22
Detoxification	77.8	0.36	0.28
Long-context QA	69.2	0.07	0.50
Translation en-ru	65.8	0.10	0.63
Instruction following	55.9	0.09	0.19
RA-QA: SyllabusQA	77.8	0.16	0.55
RA-QA: MKQA (ru)	75.8	0.17	0.44
RA-QA: BioASQ	66.7	0.08	0.31
RA-QA: SearchQA	38.4	0.15	0.16
RA-QA: Writing	30.7	0.00	0.11
RA-QA: Lifestyle	19.0	0.00	0.32

Table 6: Per-dataset results for GPT-4o as *LLM-as-a-qualitative-judge*.

You are an expert in analysing the failure cases in natural language generation tasks

You are given a question, ground truth label(s), and the answer generated by an LLM.

The generated answer was not accepted by the automatic evaluation measured with metric {metric info}

Read all these materials and reply what is the particular failure case in this example, i. e. why exactly the generated response was not accepted.

The problem can be in any part of the pipeline, including the question itself or any aspects of the system outputs.

IMPORTANT: identify ONE, MOST IMPORTANT, SPECIFIC, CLEARLY VISIBLE issue in each test case.

Question:  
 \*\*\*  
 {question}  
 \*\*\*

Ground truth label(s):  
 \*\*\*  
 {label}  
 \*\*\*

LLM-generated answer:  
 \*\*\*  
 {answer}  
 \*\*\*

So what is the particular failure in this example? First output a detailed analysis, and then output a final summary of the failure in one or two sentences after a special separator "Summary:".

Figure 7: Prompt used for per-instance analysis. The presented version of the prompt is for text LLM outputs, the prompt can be easily changed if JSON outputs are supported by an LLM. For RA-QA, we also include retrieved documents in the prompt.

You are an expert in analysing the failure cases of natural language generation systems.

You already performed per-example analysis, where for each example, you were given a question, ground truth label(s), and the answer generated by an LLM. The generated answer in all examples was not accepted by the automatic evaluation.

You already read all these materials and formulated what was the particular failure case in each example, i.e. which part of the pipeline failed so that the generated response was not accepted.

Now your task is to summarize all your per-example analyses into a concise overall summary of failure cases for the given dataset.

Summarize what are the various failure types in this dataset (provide the overall count of each error type and also ids of all examples of each error type).

Please try to be very specific in determining error types, avoid too much generic error types. On the contrary, determine as much as possible FINE GRAINED error types.

Furthermore, provide a comment for each error type explaining the essence of this error type in a bit more details (in the context of this particular dataset).

```
*** Per-example analysis which you generated before ***  
{analysis}  
*** per-example analysis ended ***
```

Summarize all your per-example analyses into a concise overall summary of failure cases.

Generate a json with the only key "summary", and a value is a dict of error types. Each value in this dict (corresponding to one detected error type) is a dictionary with keys "error\_name", "error\_description", "indexes" (indexes of all examples with this error type), and "num\_examples" (overall count of this error type).

Figure 8: Prompt used for issue clustering with direct prompting.

You are an expert in analysing the failure cases of natural language generation systems.

You already performed per-example analysis, where for each example, you were given a question, ground truth label(s), and the answer generated by an LLM. The generated answer in all examples was not accepted by the automatic evaluation.

You already read all these materials and formulated what was the particular failure case in each example, i.e. which part of the pipeline failed so that the generated response was not accepted.

Now your task is to summarize all your per-example analyses into a concise overall summary of failure cases for the given dataset.

You are using a SPECIAL CUMULATIVE ALGORITHM as follows. You are going through examples one by one and accumulate discovered error cases in a special pool. For each example, you check if any of the already discovered error types from the pool fits it, and if so, you assign this error type to this example. If none of already existing error types fit the current example, you create a new error type and add it to a pool.

A pool of already discovered error cases:

\*\*\*

{error cases}

\*\*\*

Analysis of a current example:

\*\*\*

{analysis}

\*\*\*

Do you think any of the already discovered error cases fit the current example? If yes, output "Decision:" and a key marking the chosen error case, e.g. "Decision: type\_0". Do it only if you are REALLY sure that the chosen error case fits the current example! DO NOT output cluster name. If not, output "Decision: None". DO NOT output anything else.

Figure 9: Prompt used for classifying each instance in the cumulative clustering strategy. The presented version of the prompt is for text LLM outputs, the prompt can be easily changed if JSON outputs are supported by an LLM.

You are an expert in analysing the failure cases of natural language generation systems.

You already performed per-example analysis, where for each example, you were given a question, ground truth label(s), and the answer generated by an LLM. The generated answer in all examples was not accepted by the automatic evaluation.

You already read all these materials and formulated what was the particular failure case in each example, i.e. which part of the pipeline failed so that the generated response was not accepted.

Now your task is to summarize all your per-example analyses into a concise overall summary of failure cases for the given dataset.

You are using a SPECIAL CUMULATIVE ALGORITHM as follows. You are going through examples one by one and accumulate discovered error cases in a special pool. For each example, you check if any of the already discovered error types from the pool fits it, and if so, you assign this error type to this example. If none of already existing error types fit the current example, you create a new error type and add it to a pool.

A pool of already discovered error cases:

\*\*\*

{error cases}

\*\*\*

Analysis of a current example:

\*\*\*

{analysis}

\*\*\*

You decided to create a new error type for a given example, not yet present in a pool. Now you need to generate SHORT LABEL and a 1 or 2 SENTENCE DESCRIPTION for this new error type. Please try to be very specific in determining a FINE GRAINED error type, avoid too much generic error types. At the same time, it is important that the generated label and description of the error type can be GENERALIZED to other examples, i.e. avoid references to the particular content of the current example (names, dates, etc): anything related ONLY to this example SHOULD NOT be present in the description and label.

Output answer is the following format: "SHORT LABEL: DESCRIPTION", do not output anything else!

Figure 10: Prompt used for generating a new issue type in the cumulative clustering strategy. The presented version of the prompt is for text LLM outputs, the prompt can be easily changed if JSON outputs are supported by an LLM.

Situation: Two experts are inspecting examples in natural language generation for a particular dataset.

You will be given 2 sentences, which represent the conclusions of the two experts about the same example, i.e. what is a failure case in the example they were given.

Your task is to determine if the experts describe the same failure.

Expert 1 conclusion: E1

Expert 2 conclusion: E2

Do the experts describe the same failure? Output only one word 'Yes' or 'No'.

Figure 11: Prompt used for evaluation. The presented version of the prompt is for text LLM outputs, the prompt can be easily changed if JSON outputs are supported by an LLM.

## H Examples

The following pages present examples of task instances, per-instance analysis, generated error reports, and clustering confusion matrices, for all 12 considered datasets. Clusters of size 1 are shown in confusion matrices but omitted in error reports, for space purposes.

# H.1 Semantic parsing (PIZZA dataset)

## Task illustration

<b>Example user input</b>
This task is about pizza ordering (PIZZAORDER) and drink ordering (DRINKORDER). You will be given a sentence expressing one or more orders, and you have to generate a sequence of structured representations for the orders.
<FORMAT SPECIFICATIONS>
Following the task specifications provided above, generate the structured output corresponding to the following input sentence. The output must be in a single line with all the orders mentioned in the input sentence and must be in a format described above.
Input:
I would like one small combination pizza without onions and a large diet pepsi
<b>Example ground truth answer</b>
(DRINKORDER (DRINKTYPE diet_pepsi ) (NUMBER 1 ) (SIZE large ) ) (PIZZAORDER (NOT (TOPPING onions ) ) (NUMBER 1 ) (SIZE small ) (STYLE combination ) )
<b>Task details</b>
Metric: LLMEval: it uses an LLM, which is provided with the ground truth label(s) and the generated response, and is asked to output 'yes' or 'no', i.e. whether the generated response is correct or not. We assume that the ground truth labels provided in the dataset are correct and cannot be the source of the pipeline failure.

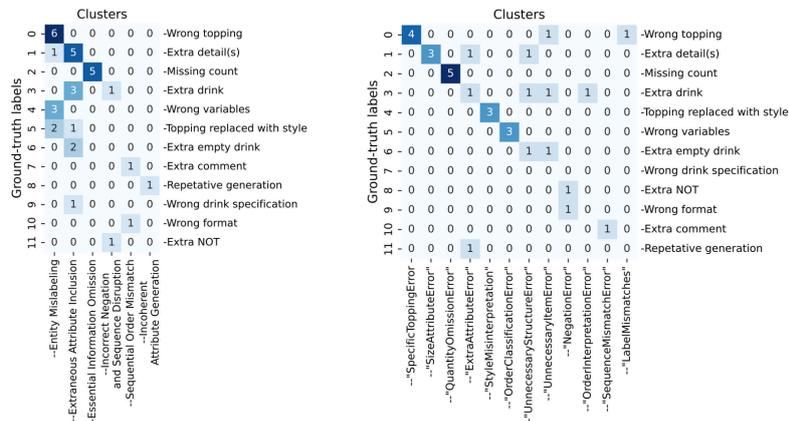
## Examples of per-instance analysis illustration

Human expert	GPT4o (accuracy: 94%)	Eval	Qwen-2.5-7B (accuracy: 88%)	Eval
An extra generated item for drinks with an invalid count 0	The LLM-generated answer incorrectly includes an unnecessary DRINKORDER structure not present in the input.	Yes	The primary issue is the incorrect inclusion of 'DRINKORDER' in the generated answer, as the input sentence does not mention any drinks.	Yes
Missing (NUMBER 1), i.e. the number of ordered pizzas	The generated answer lacks the necessary '(NUMBER 1)' component, indicating the quantity of the pizza, resulting in the deviation from the expected format.	Yes	Summary: The generated response is missing the 'NUMBER' attribute, which specifies the quantity of pizzas ordered, leading to an incomplete structured representation.	Yes
Extra generated detail in a drink, not present in the user request: (SIZE medium )	The failure was due to the incorrect inclusion of a 'SIZE' attribute in the 'DRINKORDER' of the LLM-generated response.	Yes	The LLM-generated answer incorrectly includes the 'SIZE' attribute under both the 'DRINKORDER' and 'PIZZAORDER' structures, violating the specified format where the 'SIZE' attribute should only appear under the 'PIZZAORDER'.	Yes
Extra generated drink not present in the user order: (DRINKORDER (DRINKTYPE pineapple_soda) (NUMBER 1))	The generated response incorrectly interprets the pizza order as a drink order and wrongly negates the specified pizza style.	No	The primary issue is the incorrect identification of the order type, where the LLM-generated answer includes a 'DRINKORDER' instead of a 'PIZZAORDER', leading to a fundamentally flawed response that does not meet the task requirements.	No
An extra comment is generated in addition to the correct order.	The generated response is incorrect due to the swapped sequence of 'sprite' orders, mismatching the order of entries specified in the ground truth labels.	No	The primary failure in the LLM-generated answer is the incorrect count of large sprite drinks, where the LLM incorrectly reports two large sprites instead of the correct one.	No

## An example of a final report generated by LLM-as-a-qualitative-judge (GPT4o)

- \* **Entity Mislabeling:** The generated answer contains incorrect labeling of entities, leading to a mismatch with the expected ground truth label. (12)
- \* **Extraneous Attribute Inclusion:** The generated answer includes attributes that are not part of the ground truth requirements, leading to discrepancies in the expected output. (12)
- \* **Essential Information Omission:** The generated response fails to include critical quantifiable details necessary for completing a structured output. (5)
- \* **Sequential Order Mismatch:** The generated response fails to match the required sequence of elements as specified in the ground truth, leading to erroneous outputs. (2)
- \* **Incorrect Negation and Sequence Disruption:** The generated response incorrectly negates an intended attribute and alters the sequence of elements, resulting in a deviation from the expected output. (2)

## Confusion matrices for issue type clustering for GPT4o (left) and Qwen-2.5-7B (right)



## H.2 Long context QA (Elitr-Bench dataset)

### Task illustration

Example user input
You are given a meeting transcript. Your task is to answer a question given below about this meeting. Do not output anything else.
<TRANSCRIPT>
Question:
Who is interested in applying to a PhD?
Example ground truth answer
An Indian student known by [PERSON18]
Task details
Metric: LLMEval: it uses an LLM, which is provided with the ground truth label(s) and the generated response, and is asked to output 'yes' or 'no', i.e. whether the generated response is correct or not. We assume that the ground truth labels provided in the dataset are correct and cannot be the source of the pipeline failure.

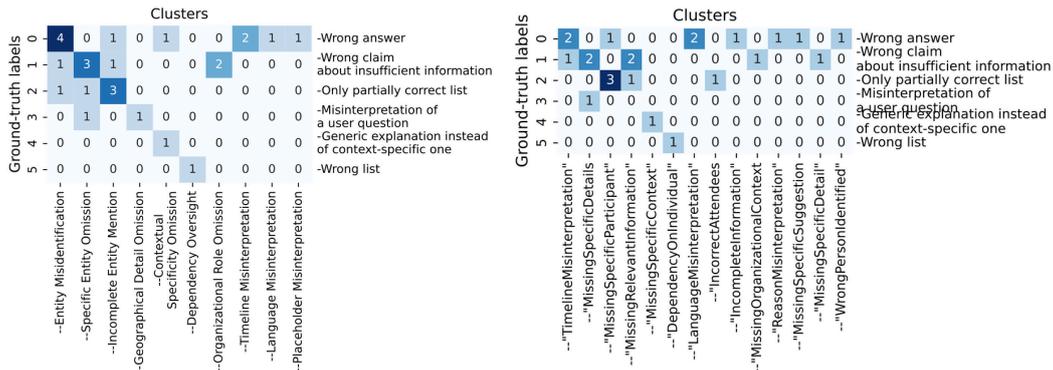
### Examples of per-instance analysis illustration

Human expert	GPT4o (accuracy: 69%)	Eval	Qwen-2.5-7B (accuracy: 65%)	Eval
An LLM outputs a partially correct list of persons.	The generated response failed to mention [PERSON15] as involved in the specific call, as indicated in the ground truth labels.	Yes	The generated response fails to identify a specific person who will have a call to discuss multi-source experiments, instead providing details about the sequence of discussions and tasks.	No
Wrong answer: end of March instead of end of February.	The LLM failed to extract and correctly interpret the timeline for de-identification from the meeting transcript, leading to an incorrect response.	Yes	Summary: The generated response incorrectly specifies the completion date as the end of March instead of the end of February as per the ground truth label.	Yes
An LLM provides a generic explanation of a term "ladder climbing" instead of providing an explanation contextualized in the discussion, i.e. STT-specific process with HTML tables	The generated answer fails to address the specific context of 'ladder climbing' as discussed in the meeting, focusing instead on a generic definition not tailored to the transcript.	Yes	The generated answer failed to correctly interpret "ladder climbing" in the context of the meeting transcript, instead providing a general explanation of model improvement techniques unrelated to the specific STT system evaluation process mentioned in the ground truth label.	Yes
An LLM mistakenly claims that the information provided in the context is insufficient to answer the given question.	The LLM-generated answer failed to extract and correctly report the specific number of people mentioned in the program committee from the transcript.	No	Summary: The LLM failed to extract and provide the specific number of people in the program committee as stated in the ground truth label.	No
The generated response does not answer the given question and talks about a different topic.	The LLM-generated answer failed to identify and cite the geographical locations of the meeting participants as required by the question.	No	The generated response failed to identify places from which participants spoke in the meeting, instead providing a summary of the discussion topics and participant actions.	No

### An example of a final report generated by LLM-as-a-qualitative-judge (GPT4o)

- \* **Entity Misidentification:** The generated response incorrectly identifies or attributes an action to an entity, leading to a mismatch with the expected answer. (6)
- \* **Incomplete Entity Mention:** The generated response fails to include all entities required by the context, resulting in an incomplete answer. (5)
- \* **Specific Entity Omission:** The generated response fails to address the request for information about a specific entity, resulting in an incomplete answer that lacks relevant details. (5)
- \* **Organizational Role Omission:** The generated response fails to accurately acknowledge the participation or role of entities in a collaborative activity, resulting in an evasive or incomplete answer. (2)
- \* **Contextual Specificity Omission:** The generated response fails to integrate specific contextual details relevant to a particular discussion, resulting in an overly generic answer that does not align with the nuanced requirements of the query. (2)
- \* **Timeline Misinterpretation:** The generated response fails to accurately identify and interpret chronological information from the given context, resulting in an incorrect timeline being communicated. (2)

### Confusion matrices for issue type clustering for GPT4o (left) and Qwen-2.5-7B (right)



### H.3 Detoxification (MultiDetox dataset)

#### Task illustration

<b>Example user input</b>
You are a professional moderator of content on social networks. You will be given an input sentence and your task is to output a refined sentence, which satisfies the following conditions: * if the given sentence is not toxic, then copy the input sentence * if the given sentence is toxic, then output a new sentence with the meaning as close as possible to the given sentence, in the same language as the given sentence.
Given sentence: {toxic sentence}
Output a refined sentence. Only output one refined sentence, do not output anything else.
<b>Example ground truth answer</b>
{detoxified sentence}
<b>Task details</b>
Metric: LLMEval: it uses an LLM, which is provided with the ground truth label(s) and the generated response, and is asked to output 'yes' or 'no', i.e. whether the generated response is correct or not. While ground truth labels provide correct high-quality answers, it does NOT mean that an LLM has to guess 100% all the information from the ground truth; LLM also may provide a valid alternative answer.

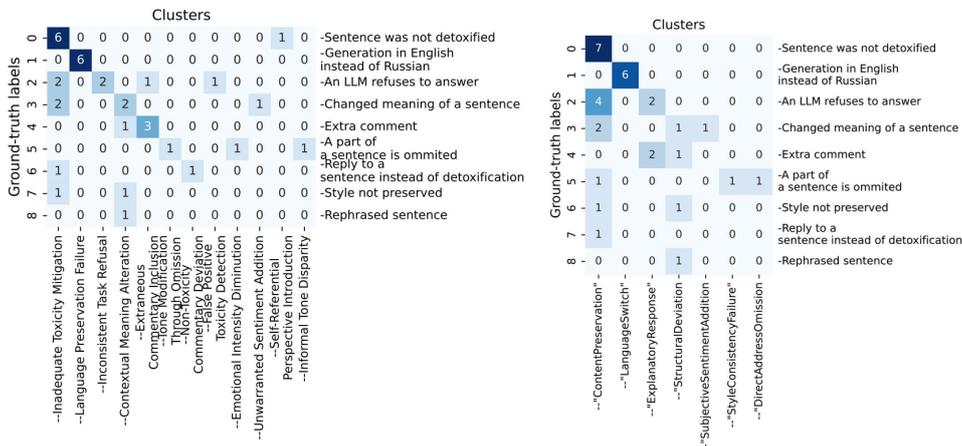
#### Examples of per-instance analysis illustration

Human expert	GPT4o (accuracy: 78%)	Eval	Qwen-2.5-7B (accuracy: 61%)	Eval
The generated sentence is in English instead of Russian.	The generated response failed by translating the Russian input into English, violating the requirement for language consistency.	Yes	The primary failure is the language mismatch between the input and output sentences, leading to a significant deviation from the required criteria.	Yes
A part of an input sentence was omitted, i.e. "Kid"	The LLM-generated response was not accepted because it omitted the word "kid," altering the tone of direct address in the sentence.	Yes	The failure lies in the loss of the specific reference to "kid" in the refined sentence, which changes the intended meaning and context of the original toxic statement.	Yes
The sentence was not fully detoxified.	The generated sentence retains offensive language from the input, failing to fully refine it into a less toxic version as required.	Yes	The primary issue is that the generated response softens the tone of the original toxic sentence, failing to preserve its intended harshness and directness.	No
An LLM refuses to answer the user's request due to its safety guideline.	The LLM response did not refine the toxic input sentence as required by the task.	No	The primary failure is that the generated answer does not provide a non-toxic sentence with a similar meaning to the input, instead offering a generic and irrelevant response.	No
Model did not follow the instructions to only output the refined sentence, i.e. it also output a comment.	The generated response failed because it added assumptions about the tone and content not present in the original sentence, violating the requirement to maintain the original meaning closely.	No	The failure lies in the introduction of an unnecessary opinion and the addition of subjective language, which deviates from the requirement to maintain the exact meaning of the original sentence in a more refined manner without altering its core message.	Yes

#### An example of a final report generated by LLM-as-a-qualitative-judge (GPT4o)

- \* **Inadequate Toxicity Mitigation:** The system fails to reduce or neutralize the presence of toxic sentiment in the generated response while maintaining the original intended meaning. (12)
- \* **Language Preservation Failure:** The system fails to maintain the original language of the input when generating a response, resulting in an unwanted translation or language shift. (6)
- \* **Contextual Meaning Alteration:** The system significantly changes the original meaning of the input while attempting to refine it, resulting in a response that deviates from the intended context. (5)
- \* **Extraneous Commentary Inclusion:** The system introduces additional commentary or information that was not requested, instead of providing a direct and concise response to the task. (4)
- \* **Inconsistent Task Refusal:** The system declines to perform a task based on content sensitivity without explicitly following the task instructions, resulting in a failure to generate the required output. (2)

#### Confusion matrices for issue type clustering for GPT4o (left) and Qwen-2.5-7B (right)



## H.4 Translation en-ru (WMT'22 dataset)

### Task illustration

<b>Example user input</b>
Translate a sentence from English to Russian, do not output anything else:  I can see here that there is no rider assigned yet to your order.
<b>Example ground truth answer</b>
Я вижу, что на ваш заказ еще не назначен водитель.
<b>Task details</b>
Metric: LLMEval: it uses an LLM, which is provided with the ground truth label(s) and the generated response, and is asked to output 'yes' or 'no', i.e. whether the generated response is correct or not. While ground truth labels provide correct high-quality answers, it does NOT mean that an LLM has to guess 100% all the information from the ground truth; LLM also may provide a valid alternative answer.

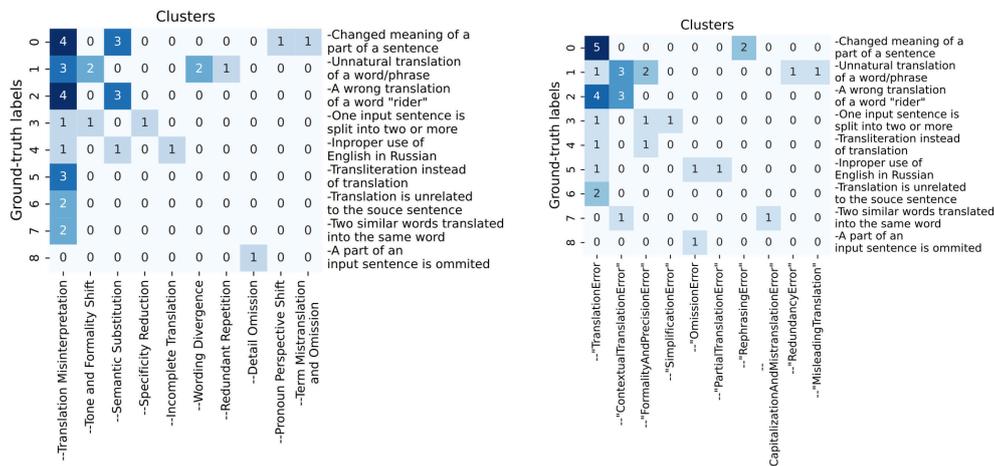
### Examples of per-instance analysis illustration

Human expert	GPT4o (accuracy: 66%)	Eval	Qwen-2.5-7B (accuracy: 71%)	Eval
Mistranslation: the wrong word is used to translate the word "rider".	The error was due to the LLM's mistranslation of 'rider', using a term more aligned with racing rather than the intended meaning of 'driver' for transportation services.	Yes	The primary issue lies in the translation of "rider" as "гонщиком" instead of "водителем", which is less precise and potentially misleading in the context of the original sentence.	Yes
Mistranslation: the wrong word is used to translate the word "generic".	The LLM-generated translation uses a less common equivalent for 'generic' and a less standard term for 'cord', reducing clarity.	Yes	The primary failure in the generated response is the incorrect translation of "generic" as "родовой" instead of "универсальный."	Yes
Inappropriate terminology: the word "rider" is transliterated instead of being translated into a proper Russian term	The LLM incorrectly translated 'rider' as 'райдер' instead of 'водитель', leading to an inaccurate translation.	Yes	The primary issue is the incorrect use of "райдер" instead of "водитель" in the translation, which slightly deviates from the ground truth label and reduces the accuracy of the response.	Yes
One input sentence is split into two output sentences	The LLM mistranslated a key phrase, altering the meaning of the sentence.	No	Summary: The LLM failed to correctly translate the phrase "once the order has been placed" to "после того, как заказ был размещен," resulting in a significant change in the meaning of the sentence.	No
Two words similar in meaning, Repair/restore, are translated into the same word in Russian.	The failure is due to incorrect capitalization and slight mistranslation in the LLM-generated response.	No	The primary issue is the incorrect placement of the verb "коснитесь" (tap/touch) between the two phrases, which disrupts the intended meaning and flow of the command.	No

### An example of a final report generated by LLM-as-a-qualitative-judge (GPT4o)

- \* **Translation Misinterpretation:** The system incorrectly translates specific terms or phrases, leading to inaccurate representations of the original meaning in the generated answer. (20)
- \* **Semantic Substitution:** The generated translation substitutes critical terms with incorrect alternatives, leading to a change in the intended meaning of the sentence. (7)
- \* **Tone and Formality Shift:** The generated translation alters the tone and level of formality compared to the ground truth, leading to a failure in matching the expected evaluative criteria. (3)
- \* **Wording Divergence:** The generated answer diverges in wording from the ground truth, causing a mismatch that leads to incorrect evaluation outcomes. (2)

### Confusion matrices for issue type clustering for GPT4o (left) and Qwen-2.5-7B (right)



## H.5 Instruction following (FLASK dataset)

### Task illustration

<b>Example user input</b>
What is the name of the actress who played the role of Rachel Green in the popular TV show "Friends"?
<b>Example ground truth answer</b>
The actress who played the role of Rachel Green in the popular TV show "Friends" is Jennifer Aniston.
<b>Task details</b>
Metric: LLM Eval: it uses an LLM, which is provided with the ground truth label(s) and the generated response, and is asked to output 'yes' or 'no', i.e. whether the generated response is correct or not.

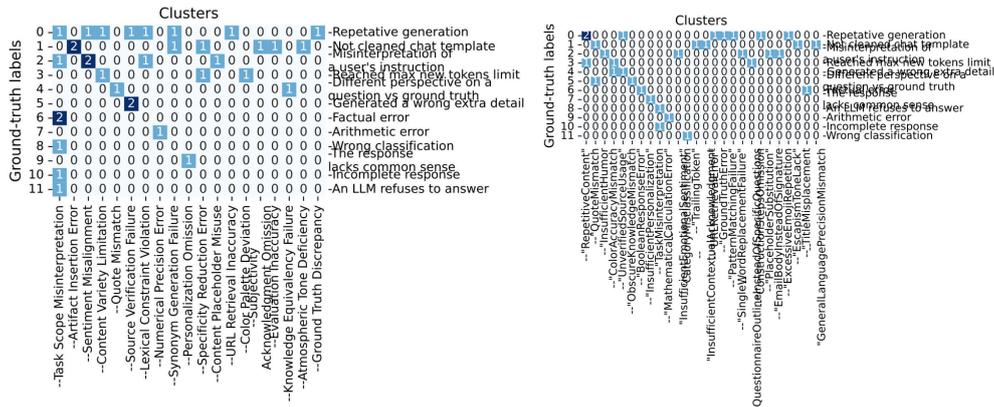
### Examples of per-instance analysis illustration

Human expert	GPT4o (accuracy: 56%)	Eval	Qwen-2.5-7B (accuracy: 47%)	Eval
An LLM generated a wrong reason for a meeting, i.e. it was not discussed in the conversation.	The generated answer was rejected because it included an unsupported detail about the meeting's purpose that was not in the ground truth.	Yes	Summary: The generated response included an unnecessary and incorrect detail about the reason for the meeting, which was not present in the input or the ground truth labels.	Yes
The provided list is not full and leaves out skills as technical writing, debugging etc	The generated response listed technical tools and languages instead of general software engineering skills, deviating from the ground truth.	Yes	Summary: The generated response failed to list the specified skills and instead focused on technical tools and platforms, missing the core requirement of the question.	Yes
Wrong classification: the correct label is "demo", not "news".	The LLM misclassified a tweet about a tutorial as 'news' instead of the correct category of 'demo' by focusing on the informational aspect rather than its promotional intent.	Yes	The LLM failed to correctly identify the tweet as a "demo" and instead classified it as "news", misunderstanding the nature of the content as a promotional demonstration rather than a report on a recent event or development.	Yes
The response lacks common sense: it should be shorter and should not say that we are proud of our customer policies.	The generated response failed due to a lack of personalized acknowledgment and appreciation for the customer's repeat visits.	No	The primary failure in the LLM-generated response is its lack of personal engagement and direct gratitude, failing to match the warm and encouraging tone of the ground truth label.	No
Generation loops in repeating the same phrase	The LLM-generated answer incorrectly focuses on a random list of materials rather than the necessary construction steps, missing the core functional details of the steam engine project.	No	The primary failure is that the generated response does not provide any of the required steps for constructing a steam engine using an empty beer can, instead listing unrelated hardware components.	No

### An example of a final report generated by LLM-as-a-qualitative-judge (GPT4o)

- Task Misalignment:** The generated response focuses on an incorrect or unintended task, deviating from the specific task or instruction outlined in the query, resulting in irrelevant or inappropriate output. (5)
- Content Variety Restriction:** The generated response fails to capture the richness and diversity of themes found in the reference material, resulting in a focus that is overly narrow or singular in perspective. (4)
- Phrase Granularity Mismatch:** The generated response provides multi-word phrases instead of the required single-word inputs, leading to discrepancies in response granularity and format expectations. (3)
- Language Specificity Deficiency:** The generated response employs vague or less precise language compared to the ground truth, resulting in a failure to meet specific language expectations or instructions. (3)
- Sentiment Misalignment with Emojis:** The generated response does not accurately reflect the intended emotional sentiment due to inappropriate or missing emojis, leading to a mismatch in tone or context. (2)
- Unverified Content Recurrence:** The generated response repeatedly includes sentences or information not substantiated by the provided source material, leading to issues with content accuracy and diversity. (2)
- Attribute Annotation Error:** The generated response inaccurately assigns or introduces attributes that are absent from the ground truth or input data, leading to errors in attribute extraction or assignment tasks. (2)
- Evaluation Misjudgment:** The automated evaluation process inaccurately labels a correct generated response as incorrect due to a misalignment between the evaluation criteria and the correct output. (2)
- Artifact Inclusion:** The generated response contains unintended or residual text artifacts, which disrupt the intended output format and clarity. (2)

### Confusion matrices for issue type clustering for GPT4o (left) and Qwen-2.5-7B (right)



## H.6 Grade school math (GSM8K dataset)

### Task illustration

Example user input
James decides to run 3 sprints 3 times a week. He runs 60 meters each sprint. How many total meters does he run a week?
Example ground truth answer
He sprints $3 \times 3 = 9$ times So he runs $9 \times 60 = 540$ meters ### 540
Task details
Metric: Exact Match: it checks if the answer highlighted in <code>\emph{}</code> matches exactly the ground truth answer. We assume that the ground truth labels provided in the dataset are correct and cannot be the source of the pipeline failure. The LLM answer was obtained using retrieval-augmented generation and chain-of-thought prompt.

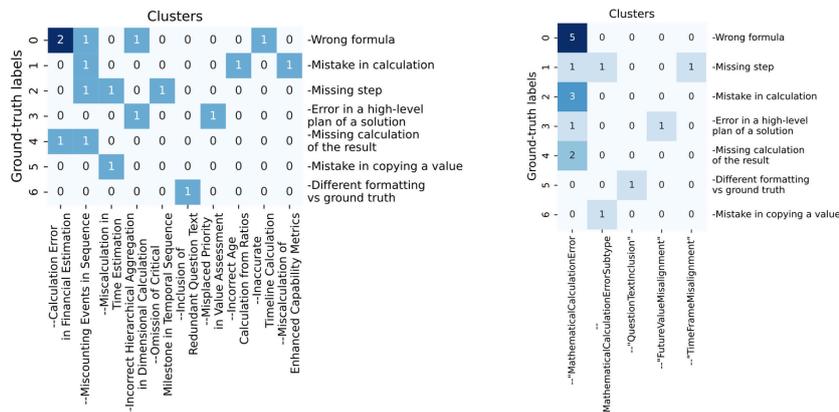
### Examples of per-instance analysis illustration

Human expert	GPT4o (accuracy: 88%)	Eval	Qwen-2.5-7B (accuracy: 82%)	Eval
A wrong formula in step 3: should be $20 * 2$ instead of $20 * 10$ .	The LLM miscalculated the production cost by incorrectly applying the supply cost to each individual candle, resulting in an erroneous net profit calculation.	Yes	The LLM incorrectly calculated the cost per candle as \$10.00 instead of \$1.00, leading to an incorrect total cost and net profit.	Yes
The last step misses a multiplication of 126 by 6 and hence outputs a wrong final amount.	The LLM-generated response miscalculated the total annual cost by failing to multiply the discounted monthly rate by the number of months in the second half of the year.	Yes	The LLM failed to correctly compute the total cost for the second half of the year, resulting in an incorrect final total amount of \$966 instead of the correct \$1596.	Yes
Arithmetic error in step 3: should be 1200 instead of 1500	The LLM miscalculated the enhanced throwing distance with the gemstone, leading to an incorrect safe distance being reported.	Yes	The primary failure in the LLM-generated answer is the incorrect calculation of the throwing distance with the gemstone, resulting in a distance of 1500 feet instead of the correct 1200 feet.	Yes
Difference in formatting between the generated answer and the ground truth answer: The final answer is represented with zeros after the decimal point, .00, while the ground truth answer is represented in an integer format, resulting in an absence of a verbatim match.	The generated answer fails the exact match metric because it unnecessarily includes the verbatim question text, causing a mismatch with the expected concise format of the ground truth.	No	The failure case is due to the lack of explicit highlighting of the final total (\$7.00) in the same manner as the ground truth label.	No

### An example of a final report generated by LLM-as-a-qualitative-judge (GPT4o)

- \* Miscounting Events in Sequence:** The LLM makes a mistake in counting or aggregating the number of occurrences in a sequence of events, leading to inaccurate calculations or assertions. (4)
- \* Calculation Error in Financial Estimation:** The LLM incorrectly performs computations related to financial metrics, such as costs or profits, due to improper application of inputs or misunderstanding of financial principles. (3)
- \* Miscalculation in Time Estimation:** The LLM incorrectly calculates or estimates time-related metrics, such as progress percentages or total time, due to erroneous input handling or flawed understanding of the temporal aspects involved. (2)
- \* Incorrect Hierarchical Aggregation in Dimensional Calculation:** The LLM miscalculates dimensional metrics by aggregating levels in the wrong direction, leading to errors in determining total and average values. (2)

### Confusion matrices for issue type clustering for GPT4o (left) and Qwen-2.5-7B (right)



## H.7 Factoid QA in Russian (MKQA dataset)

### Task illustration

<b>Example user input</b>
какой игрок сделал хет-трик в обеих подачах контрольного матча
<b>Example ground truth answer</b>
['Питер Сиддл', 'Джимми Мэтьюз']
<b>Task details</b>
Substring match: it measures whether any of the ground truth responses is included in the model response. We assume that the ground truth labels provided in the dataset are correct and cannot be the source of the pipeline failure. The LLM answer was obtained using retrieval-augmented generation.

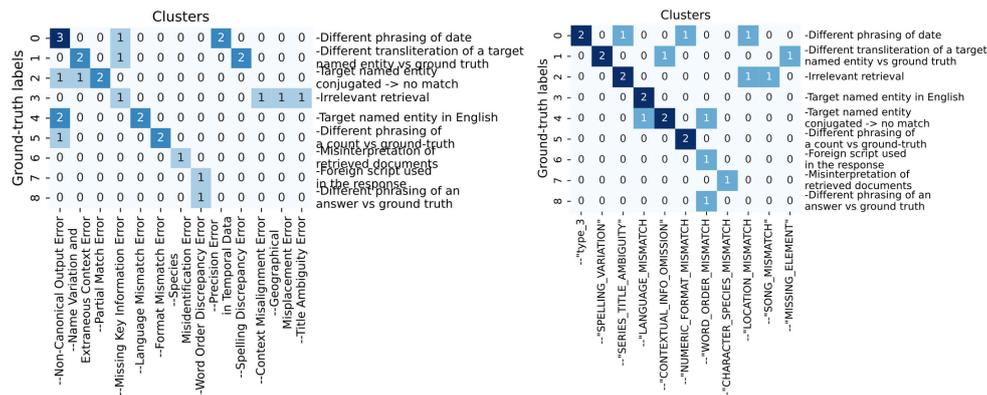
### Examples of per-instance analysis illustration

Human expert	GPT4o (accuracy: 69%)	Eval	Qwen-2.5-7B (accuracy: 66%)	Eval
The correct answer is conjugated in a sentence resulting in an absence of a verbatim match	The LLM response failed the Substring match test because it included additional context, making it not a strict substring of any of the ground truth labels.	Yes	The failure case is due to the substring match metric not recognizing the ground truth labels "Бразилия" and "Португалия" within the generated answer, even though the answer is factually correct.	Yes
different phrasing of the target date in the generated response compared to the ground-truth label	The generated answer was not accepted due to failing a strict substring match metric, as it included extra textual information alongside the correct date.	Yes	The generated response includes unnecessary details about the publication in different countries, missing the specific date '1997-06-26' that the substring match metric requires.	Yes
different phrasing of the target number (written versus digit) in the generated response compared to the ground-truth label	The LLM response uses words instead of the expected numeric format, causing a mismatch with the ground truth label.	Yes	The generated answer lacks the explicit mention of the number "3.0" which is required by the ground truth label, even though the correct years of victory were provided.	Yes
the target named entity written in English rather than in Russian, while the ground truth is written in Russian	The generated response includes extra information about Instagram's own account, which prevents it from matching the ground truth label focused solely on Cristiano Ronaldo.	No	The generated response included unnecessary details about the Instagram account's follower count, which caused it to fail the substring match criterion despite being factually correct.	No
different transliteration of the target named entity (into the Cyrillic alphabet) in the generated response compared to the ground-truth label	The generated response only contains one of the two required ground truth names, failing the substring match metric.	No	The generated answer includes unnecessary details that are not present in the ground truth labels, leading to a mismatch in the substring match evaluation.	No

### An example of a final report generated by LLM-as-a-qualitative-judge (GPT4o)

- \* **Non-Canonical Output Error:** The generated response includes extraneous words or adopts an unconventional format that deviates from the expected canonical representation, hindering accurate evaluation against the ground truth. (7)
- \* **Missing Key Information Error:** The generated response omits essential information that is present in the ground truth, resulting in an incomplete answer. (3)
- \* **Name Variation and Extraneous Context Error:** The generated response contains minor variations in names and includes additional contextual information not present in the ground truth, leading to a mismatch in evaluation. (3)
- \* **Format Mismatch Error:** The generated response uses an incorrect format, such as substituting words for the expected numeric format, leading to a discrepancy with the ground truth label. (2)
- \* **Precision Error in Temporal Data:** The generated response provides an approximate temporal detail instead of the exact value required, leading to a mismatch with the specific ground truth information. (2)
- \* **Word Order Discrepancy Error:** The generated response contains the correct information but in a different word order than the ground truth, resulting in a non-match by the substring match metric. (2)
- \* **Spelling Discrepancy Error:** The generated response contains a spelling variation in critical names or terms that results in a failure to match the ground truth label, despite conveying the intended information. (2)
- \* **Language Mismatch Error:** The generated response is in a different language than the ground truth, causing a failure in matching the correct answer in automatic evaluation. (2)
- \* **Partial Match Error:** The generated response includes the correct answer within a larger text, leading to a failure in exact substring matching during evaluation. (2)

### Confusion matrices for issue type clustering for GPT4o (left) and Qwen-2.5-7B (right)



## H.8 Biomedical QA (BioASQ dataset)

### Task illustration

<b>Example user input</b>
Which protein interacts with the Ragulator-RAG GTPases to control mTOR activity?
<b>Example ground truth answer</b>
['Amino Acid-Dependent Lysosomal Membrane Protein SLC38A9']
<b>Task details</b>
Metric: Substring match: it measures whether any of the ground truth responses is included in the model response. We assume that the ground truth labels provided in the dataset are correct and cannot be the source of the pipeline failure. The LLM answer was obtained using retrieval-augmented generation.

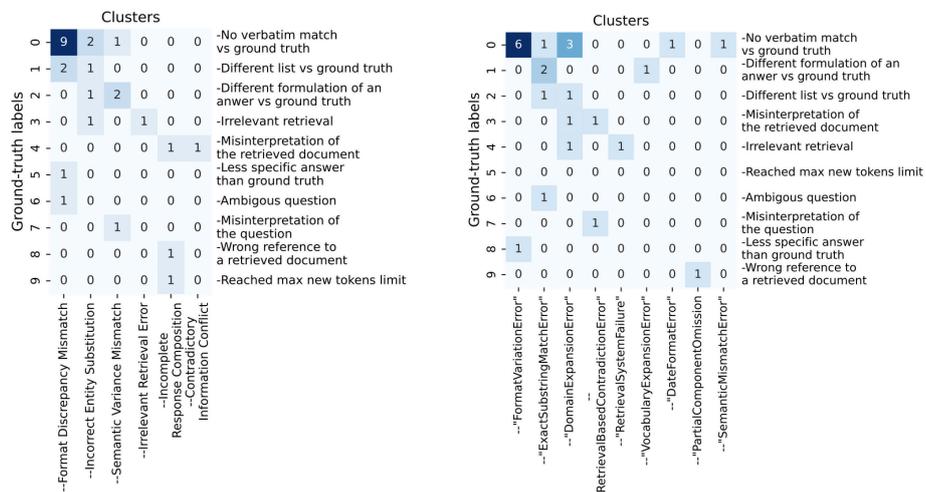
### Examples of per-instance analysis illustration

Human expert	GPT4o (accuracy: 67%)	Eval	Qwen-2.5-7B (accuracy: 63%)	Eval
The generated response as well as retrieved documents contain a different list of diseases than the ground truth list	The generated answer includes 'multiple sclerosis', a disease not present in the ground truth labels, causing a mismatch.	Yes	The primary failure in this example is the omission of "obsessive-compulsive disorder" from the list of diseases that can be treated with focused ultrasound thalamotomy.	Yes
match metric error: correct answer is judged wrong because of the absence of the verbatim match	The generated answer failed the substring match metric because it listed the correct tools separately with additional details, rather than as a continuous substring matching the ground truth label.	Yes	The failure case arises from the detailed description of 'EPD' in the generated answer, which does not strictly match the concise format required by the ground truth label.	Yes
Misinterpretation of the retrieved document: Document 4 does not mention atemoya juice while the generated response points to this document as a reference	The LLM-generated answer contradicts the ground truth 'no' due to reliance on potentially misleading retrieved information.	Yes	The LLM-generated answer incorrectly affirms that atemoya juice inhibits the CYP3A4 enzyme, contradicting the ground truth label which states 'no'.	Yes
Ambiguous question: the generated response provides one way to answer the question and the ground truth label provides another way.	The LLM-generated response failed to include the exact term 'Leishmania species' required for a substring match with the ground truth.	No	The generated answer failed to include the term "Leishmania Species," which is required by the ground truth label for the substring match to succeed.	No
The generated response is cropped	The LLM-generated answer failed due to being too detailed and not exactly matching the concise ground truth label 'yes' required by the evaluation metric.	No	The generated response fails to directly affirm the existence of data supporting the improvement of symptom severity in ALS patients through TRH administration, as required by the ground truth label.	No

### An example of a final report generated by LLM-as-a-qualitative-judge (GPT4o)

- \* **Format Discrepancy Mismatch:** The generated answer fails to match the ground truth due to differences in formatting or wording while retaining the same semantic meaning. (13)
- \* **Incorrect Entity Substitution:** The generated answer inaccurately replaces or incorporates an incorrect entity or detail, deviating from the ground truth despite correct interpretation of the retrieved content. (5)
- \* **Semantic Variance Mismatch:** The generated answer diverges in semantic meaning from the ground truth, leading to a mismatch that goes beyond mere formatting or wording differences. (4)
- \* **Incomplete Response Composition:** The generated answer omits required components of the ground truth, resulting in a response that only partially addresses the complete expected answer. (3)

### Confusion matrices for issue type clustering for GPT4o (left) and Qwen-2.5-7B (right)



## H.9 Lifestyle forum QA (RobustQA dataset)

### Task illustration

<b>Example user input</b>
how to reach this old hotel stuck in the middle of narita airport runways?
<b>Example ground truth answer</b>
['Upon arriving at Higashi-Narita Station, you should enter a tunnel and expect to see the pension on your left, as you emerge from the tunnel.']
<b>Task details</b>
Metric: LLMEval: it uses an LLM, which is provided with the ground truth label(s) and the generated response, and is asked to output 'yes' or 'no', i.e. whether the generated response is correct or not. While ground truth labels provide correct high-quality answers, it does NOT mean that an LLM has to guess 100% all the information from the ground truth; LLM also may provide a valid alternative point of view. The LLM answer was obtained using retrieval-augmented generation.

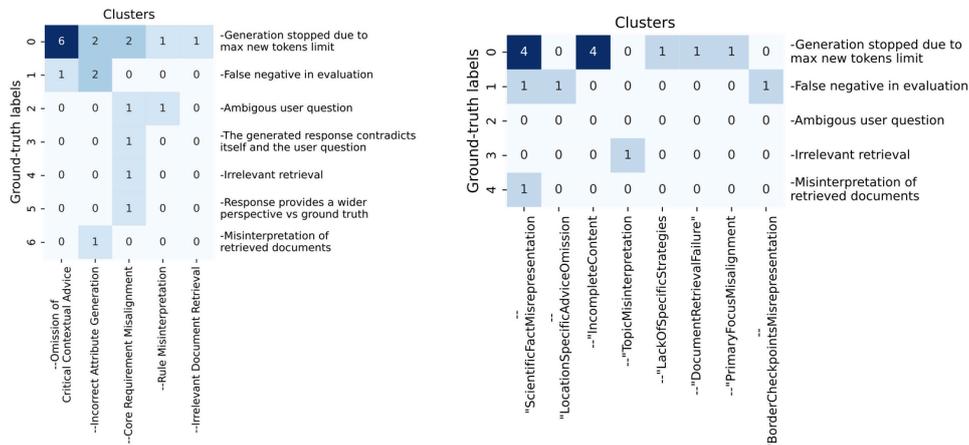
### Examples of per-instance analysis illustration

Human expert	GPT4o (accuracy: 19%)	Eval	Qwen-2.5-7B (accuracy: 14%)	Eval
The generated response provides a wider perspective on the question, offering more options to a user than a ground truth response, which potentially led to the response being not accepted by the automatic evaluation.	The LLM-generated response addresses alternatives for condensed milk instead of the method to make homemade condensed milk as per the ground truth, causing a mismatch.	Yes	The generated response failed to provide a direct, actionable alternative to condensed milk as specified in the ground truth label, instead offering general information and references to external documents.	No
The generated response contradicts itself and the user question: it says to use the toilet brush while it is asked not to use it	The LLM-generated answer fails to address the core requirement of cleaning without a toilet brush, contradicting the question.	Yes	The failure case is that the generated answer assumes the presence of a toilet brush, which contradicts the question's premise of not having a toilet brush available.	Yes
Misinterpretation of the retrieved documents: correct information in the retrieved documents is synthesized into a wrong answer	The LLM answer incorrectly claims that cooling a boiled egg quickly can revert it to its natural state, which contradicts known scientific facts and the provided ground truth.	Yes	The primary failure case is the incorrect assertion that there are ways to bring a cooked egg back to its natural state, contradicting the ground truth label which states that such a straightforward reversal is not possible.	Yes
Error in evaluation: correct answer judged as wrong	The generated response incorrectly suggests that liqueurs can lack added sugar, which contradicts the essential defining characteristic of liqueurs.	No	The primary issue is the misclassification of vermouth as a non-distilled fortified wine, which contradicts the correct definition provided in the ground truth label.	Yes
generation was stopped too early because of the reached maximum new tokens limit	The generated answer provides an incomplete and misaligned overview compared to the detailed options and context provided in the ground truth.	No	The generated response fails to include key information about Just Right Menus, MacGourmet, SousChef, Yum, and the use of simple text files with Dropbox synchronization, which are explicitly mentioned in the ground truth labels.	No

### An example of a final report generated by LLM-as-a-qualitative-judge (GPT4o)

- \* **Omission of Critical Contextual Advice:** The response fails to include essential contextual advice or recommendations relevant to the specific needs or circumstances of the subject matter, focusing instead on generic or peripheral information. (7)
- \* **Core Requirement Misalignment:** The response fails to align with the central requirement or intent of the query, resulting in a response that contradicts or overlooks the primary objective sought by the user. (6)
- \* **Incorrect Attribute Generation:** The generated response provides incorrect or contradictory information about a defining attribute of the discussed subject, conflicting with established facts. (5)
- \* **Rule Misinterpretation:** The response is based on a misinterpretation of formal rules or guidelines, leading to an inaccurate conclusion or recommendation. (2)

### Confusion matrices for issue type clustering for GPT4o (left) and Qwen-2.5-7B (right)



## H.10 Writing forum QA (RobustQA dataset)

### Task illustration

<b>Example user input</b>
what is the difference between everyday and every day?
<b>Example ground truth answer</b>
["The term 'everyday' is an adjective that describes something as commonplace or routinely/typically used, and 'every day' is an adverb which refers to something that occurs every single day."]
<b>Task details</b>
Metric: LLM Eval: it uses an LLM, which is provided with the ground truth label(s) and the generated response, and is asked to output 'yes' or 'no', i.e. whether the generated response is correct or not. While ground truth labels provide correct high-quality answers, it does NOT mean that an LLM has to guess 100% all the information from the ground truth; LLM also may provide a valid alternative point of view. The LLM answer was obtained using retrieval-augmented generation and chain-of-thought prompt.

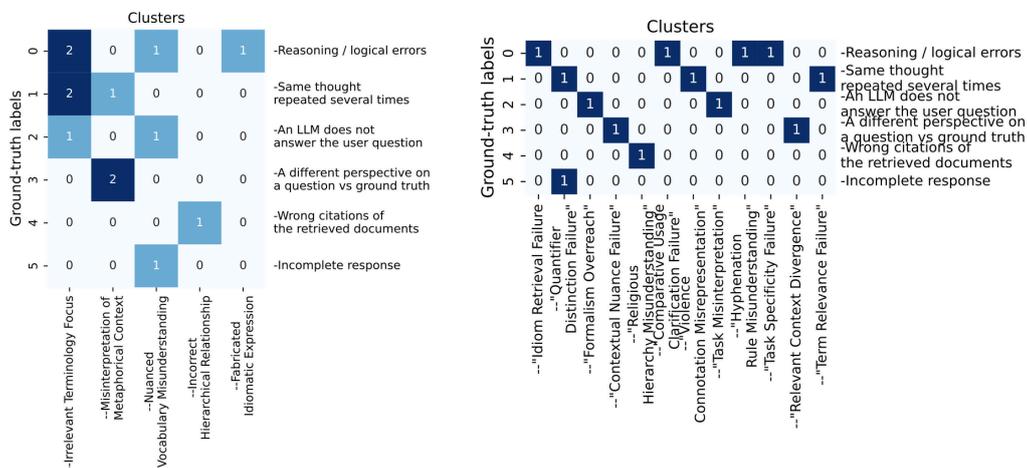
### Examples of per-instance analysis illustration

Human expert	GPT4o (accuracy: 31%)	Eval	Qwen-2.5-7B (accuracy: 31%)	Eval
The generated response is incomplete: it misses a possible interpretation of "a couple" as an indefinite small number	The LLM answer failed to appropriately highlight the quantitative meanings of "a couple" and "a few", focusing too specifically on contextual uses like relationships.	Yes	The primary failure in the LLM-generated response is the lack of emphasis on the shared and overlapping value ranges of "a few" and "several," as well as the contextual flexibility of these terms, which are crucial aspects highlighted in the ground truth labels.	No
The generated answer is correct but considers less options than are present in the ground-truth label.	The generated response fails to capture the nuanced connotations of "non-geek" discussed in the ground truth labels and offers terms that are less contextually relevant or specific.	Yes	The primary failure in the LLM-generated answer is its omission of several key terms and the lack of contextual nuance present in the ground truth labels, leading to an incomplete and less comprehensive response.	Yes
An LLM does not answer the user question	The generated response failed because it misinterpreted the question, addressing the wrong context by focusing on sentence construction rather than the correct usage of 'to-do' or 'todo'.	Yes	The LLM-generated answer fails to address the question about the use of "to-do" or "todo" as a single word, instead providing an unrelated sentence.	Yes
An LLM makes many reasoning mistakes, including wrong reciting of information from the provided documents, wrong reasoning steps, and vague final conclusion.	The generated response lacks a clear explanation of the nuanced usage and implication of 'yet another' compared to 'another'.	No	The failure lies in the LLM's inability to capture the nuanced meaning of "yet another," specifically its implication of repetition and potential annoyance, which is crucial for understanding the difference between "yet another" and "another."	No
Wrong citations of the provided documents lead to a wrong final answer	The generated answer incorrectly states that a basilica is subordinate to a bishop, conflicting with the established relationship where the cathedral serves as the bishop's seat.	No	The primary failure in the generated response is the misinterpretation of the key definitions of basilica and cathedral, particularly regarding the role of the bishop and the hierarchical relationship between these types of churches.	No

### An example of a final report generated by LLM-as-a-qualitative-judge (GPT4o)

- \* **Irrelevant Terminology Focus:** The generated response incorrectly prioritizes irrelevant linguistic details over key terms or concepts, resulting in an inaccurate representation of the main subject or context. (5)
- \* **Misinterpretation of Metaphorical Context:** The generated response misinterprets metaphorical or emotional nuances in retrieved documents, resulting in a divergence from the intended thematic context. (3)
- \* **Nuanced Vocabulary Misunderstanding:** The generated response fails to articulate subtle distinctions in vocabulary usage, leading to a shallow or misleading interpretation of context-specific language nuances. (3)

### Confusion matrices for issue type clustering for GPT4o (left) and Qwen-2.5-7B (right)



## H.11 Search engine queries (SearchQA dataset)

### Task illustration

<b>Example user input</b>
I'm going to hop in my Volvo for a trip from Jokkmokk to Skovde in this country
<b>Example ground truth answer</b>
['Sweden']
<b>Task details</b>
Metric: LLMEval: it uses an LLM, which is provided with the ground truth label(s) and the generated response, and is asked to output 'yes' or 'no', i.e. whether the generated response is correct or not. We assume that the ground truth labels provided in the dataset are correct and cannot be the source of the pipeline failure. The LLM answer was obtained using retrieval-augmented generation.

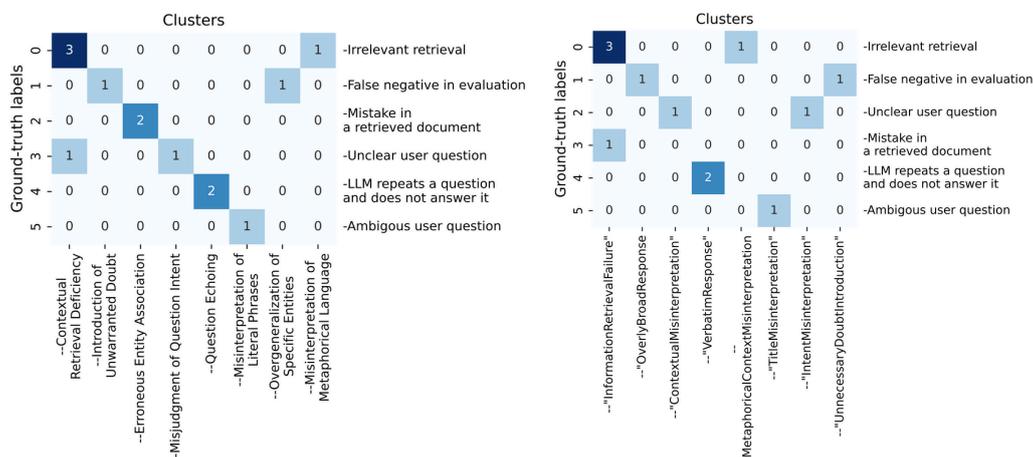
### Examples of per-instance analysis illustration

Human expert	GPT4o (accuracy: 54%)	Eval	Qwen-2.5-7B (accuracy: 8%)	Eval
the retrieved documents do not contain the correct answer	The retrieved documents did not contain the correct information about Australia's 1999 referendum on the British Commonwealth, resulting in an incorrect generated answer.	Yes	The generated answer failed to correctly identify the specific collective entity (the British Commonwealth) from which Australia was considering withdrawal, instead providing a vague description of a collective security system.	No
an LLM repeats a question without answering it	The generated response fails to answer the question as it repeats the question verbatim without providing the required answer 'Sweden'.	Yes	The LLM failed to extract and incorporate the relevant information from the context into its response, specifically missing the identification of the country as 'Sweden'.	No
a mistake in a retrieved document which is propagated to the llm response	The retrieval step supplied incorrect information about Felipe Calderon, erroneously linking him to Nicaragua, causing the LLM to generate an incorrect answer.	Yes	The primary issue is the incorrect identification of the country associated with President Felipe Calderón, likely due to the LLM retrieving and misinterpreting information that linked him to Nicaragua instead of Mexico.	Yes
Unclear user question	The LLM failed by not providing the author's name and misinterpreting the context, leading to incorrect pronoun usage.	No	The LLM failed to correctly attribute the drowning incident to the right character, misquoting "Roberta Alden" instead of "Grace Brown."	No
error of LLM evaluation, because the provided answer is correct	The generated answer unnecessarily introduces doubt about the correct information, causing it to be considered incorrect.	No	The generated response included unnecessary details and did not directly state the name "William Blake", leading to it being marked incorrect by the evaluation metric.	No

### An example of a final report generated by LLM-as-a-qualitative-judge (GPT4o)

- \* **Contextual Retrieval Deficiency:** The generated response fails due to a lack of sufficient context from the retrieved documents, resulting in a misunderstanding of the question's requirements. (4)
- \* **Erroneous Entity Association:** The retrieval step provides incorrect associations between entities, leading to the generation of inaccurate responses due to misinformation. (2)
- \* **Question Echoing:** The generated response echoes or repeats the question without providing a substantive answer, resulting in a failure to address the query with the required information. (2)

### Confusion matrices for issue type clustering for GPT4o (left) and Qwen-2.5-7B (right)



## H.12 Educational QA (SyllabusQA dataset)

### Task illustration

<b>Example user input</b>
ENG 204 syllabus_FA 2021: I won't be in class on 11/16. What topic will I miss?
<b>Example ground truth answer</b>
['You will likely miss the seminar about Franny Choi's "Soft Science" on 11/16. ']
<b>Task details</b>
Metric: LLMEval: it uses an LLM, which is provided with the ground truth label(s) and the generated response, and is asked to output 'yes' or 'no', i.e. whether the generated response is correct or not. We assume that the ground truth labels provided in the dataset are correct and cannot be the source of the pipeline failure. The LLM answer was obtained using retrieval-augmented generation.

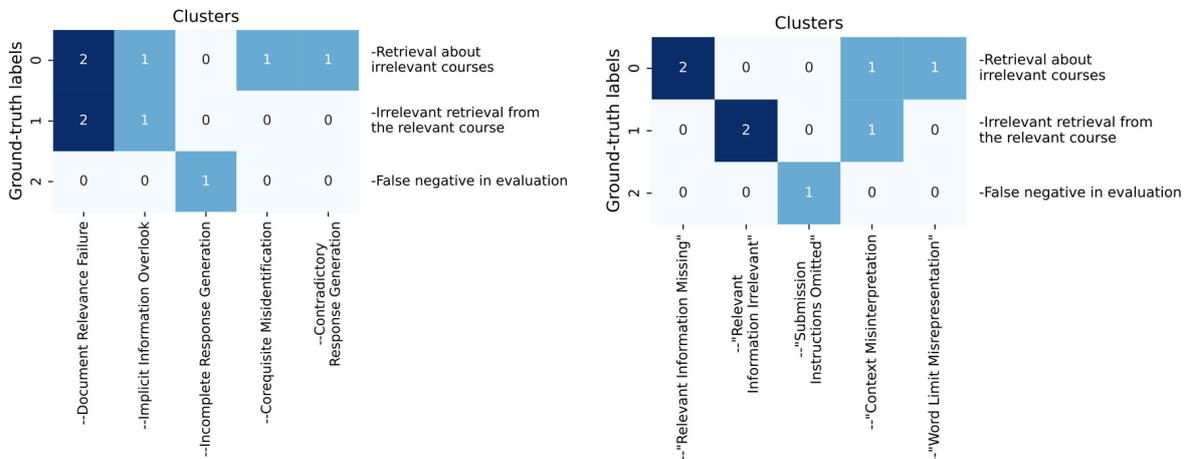
### Examples of per-instance analysis illustration

Human expert	GPT4o (accuracy: 33%)	Eval	Qwen-2.5-7B (accuracy: 22%)	Eval
The retrieved documents do not provide relevant information about the course to answer the given question	The failure was caused by the retrieval module providing irrelevant documents, leading to a response that did not specifically address the question about grade calculation.	Yes	The primary issue is the irrelevance of the retrieved documents and the subsequent generation of an answer that does not match the provided ground truth label.	Yes
The retrieved documents are about wrong courses which leads to a wrong response	The LLM-generated answer incorrectly claims there is no word limit, while the ground truth specifies a word limit of about 100 words.	No	The generated response failed to address the specific word limit mentioned in the ground truth label, instead providing unrelated information about the course syllabus.	No
Error in evaluation: a correct answer is judged as wrong	The generated response lacks a clear statement on submitting online homework via WileyPLUS, resulting in an incomplete answer.	No	The primary issue is the lack of specificity and clarity in the LLM-generated answer, which includes unnecessary information and fails to clearly distinguish between the two types of homework as specified in the ground truth label.	No

### An example of a final report generated by LLM-as-a-qualitative-judge (GPT4o)

- \* **Document Relevance Failure:** The retrieval module supplies documents that are not pertinent to the question, leading to the generation of answers that fail to address the required topic or query context. (4)
- \* **Implicit Information Overlook:** The language model fails to detect or interpret implicitly stated information in the source materials, resulting in an incorrect understanding or omission of essential details in the generated response. (2)

### Confusion matrices for issue type clustering for GPT4o (left) and Qwen-2.5-7B (right)



# Cross-Lingual Stability of LLM Judges Under Controlled Generation: Evidence from Finno-Ugric Languages

Isaac Chung<sup>1</sup>, Linda Freienthal<sup>1</sup>

<sup>1</sup>Zendesk

first.last@zendesk.com

## Abstract

Cross-lingual evaluation of large language models (LLMs) typically conflates two sources of variance: genuine model performance differences and measurement instability. We investigate evaluation reliability by holding generation conditions constant while varying target language. Using synthetic customer-support dialogues generated with identical parameters across Estonian, Finnish, and Hungarian, we test whether automatic metrics and LLM-as-a-judge scoring produce stable model rankings across these morphologically rich, related Finno-Ugric languages. With a small set of Estonian native speaker annotations as a reference point, we find systematic ranking instabilities: surface-level metrics (lexical diversity, surface and semantic similarity) maintain cross-language stability, but pragmatic judgments (coherence, instruction-following) exhibit rank inversions and near-zero correlations. Because generation is controlled, these inconsistencies reflect how judge scoring behaves differently across languages rather than true model differences.

This controlled design provides a diagnostic probe: evaluation methods that fail to maintain stability under identical generation conditions signal transfer failure before deployment. Our findings suggest that zero-shot judge transfer is unreliable for discourse-level assessment in morphologically rich languages, motivating language-specific calibration against targeted human baselines. We release our controlled generation protocol, synthetic data, and evaluation framework to enable replication across language families at <https://github.com/isaac-chung/cross-lingual-stability-judges>.

## 1 Introduction

Evaluating large language models (LLMs) in morphologically rich, underrepresented languages faces a paradox: the places that most need reliable

evaluation have the least human supervision. Recent benchmarks for Finno-Ugric languages like Estonian (Lillepalu and Alumäe, 2025), Finnish (Luukkonen et al., 2023), and Hungarian (Yang et al., 2025b) extend coverage beyond English, yet largely inherit high-resource evaluation practices—emphasizing single-turn tasks and assuming the validity of automatic or model-based scoring whose behavior in conversational settings remains poorly understood.



Figure 1: Example opening messages in each language from the generated dialogues. In English, it reads ‘Good day! You have spoken to Klaus Customer Support, Martin here. How can I help you today?’.

We address this validation trap through controlled diagnostic testing: generating dialogues with identical parameters across Estonian, Finnish, and Hungarian to probe judge behavior. If rankings destabilize when only language varies, the method will fail on natural data.

Recent multilingual judge studies reveal systematic inconsistency across languages (Fleiss’  $\kappa \approx 0.3$  across 25 languages; Fu and Liu 2025), yet the sources of this instability remain poorly understood. Our controlled generation isolates evaluation behavior from content variation to diagnose transfer failures.

Using synthetic customer-support dialogues generated with identical parameters, we first verify generation consistency through surface-level calculated metrics (lexical diversity, surface similarity, semantic similarity), then test whether LLM-as-a-judge pragmatic assessments maintain cross-language ranking stability. This two-stage design isolates judge behavior: if surface properties are comparable but judge rankings diverge, the instability originates in the evaluation process rather than content variation. Our contributions are:

1. We demonstrate that LLM-as-a-judge coherence assessment exhibits systematic rank inversions ( $\tau \approx 0$ ) across morphologically rich languages under controlled generation, while surface metrics maintain stability ( $\tau \geq 0.76$ ).
2. We provide a diagnostic methodology for detecting cross-linguistic ranking instabilities before large-scale deployment, validated through judge ablation and prompt-language sensitivity checks.
3. We release our controlled generation protocol, synthetic dialogues, and evaluation prompts to enable replication studies in other language families.

## 2 Methods

### 2.1 Dialogue Generation

We generate 10K synthetic customer-support dialogues per language using parametrized templates with identical distributions across Estonian, Finnish, Hungarian, and English (40+ industries, 20+ problem types; full specifications in Appendix B). While this setup ensures semantical alignment in the prompts, we recognize that the resulting dialogue quality may vary due to the models’ varying linguistic proficiencies, which may introduce subtle content variance across languages. English serves as a high-resource and typologically distinct anchor. By comparing Finno-Ugric outputs to this baseline, we can observe how model performance shifts when the same scenario is realized in lower-resource linguistic contexts. Dialogues are generated end-to-end in single API calls to enable discourse-level evaluation. Code and dataset is released at <https://github.com/isaac-chung/cross-lingual-stability-judges>.

### 2.2 Human Annotation

Three native Estonian speakers independently annotate 100 dialogues for coherence (conversation-

level consistency) and fluency (grammatical naturalness). Inter-annotator agreement is fair to moderate ( $\kappa = .385$  coherence,  $\kappa = .321$  fluency), reflecting conversational evaluation subjectivity. This moderate agreement bounds expectations for automated cross-linguistic consistency—recent work shows LLM judges achieve even lower cross-language agreement (Fu and Liu, 2025), highlighting the challenge of zero-shot evaluation transfer. These judgments provide a reference for interpreting automatic and judge patterns (Appendix C).

### 2.3 Evaluation Framework

We first verify generation consistency via *surface-level* calculated metrics (TTR, MATTR, self-BLEU, semantic similarity; Appendix A), then test whether LLM-as-a-judge scoring maintains cross-language ranking stability. This two-stage design isolates judge behavior: if surface properties are comparable but judge rankings diverge, instability originates in evaluation transfer. We note, however, that this design also captures the inherent variability of generator performance across languages, allowing us to observe how the entire evaluation pipeline reacts to shifting linguistic contexts.

We use gpt-5-mini with default reasoning effort as an automatic judge to evaluate 100 conversations per model per language. Guided by existing works (Barbu et al., 2025; Bae et al., 2022; Finke et al., 2025), the judge assigns scores for **Grammar (G)**, **Readability (R)**, **Coherence (C)**, and **Fluency (F)**. Additionally, we measure **Label Recovery Accuracy (LRA)**, which assesses instruction-following and semantic consistency by attempting to recover generation parameters from dialogue content. We categorize G, R, and F as *surface-level* judge metrics—evaluating grammatical correctness, lexical choice, and sentence-level naturalness—and C and LRA as *pragmatic* dimensions requiring discourse-level reasoning about conversation flow and instruction alignment. The judge operates zero-shot with English meta-prompts (Appendix D). A sensitivity check using native-language meta-prompts for Estonian showed negligible variance from English-prompt results (difference  $< 0.05$ ; see Section 3.5 for details). An ablation across three judge models in Appendix G suggests that task difficulty stems from ground-truth ambiguity rather than judge capability with minimal scoring variance ( $\Delta < 0.02$ ), supporting our choice of the cost-effective baseline model.

For each metric, we compute per-language model rankings and quantify agreement using Kendall  $\tau$  (95% bootstrap CIs,  $N = 1,500$ ). Rank inversions are tested via permutation. While our generation is controlled at the parameter level, observed instabilities reveal how the evaluation pipeline, comprising both the generator’s output quality and the judge’s scoring logic, becomes fragile when transferred to non-English contexts.

## 2.4 Generator Models

We use gpt-4.1-mini, Llama-3.3-70B-Instruct (Grattafiori et al., 2024), Mixtral-8x7B-Instruct (Jiang et al., 2024), Command-R, Llama-3.1-8B-Instruct (Grattafiori et al., 2024), and Claude Sonnet 4, all accessed via Amazon Bedrock.<sup>1</sup>

## 3 Results

Our results focus on identifying systematic reliability failures in evaluation transfer rather than comparing model performance.

### 3.1 Automatic metrics reveal stable semantic content despite surface variation

Automatic metrics (explained in Appendix A) reveal a nuanced picture in Table 2: while semantic similarity remains stable across languages (mean differences  $< .03$ ), surface-level metrics show systematic language effects. Estonian consistently exhibits higher lexical diversity (MATTR: .48-.80) and lower repetition (Full Self-BLEU: .05-.14) compared to Finnish (MATTR: .45-.70, Self-BLEU: .11-.30) and Hungarian (MATTR: .49-.76, Self-BLEU: .22-.35) across all models. These patterns likely reflect morphological complexity differences rather than generation quality variance.

Beyond language effects, models differ notably in lexical diversity: Llama3.1-8B shows lower MATTR (.45-.49) than Mixtral-8x7B (.70-.80). Despite these surface differences, semantic similarity remains consistent across languages.

Crucially, semantic similarity scores remain remarkably consistent (.89-.94 across all models and languages), confirming that underlying *content quality* is comparable despite surface variation. This dissociation validates our experimental design for judge evaluation: generation produces semantically equivalent dialogues, but surface properties differ systematically by language.

<sup>1</sup><https://aws.amazon.com/bedrock/>

### 3.2 Human annotation provides a noisy reference point

Estonian annotations yield mean scores of  $.842 \pm .367$  (coherence, on binary scale) and  $2.108 \pm .696$  (fluency, on 0-3 scale), with fair-to-moderate agreement ( $\kappa = .385$ ,  $\kappa = .321$ ). Annotators report task-level coherence but reduced linguistic naturalness (Appendix C). This moderate agreement bounds expectations for automated cross-linguistic consistency.

Annotators noted that dialogues were logically coherent but linguistically unnatural. Common feedback included overly formal tone, expressions that feel translated from English, and phrasing resembling ‘B2 level speaker, not a native.’ Frequent coherence issues included inconsistent customer names and illogical scenarios. Examples with annotator feedback are provided in Appendix C.

### 3.3 LLM-as-a-judge scores diverge from human judgments and destabilize across languages

Table 1 shows that LLM-as-a-judge evaluations align imperfectly with human judgments in Estonian, and exhibit significant instability when extended to Finnish and Hungarian. While G and R scores remain relatively stable, scores for C, F, and LRA exhibit substantial variance across languages and models. English (Appendix D) shows ceiling effects ( $C \approx 2.98$ – $3.00$ ), limiting discriminative power but maintaining moderate ranking stability.

While surface metrics remain stable, coherence rankings scramble across language pairs, indicating that discourse-level assessment logic does not transfer reliably across morphologically rich languages. Label recovery accuracy (LRA) results are provided in Appendix D.

### 3.4 Ranking stability reveals coherence breakdown

We quantify evaluation stability in Figure 2. The results reveal a sharp divide between surface-level and pragmatic assessment. Surface-level metrics (G, R, F) exhibit high cross-language stability ( $\tau \geq .70$ ) with minimal rank inversions (1–3 per pair). However, Coherence shows systematic breakdown: near-zero or negative correlations across Finno-Ugric language pairs ( $\tau = -.06$  for et–hu,  $\tau = -.17$  for fi–hu), with significant inversions ( $p = .02$ ) for et–hu. English Coherence scores show ceiling effects (mean  $\approx 2.98$ – $3.00$ ),

Model	Grammar (G)			Readability (R)			Coherence (C)			Fluency (F)			LRA		
	et	fi	hu	et	fi	hu									
gpt-4.1-mini	<b>3.17</b> ±.55	<b>3.51</b> ±.52	<b>3.57</b> ±.50	3.63±.48	<b>3.86</b> ±.34	<b>3.85</b> ±.36	<b>2.99</b> ±.09	2.97±.16	2.99±.10	<b>2.35</b> ±.48	<b>2.56</b> ±.50	<b>2.66</b> ±.48	<b>.62</b> ±.09	.34±.25	.36±.26
Llama3.3-70B-Inst.	2.39±.52	3.03±.58	3.04±.57	2.92±.43	3.44±.51	3.43±.49	2.93±.25	2.89±.34	2.94±.23	1.92±.32	2.22±.47	2.18±.41	.36±.20	<b>.62</b> ±.13	.59±.11
Mixtral-8x7B-Inst.	1.63±.48	2.32±.60	2.36±.64	1.99±.36	2.77±.61	2.70±.60	2.72±.45	2.81±.39	2.79±.43	1.17±.37	1.82±.49	1.78±.51	.33±.24	.34±.25	.33±.22
Command-R	1.50±.61	1.61±.54	1.65±.51	1.81±.56	2.01±.45	2.04±.43	2.56±.56	2.64±.48	2.57±.51	1.04±.31	1.13±.34	1.16±.36	.40±.20	.38±.22	.34±.22
Llama3.1-8B-Inst.	1.61±.49	2.22±.44	2.23±.52	1.87±.36	2.66±.48	2.65±.50	2.34±.51	2.62±.49	2.63±.48	1.06±.24	1.81±.40	1.75±.45	.30±.21	.30±.23	.42±.20
claude-sonnet-4	3.04±.50	3.18±.50	3.19±.47	<b>3.63</b> ±.48	3.81±.40	3.78±.41	2.98±.13	<b>2.99</b> ±.09	<b>2.99</b> ±.09	2.29±.45	2.45±.51	2.47±.50	.45±.17	.36±.23	<b>.75</b> ±.14

Table 1: LLM-as-a-judge evaluation of generated Estonian (et), Finnish (fi), and Hungarian (hu) dialogues. The best scores per metric and language are **bolded**.

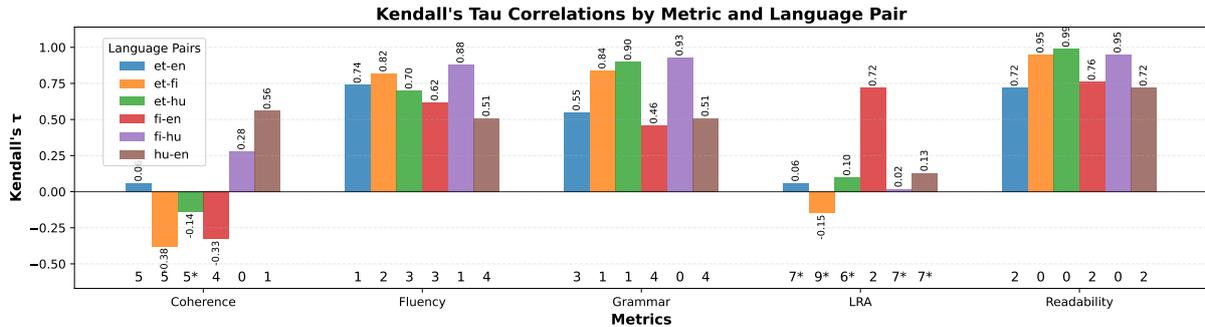


Figure 2: Cross-language ranking stability measured by Kendall’s  $\tau$ . Error bars show 95% bootstrap confidence intervals. Numbers below bars indicate rank inversions (out of 15 possible pairwise inversions among 6 models); asterisks denote statistical significance via permutation test (\*  $p < 0.05$ ). Surface-level metrics (Grammar, Readability, Fluency) maintain high stability ( $\tau \geq 0.62$ ) with minimal inversions. Pragmatic dimensions show systematic breakdown: Coherence exhibits near-zero or negative correlations, and LRA shows significant rank scrambling across all Finno-Ugric pairs (9\*, 6\*, 7\* inversions). English pairs included for context, though ceiling effects limit their informativeness for Coherence.

preventing meaningful ranking comparisons with English. Our analysis therefore focuses on Finno-Ugric pairs, where score variance allows for meaningful ranking comparisons.

Since generation parameters are held constant and automatic metrics confirm comparable generation quality, these Coherence rank inversions point to judge transfer failure at the discourse level. The judge’s internal discourse-level assessment logic collapses when transferred across morphologically rich languages, even among closely related language pairs. As sensitivity checks confirm that scores are robust to meta-prompt language (subsection 3.5), this instability represents a fundamental breakdown in cross-linguistic evaluation reliability rather than a prompt engineering problem. These findings indicate that discourse coherence assessment—unlike surface-level grammatical or lexical evaluation—cannot be zero-shot transferred and requires language-specific calibration before deployment. Full stability analysis is provided in Appendix E.

### 3.5 Meta-prompt language sensitivity

To ensure that the use of English-centric meta-prompts did not introduce instruction-language bias into our results, we conducted a sensitivity study on the Estonian calibration set ( $N = 100$ ). We re-evaluated the dialogues from all six generator models using a version of the LLM-as-a-judge system prompt translated into Estonian by a native speaker.

Scores produced by the native-language prompt are nearly identical to those produced by the English meta-prompt. Results suggest that the judge’s evaluation behavior is driven by its internal representation of the target language rather than the language of the instructions. Detailed results can be found in Appendix F. This rules out prompt language as the source of instability. The underlying cause remains as discussed in Section 3.4.

### 3.6 Ablation: Judge Model

To test whether instability is specific to our chosen judge, we compared six judge models (GPT-5-mini, GPT-5.1, GPT-5.1-high, Qwen3-32B (Yang et al., 2025a), Llama-4-Maverick (Meta AI, 2025), GPT-OSS-120B (OpenAI et al., 2025)) on Finnish

Model	TTR			MATTR			Full Self-BLEU			Agent Self-BLEU			Client Self-BLEU			Intra Model Sim		
	et	fi	hu	et	fi	hu	et	fi	hu	et	fi	hu	et	fi	hu	et	fi	hu
gpt-4.1-mini	.80±.07	.63±.07	.64±.07	.81±.06	.68±.06	.71±.04	.10	.19	.22	.15	.19	.20	.12	.11	.15	.93±.01	.92±.02	.93±.02
Llama3.3-70B-Inst.	.67±.11	.52±.11	.55±.11	.67±.11	.55±.09	.59±.08	.14	.30	.35	.26	.34	.36	.25	.18	.25	.94±.01	.93±.02	.93±.02
Mixtral-8x7B-Inst.	.80±.10	.70±.08	.76±.08	.80±.10	.70±.08	.76±.08	.07	.21	.23	.15	.22	.20	.15	.15	.16	.91±.02	.91±.02	.90±.02
Command-R	.74±.10	.64±.09	.60±.09	.75±.09	.67±.07	.72±.06	.07	.11	.31	.11	.10	.33	.09	.07	.25	.92±.02	.91±.02	.89±.03
Llama3.1-8B-Inst.	.48±.18	.42±.12	.45±.12	.48±.19	.45±.09	.49±.09	.05	.24	.29	.43	.24	.28	.43	.12	.18	.93±.01	.93±.01	.93±.01
claude-sonnet-4	.73±.09	.59±.10	.61±.09	.78±.06	.68±.05	.72±.05	.12	.19	.26	.19	.19	.25	.12	.13	.19	.92±.02	.92±.02	.92±.02

Table 2: Automatic metrics for generated Estonian (et), Finnish (fi), and Hungarian (hu) dialogues. TTR, MATTR and Intra Model Similarity show their standard deviation as well.

dialogues. All judges exhibit near-identical performance patterns with minimal variance ( $\Delta < 0.02$  across categories). This suggests the instability is systematic rather than judge-specific. Full details in [Appendix G](#).

## 4 Discussion and Outlook

**Surface-level evaluation transfers; discourse assessment does not.** Practitioners can deploy judge-based surface assessments (grammar, readability, fluency) for cross-linguistic comparison with confidence ( $\tau \geq 0.70$  across Finno-Ugric pairs). Discourse coherence exhibits systematic breakdown ( $\tau \approx 0$ ) even among related languages, requiring language-specific calibration.

**Controlled stability as a validity gate.** Our diagnostic approach provides a negative check: if an LLM judge produces inconsistent model rankings across languages under identical generation conditions, they will fare worse on natural data. This motivates a staged workflow: (1) verify generation consistency with automatic metrics, (2) collect a small expert sample ( $N \sim 100$ ) in the target language, (3) test judge-human ranking alignment, (4) calibrate if correlations are weak. This prioritizes measurement reliability while respecting resource constraints in underrepresented language communities.

### Limitations

Synthetic dialogues enable controlled evaluation but may exhibit stylistic homogeneity and phrasing not present in real data. Validation on natural customer support scenarios is needed to confirm ranking instabilities persist in operational settings. Surface-level ranking stability suggests comparable generation quality across languages, making judge transfer failure the more likely explanation for Coherence instability. However, we cannot completely rule out discourse-level quality differences that surface metrics do not capture.

Human calibration is restricted to Estonian ( $N =$

100). Our controlled generation does not require multilingual human labels to detect ranking problems: if model rankings change when only language varies, the judge is unreliable. The Estonian annotations serve only to confirm that synthetic dialogues vary semantically and evaluate the fluency of a subset of the synthetic dialogues.

We examine customer support dialogues in three related Finno-Ugric languages. While judge ablation ([Appendix G](#)) confirms scoring stability across GPT-5 variants, our findings may not hold for non-commercial models, other conversational domains, or linguistically distant languages. We focus on discourse coherence; other aspects like politeness conventions and language-specific grammatical patterns remain unexplored.

### Acknowledgments

We thank Mervi Sepp Rei, Reimo Priidik, Martin Küngas, and Andreas Pung for labeling, Daniel Loureiro for his valuable feedback on the draft, and Joonathan Mägi, Mikk Müraus, and Kajetan Bocharczuk for their foundational work on the conversation generator. We thank Magda Kubit, Abdallah Akzouk, and Abhinay Kathuria for their support in open-source model inference. We thank the reviewers for their insightful feedback.

### References

- Gati Aher, Rosa I. Arriaga, and Adam Tauman Kalai. 2023. [Using large language models to simulate multiple humans and replicate human subject studies](#). *Preprint*, arXiv:2208.10264.
- Sanghwan Bae, Donghyun Kwak, Sungdong Kim, Donghoon Ham, Soyoung Kang, Sang-Woo Lee, and Woomyoung Park. 2022. [Building a role specified open-domain dialogue system leveraging large-scale language models](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2128–2150, Seattle, United States. Association for Computational Linguistics.

- Eduard Barbu, Meeri-Ly Muru, and Sten Marcus Malva. 2025. [Improving estonian text simplification through pretrained language models and custom datasets](#). *Preprint*, arXiv:2501.15624.
- Maximillian Chen, Alexandros Papangelis, Chenyang Tao, Seokhwan Kim, Andy Rosenbaum, Yang Liu, Zhou Yu, and Dilek Hakkani-Tur. 2023. [PLACES: Prompting language models for social conversation synthesis](#). In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 844–868, Dubrovnik, Croatia. Association for Computational Linguistics.
- Maximillian Chen, Alexandros Papangelis, Chenyang Tao, Andy Rosenbaum, Seokhwan Kim, Yang Liu, Zhou Yu, and Dilek Hakkani-Tur. 2022. [Weakly supervised data augmentation through prompting for dialogue understanding](#). In *NeurIPS 2022 Workshop on Synthetic Data for Empowering ML Research*.
- Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzemiński, Genta Indra Winata, Saba Sturua, Saiteja Utpala, Mathieu Ciancone, Marion Schaeffer, Diganta Misra, Shreeya Dhakal, Jonathan Rystrom, Roman Solomatin, Ömer Veysel Çağatan, and 63 others. 2025. [MMTEB: Massive multilingual text embedding benchmark](#). In *The Thirteenth International Conference on Learning Representations*.
- Lennart Finke, Thomas Doods, Mat Allen, Juan Diego Rodriguez, Noa Nabeshima, and Dan Braun. 2025. [\[tiny\] parameterized synthetic text generation with simplestories](#). In *Will Synthetic Data Finally Solve the Data Access Problem?*
- Xiyan Fu and Wei Liu. 2025. [How reliable is multilingual llm-as-a-judge?](#) *Preprint*, arXiv:2505.12201.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, L elio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, and 7 others. 2024. [Mixtral of experts](#). *Preprint*, arXiv:2401.04088.
- Kimmo Kettunen. 2014. [Can type-token ratio be used to show morphological complexity of languages?](#) *Journal of Quantitative Linguistics*, 21:223–245.
- Sven Laur, Siim Orasmaa, Dage S arg, and Paul Tammo. 2020. [Estnltk 1.6: Remastered estonian nlp pipeline](#). In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 7154–7162, Marseille, France. European Language Resources Association.
- Helena Grete Lillepalu and Tanel Alum ae. 2025. [Estonian native large language model benchmark](#). *Preprint*, arXiv:2510.21193.
- Risto Luukkonen, Ville Komulainen, Jouni Luoma, Anni Eskelinen, Jenna Kanerva, Hanna-Mari Kupari, Filip Ginter, Veronika Laippala, Niklas Muenhoff, Aleksandra Piktus, Thomas Wang, Noumane Tazi, Teven Le Scao, Thomas Wolf, Osma Suominen, Samuli Sairanen, Mikko Merioksa, Jyrki Heinonen, Aija Vahtola, and 2 others. 2023. [Fingpt: Large generative models for a small language](#). *Preprint*, arXiv:2311.05640.
- Meta AI. 2025. [Llama 4: Multimodal intelligence](#). Meta AI Blog. Accessed: 2025-01-XX.
- OpenAI, :, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, and 108 others. 2025. [gpt-oss-120b gpt-oss-20b model card](#). *Preprint*, arXiv:2508.10925.
- Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. [Stanza: A Python natural language processing toolkit for many human languages](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.
- Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. [Multilingual e5 text embeddings: A technical report](#). *arXiv preprint arXiv:2402.05672*.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025a. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.
- Haote Yang, Xingjian Wei, Jiang Wu, No emi Ligeti-Nagy, Jiaying Sun, Yinfan Wang, Zijian Gy oz o Yang, Junyuan Gao, Jingchao Wang, Bowen Jiang, Shasha Wang, Nanjun Yu, Zihao Zhang, Shixin Hong, Hongwei Liu, Wei Li, Songyang Zhang, Dahua Lin, Lijun Wu, and 2 others. 2025b. [Openhual: Evaluating large language model on hungarian specifics](#). *Preprint*, arXiv:2503.21500.
- Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. [Taxygen: A benchmarking platform for text generation models](#). In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR ’18*, page 1097–1100, New York, NY, USA. Association for Computing Machinery.

## A Automatic Metrics

Here is the description of the automatic metrics.

- **TTR and MATTR**: we compute both simple Type-Token Ratio (TTR) (unique words / total words) and Moving Average TTR (MATTR) (Kettunen, 2014) over sliding 100-token windows for length-independent measurement of morphological variety. Higher MATTR and TTR values indicate greater lexical diversity.
- **Self-BLEU** (Zhu et al., 2018): Calculated at three granularity levels: full conversations, agent responses only, and client responses only—to detect formulaic patterns. We used 4-gram BLEU and NLTK’s smoothing function (method4). Lower values indicate reduced repetition and greater diversity.
- **Intra Model Conversation Similarity** answers the question "How different are the conversations from each other?". For that we use the cosine similarity between sentence embeddings from multilingual-e5-large-instruct (Wang et al., 2024), the highest-ranked multilingual model in MMTEB (Enevoldsen et al., 2025). Lower scores indicate higher similarity (more template-like), while higher scores indicate greater conversation diversity within a model.

For calculating the values, all three languages use morphological lemmatization: EstNLTK (Laur et al., 2020) for Estonian, and Stanza (Qi et al., 2020) for Hungarian and Finnish, with language-specific stopword filtering.

See in Table 2 the results of the metrics per model and language.

Beyond language effects, we observe notable model differences. Llama3.1-8B shows substantially lower lexical diversity (TTR: .42-.48, MATTR: .45-.49) compared to Mixtral-8x7B-Inst. (TTR: .70-.80, MATTR: .70-.80), suggesting different training data characteristics or architectural effects on generation diversity. Command-R achieves the lowest agent-side self-BLEU scores (.10-.11 for et/fi), indicating reduced formulaic patterns in agent responses. However, all models maintain consistent semantic similarity scores across languages, confirming that surface-level differences do not translate to semantic quality variance.

## B Dialogue Generation

We generate synthetic customer-support dialogues using parametrized prompt templates to create controlled test conditions across languages. Parameters control industry (40+ categories), customer problem type (20+), communication channel, agent experience, agent type, and conversation length. Crucially, **we use identical parameter distributions and generation models across all three languages**, enabling us to isolate evaluation behavior from content variation. If dialogues are generated under identical conditions but judges produce different model rankings across languages, the instability originates in the evaluation process rather than genuine performance differences.

Dialogues are generated end-to-end in a single API call, enabling evaluation of global discourse coherence rather than turn-level response quality. We generate 10K conversations per language for Estonian, Finnish, Hungarian, and English (40K total), providing sufficient scale to probe evaluation stability while remaining tractable for analysis.

Table 3 and Table 4 together form the dialogue generation prompt, with values of changing parameters in curly brackets. The parameters are sampled from the fixed sets detailed in Table 5.

Each multi-turn conversation is generated individually in one go, similar to the method used in PLACES (Chen et al., 2023) but without in-context learning as we operate in a data-scarce setting for underrepresented languages. Utterance-level generation strategies (Chen et al., 2022; Aher et al., 2023) are not suitable for this study as we seek to evaluate full conversation generation capabilities of LLMs.

**\*\*Role\*\*** You are an expert generator of customer support conversations. The generated conversations must stay on topic as much as possible, and mimic real life customer support interactions as much as possible. The most important thing is for these conversations to be as realistic as possible.

**\*\*Instructions\*\*** These conversations are between professional agents and human customers. Customers have emotions, needs, and expectations. There are specific instructions for each conversation that you must follow. These are for agents to follow when interacting with customers. There are also instructions for the customer to follow. Do you **UTMOST BEST** to adhere to the following instructions for generation. If there are more than 1 agent in the conversation, the agents turns must be sequential and must **NOT** interleave. For example, if agent1 and agent2 are in the conversation, the **ALLOWED** turns can be: a) agent1, customer, agent1, customer, agent2, customer, agent2; b) agent2, customer, agent2, customer, agent1, customer, agent1; and the **BANNED** turns are: c) agent1, customer, agent2, customer, agent1, customer, agent2;

Table 3: System prompt for dialogue generation.

(User Prompt) Generate a chat conversation between a customer and and {n\_agents} support agents. The emails of the agents are: {agentemails}. The conversation must be in {language} and should be made of [{n\_messages}] messages.

'Klaus' is a company in the {industry} industry. The conversation must be tailored to the industry. For example, use products and services that are common in the industry, and use language that is common in the industry. The conversation must reference at least one issue with a service, product, or policy that is relevant to the company.

The AGENT must greet the customer. For example, using common greeting words like 'Hello' or 'Good day' in the respective language and address the customer by name, and based on the channel. The AGENT must use proper grammar and spelling, and must follow grammatical rules in the respective language. The AGENT must demonstrate empathy towards the customer and must tailor the conversation to address their problems and needs. The AGENT must use professional tone. {agent\_type} {problem} {channel} {agent\_experience}

Table 4: User prompt for dialogue generation.

## C Human Labeling

### C.1 Instructions

Table 6 shows detailed labeling instructions given to human labelers to evaluate the generated Estonian dialogues<sup>1</sup>. Agreement levels follow standard guidelines:  $\kappa > 0.8$  (excellent),  $0.6 < \kappa \leq 0.8$  (substantial),  $0.4 < \kappa \leq 0.6$  (moderate),  $0.2 < \kappa \leq 0.4$  (fair), and  $\kappa \leq 0.2$  (poor). These expert judgments provide the calibration signal necessary to validate evaluation dimensions in morphologically rich contexts.

### C.2 Feedback and Examples

Table 7 shows one agent-customer exchange from two examples taken from Estonian dialogues that have labeler feedback. The labelers mentioned that the text contained expressions that could be used in the language but do not feel natural (e.g. gives a feeling of B2 level speaker, not a native). Many phrases felt rough or one could detect the

English phrase it was translated from. This is also reflected by the fluency score, with the average of the reference label (agreement between three annotators) being  $2.108 \pm .696$  on the scale of 0-3.

Regarding logical coherence, the scores are higher: The average of reference labels in conversations is  $.842 \pm .367$  on a binary scale. Reoccurring reasons for the negative logical coherence grade were:

- Inconsistent customer names or amounts of product during the conversation.
- The described issue is illogical. E.g., a customer bought a bicycle and now wants to know how to pay or a customer needs to return an object it has not received yet.
- Hallucinated words that make the entire conversation not understandable.

**Industry:** manufacturing, energy production, energy management, energy technology, apparel retail, retail clothing stores, apparel manufacturing, fitness apparel retail, footwear retail, safety apparel manufacturing, home decor retail, home textiles retail, manufacturing tools, retail technology solutions, gaming technology services, transportation technology, transportation services, logistics and transportation, kitchen appliances manufacturing, utility management services, audio equipment manufacturing, e-commerce grocery retail, gambling and betting, e-commerce retail baby products, furniture retail, label manufacturing, cutlery manufacturing, bicycle manufacturing, telecommunications retail, pet retail, financial services, financial software development, gaming, retail, outdoor equipment retail, e-commerce jewelry manufacturing, retail fashion accessories, automotive parts retail, fintech services, games, e-commerce retail goods, automotive retail, coatings manufacturing, sporting goods manufacturing, e-commerce, beverage retailing, computer hardware manufacturing, automotive manufacturing, e-commerce electronics retail.

**Problem:** create account, delete account, edit account, switch account, check cancellation fee, delivery options, complaint, review, check invoice, get invoice, newsletter subscription, cancel order, change order, place order, check payment methods, payment issue, check refund policy, track refund, change shipping address, set up shipping address.

**Channel:** email, chat.

**Agent Experience:** junior, senior.

**Language:** Estonian, Finnish, Hungarian.

**Agent Type:** human, bot.

**Number of messages:** 4, 8, 12, 16.

Table 5: Parameter options for synthetic dialogue generation. All options are sampled with equal probability, except for message length, which is weighted to favor shorter interactions [0.4, 0.3, 0.2, 0.1].

**Does the content make sense?**  
**YES** → Questions and answers are logical, relevant to the topic.  
**NO** → Questions and answers do not interact logically OR the issue/solution would never occur in any industry OR the agent never sends an email starting with “welcome to chat, how can I help you?”

---

**Is this fluent, human-written Estonian?**  
**3** → Messages could pass as written by fluent speakers.  
**2** → Majority of messages pass as written by fluent speakers, but 1–2 odd wordings and/or 1–2 grammar mistakes (e.g., wrong verb case, pronoun confusion).  
**1** → Several odd wordings and grammar mistakes; still resembles Estonian.  
**0** → Reading this gave me an aneurysm.

Table 6: Detailed labeling instructions are given to human labelers for each question.

## D LLM As A Judge

### D.1 Instructions

Table 9 shows the full system prompt used for the LLM-as-a-judge to evaluate the linguistic and pragmatic dimensions of the generated dialogues. This zero-shot approach uses English meta-prompts to assess performance in morphologically rich languages. As discussed in the main text, the **Label Recovery Accuracy (LRA)** dimension is further utilized as a diagnostic for instruction-following and semantic consistency by attempting to extract generation parameters from the dialogue content. The prompt used to assess LRA is shown in Table 10.

### D.2 English Results

Table 8 shows LLM-as-a-judge results on the English dialogues.

### D.3 LRA Full Results

Label recovery accuracy measures the judge’s ability to extract generation parameters from dialogue content. Figure 3 shows performance across Estonian, Finnish, Hungarian, and English for all parameter categories.

**Example 1:**

AGENT: Tere päevast! Harald siin Klaus spordivarustuse tugitiimist. Kuidas saan teid täna aidata?

CUSTOMER: Tere! Tellisin hiljuti spordijalatsid, aga kahjuks pidin tellimuse tühistama. Nüüd näen, et mulle on lisatud tühistamistasu. Kas see on õigustatud?

Fluency: 1/3

Coherence: 1/1

Feedback: Too formal. "Harold siin" is too literally translated. We usually don't say that. We say "Mina olen Harold" most likely in this context.

**Example 2:**

AGENT: Tere päevast, hea klient! Tänan, et võtsite ühendust Klaus klienditoega. Kuidas saan Teid täna aidata?

CUSTOMER: Tere! Ma tellisin teie poest uue mobiiltelefoni, kuid märkasin, et tarneaadress on valesti sisestatud. Kas saaksin selle muuta enne, kui tellimus välja saadetakse?

Fluency: 1/3

Coherence: 1/1

Feedback: Too formal, usually these conversations are more casual. "kuid" and "ning" are usually not used in speech, only in some literature.

Table 7: One agent-customer exchange from two example generated Estonian dialogues that have labeler feedback. Most labeler feedback flags uncommon expressions and overly formal tone, which led to lower fluency scores in those dialogues.

Model	Grammar (G)	Readability (R)	Coherence (C)	Fluency (F)	LRA
claude-sonnet-4	<b>3.99</b> ±.09	<b>4.00</b> ±.00	2.98±.13	<b>3.00</b> ±.00	<b>.77</b> ±.16
llama3-70b-instruct	3.94±.23	3.99±.09	2.98±.13	2.99±.09	.64±.10
mixtral-8x7b-instruct	3.97±.18	3.97±.18	2.90±.29	2.97±.16	.39±.22
llama3-8b-instruct	3.96±.20	3.98±.13	2.91±.29	2.95±.22	.40±.18
command-r	3.91±.29	3.96±.19	2.95±.23	2.94±.25	.34±.22
gpt-4.1-mini	3.96±.21	<b>4.00</b> ±.00	<b>2.99</b> ±.09	2.99±.09	.37±.22

Table 8: LLM-as-a-judge evaluation of generated English dialogues. The best scores per metric are **bolded**.

## E Cross-language ranking stability

We aggregate per-language per-model means and compute rank correlations for each language pair (et-en, et-fi, et-hu, fi-en, fi-hu, hu-en). To assess whether observed order flips exceed chance, we run a permutation test (randomly reassigning language labels at the per-model level) and report 95% bootstrap confidence intervals.

For inversion count  $n = 6$  models, the maximum number of possible pairwise inversions is  $n(n - 1)/2 = 15$ . An inversion count of 0 represents perfect preservation of model ranking between two languages, while 15 represents a perfect reversal.

Across languages, surface-oriented dimensions (Grammar, Readability, Fluency) show high rank stability ( $\tau$  typically  $\geq 0.5$  with non-significant inversion counts). In contrast, pragmatic dimensions are fragile under transfer: Coherence shows

attenuated or negative agreement for pairs involving Estonian (et-en/fi/hu) with marginal/significant inversion counts, while remaining stable for fi-hu.

LRA exhibits significant inversions for several pairs, including et-en (7,  $p = 0.02$ ), et-fi (9,  $p = 0.01$ ), et-hu (6,  $p = 0.03$ ), fi-hu (7,  $p = 0.02$ ), and hu-en (7,  $p = 0.02$ ). Coherence shows marginal/significant inversions in et-en (5,  $p = 0.05$ ), et-fi (5,  $p = 0.05$ ), and et-hu (5,  $p = 0.04$ ). Because the domain and generator are held constant, instability reflects evaluation transfer rather than model content.

## F Meta-Prompt Sensitivity

Table 12 shows that scores produced by the native-language prompt are nearly identical to those produced by the English meta-prompt. For example, the maximum variance observed for any model in

any dimension remains below 0.05.

## G Appendix: Judge Model Ablation Study

We compare six judge models for Finnish conversation label recovery: GPT-5-mini (baseline), GPT-5.1 with default reasoning, GPT-5.1 with high reasoning effort, Qwen3-32B, Llama-4-Maverick, and GPT-oss-120B. Open-source LLMs are accessed via Groq<sup>2</sup>. Each judge evaluated the same six models over the sampled Finnish dialogues over the LRA categories. The same judge prompt is used from Appendix D across all judges.

Figure 4 shows accuracy by category. All six judges exhibit near-identical performance patterns, with minimal differences ( $\Delta < 0.02$  across categories). Channel classification proves easiest ( $\approx 55\text{--}57\%$ ), followed by Agent Type ( $\approx 57\%$ ), Agent Experience ( $\approx 48\text{--}51\%$ ), Problem ( $\approx 19\text{--}22\%$ ), and Industry ( $\approx 9\text{--}11\%$ ).

Inter-judge agreement was assessed using Spearman correlations across all model-category pairs. Mean correlation was 0.66 across all judge comparisons, indicating moderate-to-substantial agreement while preserving meaningful judgment variance.

Three findings emerge: (1) **Model choice has minimal impact**—GPT-5.1-high performs identically to default reasoning settings, and open-source alternatives (Qwen3-32B, Llama-4-Maverick, GPT-oss-120B) achieve comparable results to proprietary models, suggesting this structured classification task does not benefit from extended reasoning or increased model scale; (2) **Trends generalize across judges**—the performance patterns observed in our main experiments with GPT-5-mini are consistently reproduced by all five alternative judges, including open-source models; (3) **Task difficulty hierarchy is judge-invariant**—all judges struggle identically with Industry/Problem categories while succeeding on Channel/Agent classifications, suggesting difficulty stems from ground-truth ambiguity rather than judge capability.

These results validate our use of GPT-5-mini as the judge throughout our main experiments, demonstrating comparable reliability to both proprietary reasoning models and open-source alternatives.

---

<sup>2</sup><https://groq.com/>

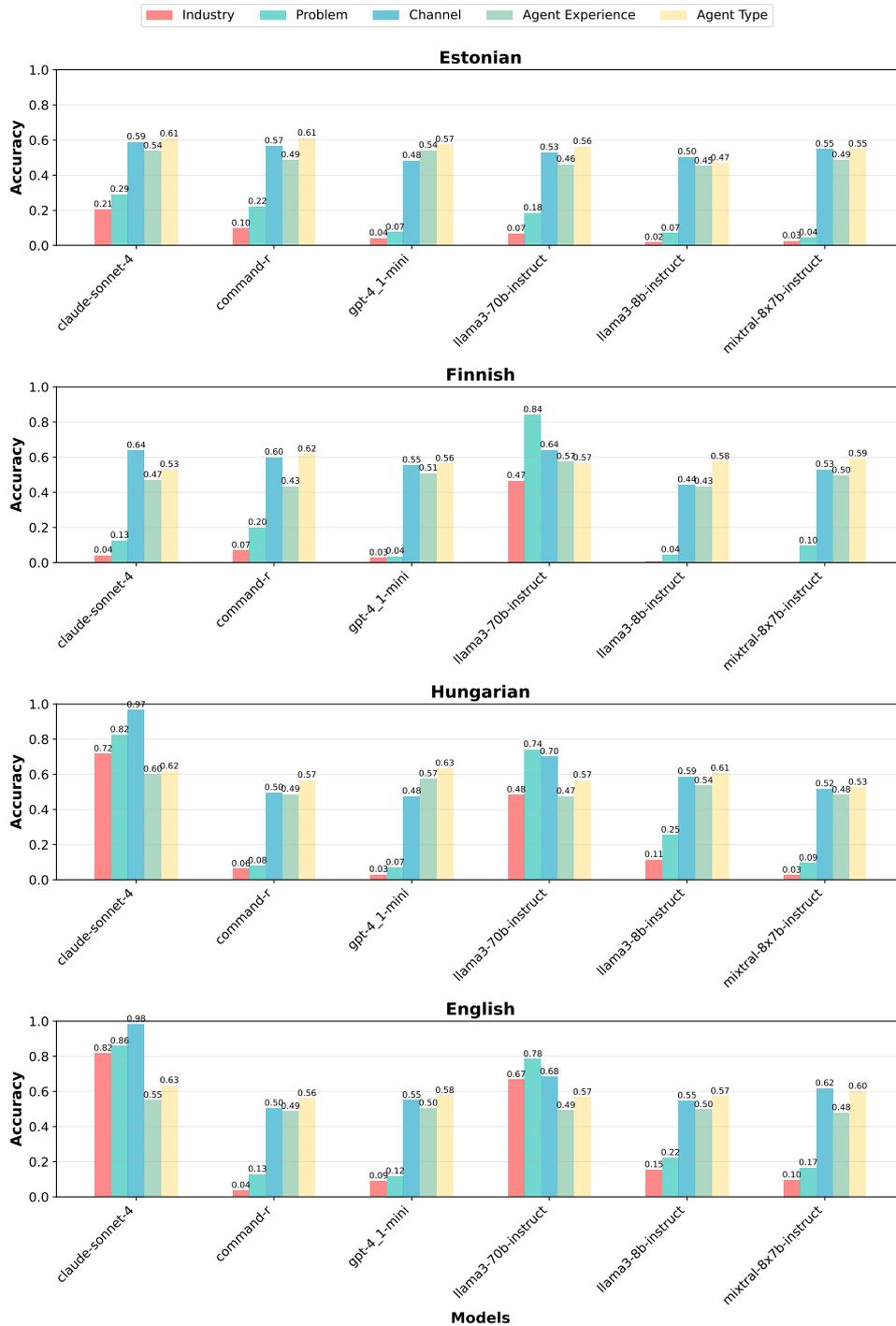


Figure 3: Label recovery accuracy (LRA) across categories by model for sampled dialogues in Estonian, Finnish, Hungarian, and English. Performance varies substantially by category complexity: simple binary parameters (Agent Experience, Agent Type) show consistent accuracy across all languages, while complex semantic categories (Industry: 40+ options, Problem: 20+ types) exhibit poor and inconsistent performance in all languages including English. This pattern suggests that complex parameter recovery may exceed current model capabilities regardless of target language, limiting LRA’s utility as a cross-linguistic diagnostic. Unlike surface metrics and coherence assessment, where clear stability differences emerge, LRA instability appears task-dependent rather than language-dependent.

### LLM Judge Prompt Template

You are an expert linguistic judge specialising in customer-support dialogues across all languages, including Finno-Ugric languages. Your task is to rate the given dialogue according to the criteria below. Provide only the requested output format.

#### Grading Criteria

##### 1. Grammaticality (G): Score 0–4

This criterion evaluates the grammatical correctness of the text, checking how free it is from grammatical errors.

- 0: Numerous grammatical mistakes; largely unreadable.
- 1: Significant errors that make parts difficult to understand.
- 2: Some errors present but overall understandable.
- 3: Minor mistakes that do not affect comprehension.
- 4: Grammatically perfect with no mistakes.

##### 2. Readability (R): Score 0–4

This criterion assesses ease of reading and natural flow, considering sentence length, word complexity, and overall coherence.

- 0: Completely incoherent and unreadable.
- 1: Very difficult to read and understand.
- 2: Readable but requires significant effort.
- 3: Mostly coherent, minor effort required.
- 4: Very easy to read, natural flow.

##### 3. Content Coherence (C): Score 0–3

3: Questions and answers are completely logical, relevant, and form a coherent dialogue flow; realistic business scenario

2: Questions and answers are mostly logical and relevant with minor coherence issues; plausible business scenario

1: Some logical connection between questions and answers but with notable coherence problems; somewhat realistic scenario

0: Questions and answers do not interact logically OR the issue/solution would never occur in any industry OR conversations lack proper structure

##### 4. Fluency (F): Score 0–3

How fluent and natural is this [Estonian/Finnish/English]?

- 3: Messages could pass as written by fluent native speakers.
- 2: Majority of messages pass as fluent, but 1–2 odd wordings and/or 1–2 grammar mistakes.
- 1: Several odd wordings and grammar mistakes; still resembles the language.
- 0: Extremely poor quality with pervasive errors.

Table 9: LLM-as-a-judge system prompt for evaluating grammar, readability, coherence, and fluency of the synthetic customer support dialogues.

### LLM Judge LRA Prompt Template

You are an expert analyst specializing in customer support conversation classification across all languages, including Finno-Ugric languages. Your task is to classify the given conversation according to the categories below.

#### CLASSIFICATION CATEGORIES

1. Industry: {all possible parameters}
2. Problem: Identify the primary issue or inquiry type: {all possible parameters}
3. Channel: Determine the communication method used
  - email: Email-based correspondence
  - chat: Live chat, instant messaging
4. Agent Experience: Assess the agent's expertise level based on responses
  - junior: Basic responses, may need escalation, limited problem-solving
  - senior: Expert responses, complex problem-solving, proactive suggestions
5. Agent Type: Determine if responses are from human or AI
  - human: Natural conversational style, empathy, contextual understanding
  - bot: Structured responses, consistent formatting, may lack nuance

Analyze the conversation carefully and provide your classification for each category along with a brief explanation.

Please classify the following customer support conversation across all required categories:

{conversation}

Provide classifications for:

1. Industry: Select from the specific industries listed in the system prompt (e.g., manufacturing, energy production, etc.)
2. Problem type: Select from the specific problem types
3. Channel: email or chat
4. Agent experience level: junior or senior
5. Agent type: human or bot

Include a brief explanation for your classification decisions.

Table 10: LLM-as-a-judge prompt for evaluating LRA of the synthetic customer support dialogues.

Metric	Pair	Kendall $\tau$ [95% CI]	Spearman $\rho$ [95% CI]	Inversions (obs, $p$ )
Coherence	et-en	0.06 [-0.30, 0.45]	0.07 [-0.39, 0.52]	5, 0.05
Coherence	et-fi	-0.38 [-0.65, 0.65]	-0.41 [-0.70, 0.70]	5, 0.05
Coherence	et-hu	-0.14 [-0.67, 0.29]	-0.16 [-0.72, 0.36]	5, 0.04
Coherence	fi-en	-0.33 [-0.58, -0.12]	-0.38 [-0.65, -0.13]	4, 0.10
Coherence	fi-hu	0.28 [0.00, 0.30]	0.29 [0.00, 0.31]	0, 1.00
Coherence	hu-en	0.56 [0.26, 0.60]	0.64 [0.37, 0.78]	1, 0.66
Fluency	et-en	0.74 [0.47, 1.00]	0.86 [0.60, 1.00]	1, 0.80
Fluency	et-fi	0.82 [0.60, 1.00]	0.92 [0.77, 1.00]	2, 0.44
Fluency	et-hu	0.70 [0.60, 0.87]	0.83 [0.77, 0.94]	3, 0.28
Fluency	fi-en	0.62 [0.33, 0.87]	0.77 [0.54, 0.94]	3, 0.27
Fluency	fi-hu	0.88 [0.87, 1.00]	0.95 [0.94, 1.00]	1, 0.81
Fluency	hu-en	0.51 [0.20, 0.73]	0.69 [0.37, 0.89]	4, 0.14
Grammar	et-en	0.55 [0.33, 0.73]	0.73 [0.54, 0.83]	3, 0.26
Grammar	et-fi	0.84 [0.73, 1.00]	0.93 [0.83, 1.00]	1, 0.78
Grammar	et-hu	0.90 [0.73, 1.00]	0.96 [0.83, 1.00]	1, 0.78
Grammar	fi-en	0.46 [0.20, 0.60]	0.66 [0.31, 0.77]	4, 0.13
Grammar	fi-hu	0.93 [0.73, 1.00]	0.97 [0.89, 1.00]	0, 1.00
Grammar	hu-en	0.51 [0.33, 0.60]	0.71 [0.49, 0.77]	4, 0.13
LRA	et-en	0.06 [-0.20, 0.33]	0.16 [-0.20, 0.49]	7, 0.02
LRA	et-fi	-0.15 [-0.47, 0.20]	-0.17 [-0.60, 0.26]	9, 0.01
LRA	et-hu	0.10 [-0.20, 0.33]	0.07 [-0.31, 0.37]	6, 0.03
LRA	fi-en	0.72 [0.33, 1.00]	0.82 [0.49, 1.00]	2, 0.38
LRA	fi-hu	0.02 [-0.33, 0.33]	0.02 [-0.31, 0.49]	7, 0.02
LRA	hu-en	0.13 [-0.07, 0.33]	0.25 [-0.09, 0.54]	7, 0.02
Readability	et-en	0.72 [0.55, 0.83]	0.84 [0.70, 0.93]	2, 0.45
Readability	et-fi	0.95 [0.87, 1.00]	0.98 [0.94, 1.00]	0, 1.00
Readability	et-hu	0.99 [0.87, 1.00]	1.00 [0.94, 1.00]	0, 1.00
Readability	fi-en	0.76 [0.55, 0.97]	0.87 [0.70, 0.99]	2, 0.43
Readability	fi-hu	0.95 [0.87, 1.00]	0.98 [0.94, 1.00]	0, 1.00
Readability	hu-en	0.72 [0.55, 0.83]	0.83 [0.70, 0.93]	2, 0.44

Table 11: Cross-language ranking stability: Kendall  $\tau$  and Spearman  $\rho$  and inversion counts (obs) with permutation  $p$ -values. Significant inversions ( $p < 0.05$ ) indicate that rankings are not preserved across languages under the given evaluation.

Model	Grammar (G)		Readability (R)		Coherence (C)		Fluency (F)	
	en	et	en	et	en	et	en	et
gpt-4.1-mini	3.17	3.18	3.63	3.60	2.99	2.99	2.35	2.36
llama-3.3-70b-inst.	2.39	2.38	2.92	2.98	2.93	2.94	1.87	1.86
claude-sonnet-4	3.04	3.08	3.63	3.64	2.98	2.98	2.29	2.32
mixtral-8x7b-inst.	1.63	1.63	1.99	1.93	2.72	2.72	1.17	1.22
llama-3.1-8b-inst.	1.61	1.64	1.87	1.88	2.34	2.24	1.06	1.09
command-r	1.50	1.52	1.81	1.83	2.56	2.45	1.04	1.07

Table 12: Comparison of LLM-as-a-judge mean scores using English (en) vs. Estonian (et) meta-prompts for the Estonian calibration set. The negligible variance confirms that evaluation stability is robust to the language of instructions.

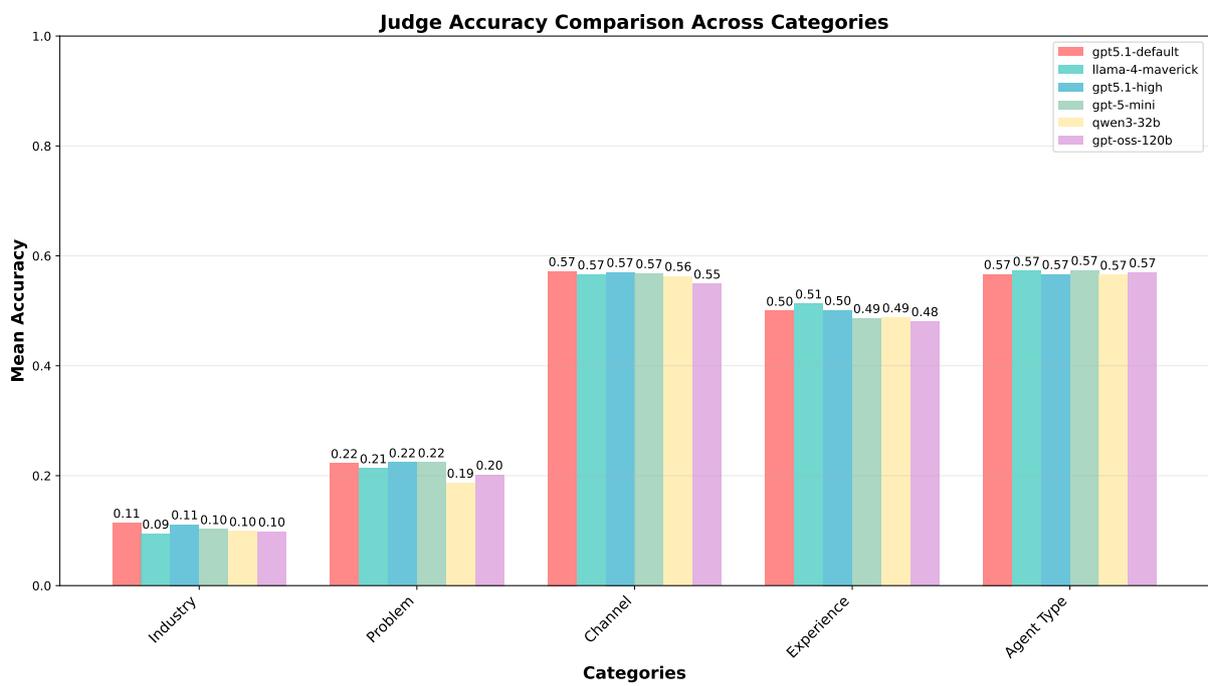


Figure 4: Comparison between different LLM judges over Finnish dialogues across LLM-as-a-judge metrics.

# Cross-lingual and cross-country approaches to argument component detection: a comparative study

Cecilia Graiff<sup>1</sup>, Chloé Clavel<sup>1</sup>, Benoît Sagot<sup>1</sup>

<sup>1</sup> ALMAnaCH, Inria Paris  
name.surname@inria.fr

## Abstract

Argument mining in multilingual settings has rarely been investigated, due to the lack of annotated resources and to the inherent difficulty of the task. We benchmark the performance of models on cross-lingual and cross-country argument component detection, focusing on political data from the US and France. To do so, we introduce FrenchPolArg, a corpus of argumentative political discourse in French, and we automatically translate already existing US-English resources. We benchmark three different cross-lingual and cross-country pipelines, and compare their results to find the best-performing one. We obtain promising results to be integrated in semi-automatic annotation workflows to reduce the time and cost of annotations.

## 1 Introduction

Two main gaps in argumentation mining research are the lack of annotated data and the poor generalizability of models across domains, languages, and datasets. Because of the high subjectivity of the task and the differences in argumentation styles, models tend to learn data rather than arguments (Feger et al., 2025). Moreover, the available datasets are mostly in English, and do not cover a broad range of topics. This paper aims at evaluating the performance of a cross-lingual pipeline for the task of argument component detection in a bilingual setting, thus contributing to the generalizability of argument retrieval by language models. We focus our analysis on political-domain data in the US and France. As can be imagined, this setting presents differences in both language and culture. While we are aware of the necessity of differentiating between the former and the latter, this paper will not investigate the relation between them, due to our choice of focusing on cross-lingual model generalizability rather than an accurate sociolinguistic analysis. To avoid an oversimplified use of

the expression “cultural differences,” we will refer to this aspect as “country-level.” We further define our work with the following research questions (RQs):

**RQ 1: How much do language differences influence the model’s performance?** We translate the English dataset ElecDeb60to20 and project the annotations, with the purpose of evaluating multilingual models on the translated version, thus assessing the performance drop. This approach allows us to focus on the cross-lingual aspect, and obtain a faithful estimation of how much adopting a cross-lingual pipeline impacts the model’s performance. This experimental setting is crucial to assess the importance of one of the main problems of the current state of the art in argumentation mining research, namely the difficulty of learning and generalizing tasks between two languages.

**RQ 2: How much do country-level differences impact the performance of language models in the argument component detection task?** While the problem of robustness and generalizability is inherent to the argument mining task, the further nuance of country-level difference is rarely taken into account, even though some existing studies focus on cross-lingual or cross-domain problems (Eger et al., 2018; Yeginbergen et al., 2024a; Schaefer et al., 2022). However, we stress the relation between language, culture, and argumentation style, thus investigating its impact on model performance. We add a further layer to the first research question by increasing the difficulty of model generalizability, adding a change of culture to the change of language.

**RQ 3: How can we improve the cross-lingual and cross-country generalizability of language models for the argument component detection task and allow their reusability on different datasets?** One of the main challenges of argu-

ment mining is the lack of available annotated data, especially in multilingual scenarios. Therefore, we deliver information about which ones among the chosen cross-lingual and cross-country pipelines help improve generalizability. This part of the work also leads the way to the creation of a semi-automatic annotation workflow to help the costly and difficult process of delivering new annotated datasets.

**Contributions** Our contributions are as follows:

1. We build FrenchPolArg, a corpus of French presidential debates and speeches, and annotate part of it to create a ground truth. We aim at providing a dataset for a domain for which, to the best of our knowledge, no annotated resource is currently available in French. Because non-English data is particularly scarce in the argument mining field, we believe that this resource addresses one of the most important gaps in the current research.
2. We automatically translate into French ElecDeb60to20 (Goffredo et al., 2023), a corpus of US presidential debates annotated for argument components. We project the original annotations of this corpus, thus obtaining an annotated resource of argumentative political text in French. We are aware that language is not the only difference between a French and an English corpus - cultural differences certainly play a role too. Our aim is to leverage these data for cross-lingual experiments.
3. We benchmark the performances of language models on cross-language and cross-country transfer, and compare the results to assess performance drop. Furthermore, we deliver information about the best performing pipeline to address language and country differences.

## 2 Related Work

Argument component detection is a subtask of argument mining that focuses on detecting claims and premises. Several studies have been published on the topic (Lawrence and Reed, 2020), but they mostly focus on a monolingual setting with no domain variation, leaving a research gap concerning multilingual and multicultural approaches. Because this task lies between the realms of argument mining and claim detection, we extend our literature review to related tasks such as cross-domain argument mining and claim detection.

### **Cross-lingual transfer learning with masked language models**

Cross-lingual transfer learning aims to enhance the performance of models on target languages by leveraging knowledge acquired from different source languages (Zhuang et al., 2021). Because some preliminary experiments with LLMs did not deliver good results, we chose to focus on BERT-based models. The better performance of BERT-based models on similar tasks, such as claim detection or fake news detection, is confirmed in the literature. Raza et al. (2024) benchmark BERT-based models and LLMs on the classification of fake news detection, and report F1 scores close to 90% for roberta-base-uncased, while the best performing LLM-based approach, which is fine-tuning Mistral-7B-v0.2, reports an F1 score of 80.23%. Azuma et al. (2025) perform an accurate comparative study of SVMs, BERT-like models, and LLMs (Mistral, LLaMA 3.2-3B, and Qwen3-4B) and determine that for specialized tasks, task-specific fine-tuning of a smaller, specialized model remains a more effective and computationally efficient approach than in-context learning or even parameter-efficient fine-tuning with larger, general-purpose generative models. Among the models leveraged for cross-lingual transfer learning, mBERT achieves very good results. mBERT follows the architecture of BERT, and was trained on with data from Wikipedia in 104 languages and without cross-lingual signal. Muller et al. (2021) perform a structural and behavioral analysis of the language-transfer capabilities of mBERT, concluding that this language model is composed by two sub-networks. The first one is a multilingual encoder, followed by a task-specific language-agnostic predictor. Wu and Dredze (2019) test mBERT on 5 different tasks (document classification, natural language inference, named entity recognition, part-of-speech tagging, and dependency parsing) from English to 38 target languages. Their results show that mBERT always achieves optimal results, sometimes state-of-the-arts results. Cross-lingual transfer for sequence labeling tasks is evaluated by García-Ferrero et al. (2022) and more specifically for argument mining by Yeginbergen et al. (2024a), as explained in the next paragraph.

**Cross-lingual argument mining** One of the first comprehensive experiments about multilingual argument component detection is (Eger et al., 2018). Similarly to the first part of our methodology, they (automatically and manually) translated into Ger-

man, French, Spanish, and Chinese an English dataset of student essays and projected the annotations. While they obtained relatively good result, the inevitable performance drop is clearly visible. (Yeginbergen et al., 2024a) demonstrate that in the case of multilingual argument mining, data-transfer methods outperform model-transfer, where “model-transfer” denotes the use of a language model’s experience in one language to generalize and apply it to another language. Their study is a further proof of the tendency of language models to learn data rather than the argument mining task itself. However, their results differ from prior findings on related sequence labeling tasks in the community (García-Ferrero et al., 2022). Our paper partially builds on their proposed approach by leveraging the automatic translation and annotation projection pipeline. While delivering different results from their paper, RQ3 partly shares their aim and structure. However, they work with the medical dataset AbsRCT (Mayer et al., 2020), consisting of randomized controlled trials retrieved from the MEDLINE database via PubMed search, whereas our dataset is political and presents a conversational argumentative context. Therefore, our works entails important differences that make our contribution unique, as no similar research exists on political and conversational data. It is important to mention that we introduce the cross-country differences, which we compare to cross-domain tasks, whereas Yeginbergen et al. (2024a) is focused on cross-lingual transfer. To do so, we build a new dataset of French political argument, leveraged as unseen test set. (Schaefer et al., 2022) experiment with claim detection with BERT and RoBERTa on several dataset combinations in order to find the best composition for training, which appears to have large corpus size, homogeneous claim proportions, and less formal text domains. The combination of different datasets is a strategy used in this paper as well, as it allows to augment the training size while decreasing the probability of the model overfitting on a single dataset. Their approach aims at finding the best dataset combination, hence their choice of experimenting only with RoBERTa. Differently from them, we include data mixing as an approach among others, and experiment with several models. Moreover, they work in a monolingual setting, while this paper merges the cross-lingual and the cross-country transfer, thus providing different insights.

**Cross-domain argument mining** While language models are able to perform cross-lingual transfer learning, it is worth noticing that a shift in language does not take into account country-related differences. For this reason, we consider the methods applied to cross-domain argumentation research to be pertinent to our work. (Daxenberger et al., 2017) were among the first to compare in-domain and cross-domain claim detection. More specifically, they compared user-generated web discourse such as blog posts, persuasive essays, online comments, Wikipedia talk pages, argumentative microtexts, and include the remaining categories as “various genres.” Their results reveal a strong performance drop in cross-domain contexts. Moreover, they observe that feature-based approaches outperformed deep learning approaches, thus revealing that the presence of lexical indicators guides the model’s choice. A similar result is achieved by (Alhamzeh et al., 2021), with slightly higher scores for traditional feature-based methods compared to DistilBERT. This paper also shows performance improvement when leveraging ensemble learning, and focuses on user-generated web content and student essays. However, while they obtain an F1 score of 0.79 on the cross-domain task, it is worth noticing that they train the model on a mix of both the corpora they investigate, thus not focusing on generalizability and allowing the model to still learn the dataset more than the task itself. In our work, we benchmark other approaches as well, to avoid biasing our interpretation of the cross-lingual transfer abilities of models. A more recent work on stance classification evidences the lack of cross-domain works, as well as the necessity for annotated cross-domain data (Yuan et al., 2024). While our setting is comparable to cross-domain under the technical point of view, the cross-country question that we are analyzing is underrepresented in the field.

### 3 Methodology

The cross-lingual aspect of this research is first focused on a bilingual study of English and French. We test three approaches on these two languages, namely (1) data transfer, (2) model transfer, and (3) data mixing. To the best of our knowledge, no annotated resource exists in French for the political domain. Therefore, we build FrenchPolArg, a French dataset of political discourse, and partially annotate it with argument components (claims

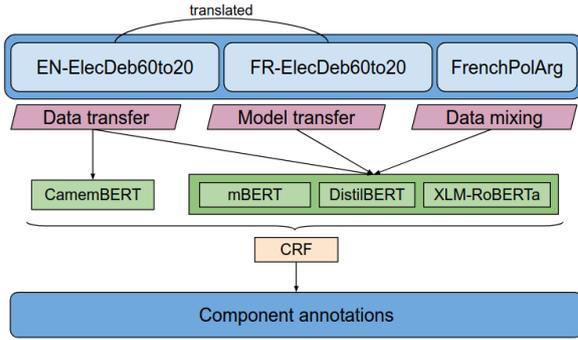


Figure 1: Pipeline of our work.

and premises) to build a ground truth. To perform cross-lingual experiments that can lead the way to a semiautomatic annotation workflow, we translate ElecDeb60to20 (Goffredo et al., 2023), a collection of presidential debates in the US from 1960 to 2020 annotated for argumentative components and relations. The implemented pipeline follows the approach described in (Yeginbergen et al., 2024b), based on machine translation (MT) and annotation projection. We translate the dataset into French, thereby generating parallel training sets used to fine-tune BERT-based classifiers for the task in the three settings mentioned above. We subsequently evaluate the model on the unseen real-world data contained in FrenchPolArg, to assess model performance and to perform a comparative analysis of the approaches. The data and the code are available [here](#). An overview of the pipeline used for this paper is available in Figure 1. In the next paragraphs, we will present the used approaches and the dataset collection, and later we will explain the experimental setting and fine-tuning steps.

### 3.1 Followed approaches

We adopt three different approaches, based on:

1. **Data transfer:** Data transfer consists in generating training data by translating ElecDeb60to20 (Goffredo et al., 2023) into French and projecting the annotations, thus obtaining FR-ElecDeb60to20. We test this setting with both multilingual and monolingual models. While we also leverage multilingual models in the model transfer strategy, we consider the latter fundamentally different, because the models are trained and tested on data in the same language.
2. **Model transfer:** Differently from the data transfer approach, model transfer consists

in leveraging the cross-lingual transfer abilities of the model itself, rather than applying a pipeline of translation and projection with the aim of generating new data containing the same knowledge in another language. We leverage multilingual pre-trained language models. We use the standard English version of ElecDeb60to20 as a training set, and test the model on its French translation to investigate the impact of language differences. We then test on FrenchPolArg to add the further layer of country differences.

3. **Data mixing:** We progressively augment the French and English versions of ElecDeb60to20 with data from FrenchPolArg, in order to iteratively improve the results. We consider the limited size of the annotated portion of FrenchPolArg as a limitation. Therefore, the data augmentation approach allows us to have sufficient data for training a BERT-based model, while diminishing the probability of the model overfitting on FrenchPolArg.

## 3.2 Dataset creation

### 3.2.1 FrenchPolArg

**Dataset collection** FrenchPolArg is composed by the transcripts of 8 presidential debates broadcasted on television in 1974, 1981, 1988, 1995, 2007, 2012, 2017, and 2022. The debates are publicly available [online](#). When a transcript already existed, it was scraped; alternatively, the YouTube video of the debate was transcribed and diarised with the WhisperX model (Bain et al., 2023), which provides fast automatic speech recognition (70x real time with large-v2) and was chosen over Whisper because it includes speaker diarisation. Moreover, transcripts of French presidential speeches and declarations were scraped from the official website of the [Élysée](#) and of [Vie Publique](#). Because this is not a conversational setting, we choose not to focus on it for this paper. As the speeches are uploaded on the website in PDF form with plenty of noisy text, the scraped content was thoroughly preprocessed, both automatically and manually. As far as we know, the only similar resource is FREDSum (Rennard et al., 2023), a corpus of French political debate annotated for summarization. While the data partially overlap, our contribution differs from FREDSum because it covers both debate and speeches, it does not entail a separation based on topics but only on debates, and it is partially an-

notated with argument components, as will be explained in the next section. Differently from FRED-Sum, we chose to not split the dataset based on the debate topics, as we aim at having a dataset that spans more broadly in order to better evaluate generalizability in this paper.

**Annotation of the test set** Due to the length and complexity of the annotation task, it was not possible to annotate the whole FrenchPolArg. This resource problem led us to choose to annotate a part of the dataset to use as ground truth to evaluate model generalizability. To minimize diachronic bias, we annotate the first, last, and middle available debate, namely the ones between François Mitterrand and Valéry Giscard d’Estaing (1981), Jacques Chirac and Lionel Jospin (1995), and Emmanuel Macron and Marine Le Pen (2022). The annotations were performed by a 25-year-old French-speaking female annotator, who also is one of the authors of this paper.

We define a set of arguments  $A$ , composed of a set of premises  $P$  and a set of claims  $C$ , where the premises have the function of supporting or attacking the claims. To take into account the faulty logical reasoning often present in natural language arguments, we allow the existence of claims that are not supported by premises, but not of premises that do not support any claim. We choose to not limit the amount of accepted claims to only one, to allow more complex argument structures. The annotation of argumentative components was performed with the INCEpTION platform and focused on the token-level. We annotated 27,299 tokens (1509 sentences) following the Beginning-Inside-Outside (BIO) scheme (Ramshaw and Marcus, 1995), which marks tokens as beginning, inside, or outside an argument component.

### 3.2.2 Translated ElecDeb60to20

**Description of the original dataset** To the best of our knowledge, the only political dataset annotated for argument components and relations is ElecDeb60to20 (Goffredo et al., 2023). This dataset contains the debates between presidency and vice-presidency candidates of the United States of America from 1960 until 2020. The raw data was originally scraped from the Commission on Presidential Debates (CPD). A first round of annotations was performed on the data from 1960 to 2016; later, the dataset was extended to comprehend the debates until 2020. The annotation task

was performed manually, reaching a high inter-annotators agreement with values such as a Krippendorff’s alpha of 0.757 for the second annotation round. Apart from its political nature, we chose this dataset because the annotation scheme is compatible with ours, explained in the next paragraphs. Moreover, the good results obtained by ElecDeb60to20 in previous research works in the argument mining field deemed it a suitable choice. The use of ElecDeb60to20 required accurate pre-processing. The part of the dataset annotated for argumentative components presented several duplicates, that we removed automatically. Duplicates with different annotations were removed manually and not automatically, in order to avoid the potential loss of information and to ensure the quality of the final dataset.

**Translation and annotation projection** Following the pipeline implemented by (Yeginbergen et al., 2024a), we automatically translate ElecDeb60to20 into French<sup>1</sup>. We leverage Opus-MT (Tiedemann and Thottingal, 2020), a set of language models pre-trained on the machine translation task. Even though Yeginbergen et al. (2024a) observe a slightly better performance of DeepL compared to Opus-MT, we choose the latter model because it is open source. While our focus in this paper is bilingual, we aim at extending this work to a multilingual setting, thus motivating the necessity of a translation service that covers a wide range of languages. Therefore, we make the translated corpus available in French. The original annotations of components were projected with SimAlign (Sabet et al., 2021), particularly suited to the task of aligning sequences in different languages, where the structure of sentences does not necessarily match. Here as well, the choice is motivated by the good results reported by Yeginbergen et al. (2024a) and García-Ferrero et al. (2022) on this task. For further clarity, we present an overview of the used datasets in Table 1.

---

<sup>1</sup>We also provide translations into German, Spanish, and Italian. This paper focuses on a bilingual study, but we plan to expand it to a multilingual setting.

Dataset Name	Dataset Description	Size (tokens)
EN-ElecDeb60to20	Annotated US presidential debates in English. (Goffredo et al., 2023).	707,976
FR-ElecDeb60to20	French translation of EN-ElecDeb60to20.	741,835
FrenchPolArg	Annotated French presidential debates.	68,392

Table 1: Overview on the used datasets.

### 3.3 Model training

For each of the above presented settings, we test several different configurations. In this section, we will present the tasks and the data splitting procedure; for an overview of the selected architectures, please refer to Appendix B.

**Fine-tuning task** The models are fine-tuned on a multiclass token classification task, following the adopted BIO annotation scheme to represent components’ boundaries. Given a sentence  $S$  consisting of  $n$  tokens, our goal is to predict for each token  $t_i$  if it is a claim ( $y_i = 1$ ), a premise ( $y_i = 2$ ), or not argumentative ( $y_i = 0$ ). Therefore, our task is originally formulated as a 3-classes token classification task, with the aim of labeling tokens as claim, premise, or non argumentative. However, because we adopt the BIO annotation scheme, our classes are expanded to include the begin (B), inside (I), and outside (O) tags. Moreover, we want to exclude the tags generated by the tokenizer, such as [SEP], [CLS], [PAD]. Therefore, our original 3-classes problem is expanded to a 6-classes problem: B-Claim, I-Claim, B-Premise, I-Premise, O for non argumentative tokens, and X for the tokenizers’ tags. Because the model’s performance on classifying the tags reached top accuracy (more than 0.99), we consider it influential and evaluate our model on the 5 remaining classes.

**Data splitting** We split the monolingual datasets respecting class proportions, as explained in detail in appendix A. For the multilingual dataset, we split equally the train, dev, and test set between the two languages to avoid cross-lingual duplicates. Because both English and French are not considered low-resource languages, we assume that we can rely on the transfer learning capabilities of multilingual models. However, we are aware of the bias caused by the fact that the amount of English in their training data is superior to any other

language. In the data mixing setting, the training dataset is built by adding samples from the target dataset to FR-ElecDeb60to20. We avoid increasing the training dataset with more than 20% of FrenchPolArg to not decrease too much the test set size. We call the thus obtained subsets FrenchPolArg-S1 and FrenchPolArg-S2, where the first one is used to increase the train dataset size, and the second one is used as test dataset. We call the increased training datasets EN-aug, which consists of EN-ElecDeb60to20 augmented with FrenchPolArg-S1, and FR-aug, which consist of FR-ElecDeb60to20 augmented with FrenchPolArg-S1. All experiments are tested on FrenchPolArg-S2.

### 3.4 Evaluation

#### 3.4.1 Evaluation of the translation and projection pipeline

We evaluate the translation and projection pipeline by manually correcting a randomly selected sample of 100 sentences (1556 tokens) and checking the token distribution. We report in Table 2 the difference in the token distribution between the original and backtranslated English versions of ElecDeb60to20. The backtranslated version is obtained by automatically translating ElecDeb60to20 into French and projecting the annotations, and then repeating the same process from French to English. These absolute and percentual values are used as an evaluation of the accuracy of the projections, together with the manual evaluation of a random sample of 100 sentences (1556 tokens). We focus on the B-tokens, which represent the amount of claims and premises in both datasets. B-Claim and B-Premise present a decrease, meaning that the annotation missed some tags, but the percentual difference in the whole dataset is very small (around 0.25%). Parallel to this quantitative analysis, we perform a qualitative one by analyzing 100 random sentences (1556 tokens). We report only two minor mistakes in the projection, and only three cases where a too literal translation hinders the understanding of FR-ElecDeb60to20 (such as “donner un coup de pied” for “kicking in”). We are aware that no perfect model exists, and consider the obtained results to be satisfactory enough to ensure the overall correctness of the translation projection pipeline. Moreover, as our aim is to test this configuration, we do not correct the errors manually, in order to obtain an unbiased evaluation of our approach.

Tag	Original	Backtranslated	Difference (%)
O	226,333	241,982	+4.2%
B-Claim	12,147	10,458	-0.26%
I-Claim	139,457	134,204	-0.25%
B-Premise	10,743	9,100	-0.26%
I-Premise	145,506	123,473	-3.48%

Table 2: Differences in the token distribution between the original and backtranslated English version of ElecDeb60to20. The percentage indicates the change in relation to the whole corpus. We use these numbers to evaluate the quality of annotation projection.

**Evaluation of the cross-lingual transfer experiments** The models are evaluated with micro and macro F1, which are common metrics chosen for sequence classification tasks. While macro F1 weighs each class equally, micro F1 weighs each sample equally, thus allowing us to perform a proper evaluation even though the classes are not equally distributed. Because the differences between micro and macro F1 were minimal, we report only the macro scores in the paper. Full tables comprehending the micro F1 scores are available in Appendix D. To avoid that the model’s ability of recognizing non-argumentative tokens (labeled as O) biases the metric of model accuracy, we also report the specific F1 scores for claims and premises in Appendix D.

## 4 Results

With our experiments, we aim at answering the following research questions:

**RQ 1: How much do language differences influence the model’s performance?** Before testing the three previously described approaches, we extensively experiment with EN-ElecDeb60to20 and FR-ElecDeb60to20 to investigate the impact of the language shift while eliminating country-related confounders. We present the results in Appendix D. As expected, we note a performance drop: the task reports an F1 score of 0.637 (rounded to 0.63) on the English dataset with the best performing model (mBERT), and of 0.46 on the French dataset, suggesting that the change in language has an important impact on the model’s performance. Interestingly, we do not report significant differences among the models’ performances, including CamemBERT, the only non-multilingual one. While the absolute values are lower, the proportional decrease described in this paper does not show important differences.

We also run the same experiment on the backtranslated version of the English corpus. The results are reported in Appendix E. We notice that while the translation to French involved a performance decrease, testing mBERT on the backtranslation does not present significant differences in performance compared to the French version. Hence, we argue that the translation, while not perfect, is not enough to explain all of the performance drop. Because the evaluation of the translation and projection pipeline was satisfactory, we exclude this possible cause, and hypothesize that the difficulty of cross-lingual transfer of the argument mining task be the reason. Therefore, we are convinced that the model’s robustness needs to be improved also under the cross-lingual point of view.

To further prove this point, we test a model trained on EN-ElecDeb60to20 on its French translation, which causes a decrease in performance from 0.637 (rounded to 0.63) to 0.559 (rounded to 0.55). The results are reported in Table 9. These results confirm the scarce ability of language models to generalize to different languages in the argument mining task, and deliver further proof of the impact of language shifts on model performance, thus answering our RQ 1.

Language	Model	Macro F1
EN	mBERT	0.63
EN	XLM-RoBERTa	0.58
FR	mBERT	0.46
FR	XLM-RoBERTa	0.46
FR	CamemBERT	0.45
FR	DistilBERT	0.44

Table 3: Experiments on the original and the translated version of ElecDeb60to20. All experiments present the same language in the training and test set. We use the base, multilingual, cased version of DistilBERT, and the base, multilingual, uncased version of BERT. All models have the addition of a CRF as last layer.

Model	Macro F1	F1 Premise	F1 Claim
mBERT	0.55	0.51	0.46
XLM-RoBERTa	0.39	0.15	0.29

Table 4: Results of the models trained on EN-ElecDeb60to20 and tested on FR-ElecDeb60to20. We use the base, multilingual, uncased version of mBERT and XLM-RoBERTa.

**RQ 2: How much do country-level differences impact the performance of language models in**

**the argument component detection task?** To include the country-level differences, we test the pipeline on FrenchPolArg, which is in the same language as FR-ElecDeb60to20, but presents a different culture. Our experiments can be seen in Table 5, while a more extensive report containing both micro and macro F1 score can be found in appendix D. We notice that the performance drop is only slightly superior to the experiments presented in Table 3, with mBERT scoring 0.55 macro F1 when tested on FR-ElecDeb60to20, and 0.50 on FrenchPolArg. While we still are in the political domain, we hypothesize that the argumentation styles might differ from one country to the other. Therefore, we interpret our experiments as a proof that while country-level differences can cause a bias, the main issue to address in the field concerns the cross-lingual differences.

**RQ 3: How can we improve the cross-lingual and cross-country generalizability of language models for the argument component detection task and allow their reusability on different datasets?** To answer this research question, we compare the three different approaches described in this paper (model transfer, data transfer, and data mixing) and test them in the above described configurations. We present a summary of all the conducted experiments in Tables 5 and 6, while a more extensive report containing both micro and macro F1 score can be found in appendix D. Differently from the results of Yeginbergen et al. (2024a), we notice that model transfer delivers better results than data transfer, with 0.50 as best results versus 0.43.

Among the model transfer experiments, mBERT scores significantly better than the other models, while still reporting a performance drop compared to the test on EN-ElecDeb60to20 reported in Table 3. We also note the very poor performance of XLM-RoBERTa, which is the only model to score better in the data transfer approach, but delivers worse results than any other model in both approaches. This result is surprising, because XLM-RoBERTa is usually reported to score better results on multilingual tasks; however, we hypothesize that the small size of the dataset might have had an impact on the performance of XLM-RoBERTa, while smaller models manage to deliver better results. We present further details about the chosen models in Appendix B, but because of time- and resource-related limitations, we could not investigate this

Model	Macro F1
<i>Model transfer</i>	
mBERT	0.50
distilBERT	0.49
XLM-RoBERTa	0.35
<i>Data Transfer</i>	
CamemBERT	0.40
mBERT	0.42
distilBERT	0.43
XLM-RoBERTa	0.40
<i>Data+Model Transfer</i>	
mBERT	0.47

Table 5: All experiments and models. The models are tested on FrenchPolArg and trained respectively on EN-ElecDeb60to20 for the model transfer setting, FR-ElecDeb60to20 for the data transfer setting, and the multilingual dataset FR+EN-ElecDeb60to20 for the data+model transfer. We use the base, multilingual, cased version of DistilBERT, and the base, multilingual, uncased version of BERT.

aspect any further.

Our results show that the data transfer approach delivers relatively poor results for all models. The best performing model in this scenario is distilBERT (F1 score of 0.43), but the difference with BERT (F1 score of 0.42) is negligible. We notice that there are no significant differences between the performances of monolingual models such as CamemBERT, and multilingual models.

Based on these results, we hypothesize that argumentative discourse is such a language- and culture-dependent process that even a correct translation significantly decreases accuracy. While the proposed approaches are not perfect, we suggest that data mixing as the best performing one with a 0.585 macro F1 score could be applied to unseen data in order to provide a first batch of annotations, to later revise manually. This semi-automatic workflow would significantly reduce the annotation cost and time.

## 5 Conclusion and Future Work

In this paper, we present an original French annotated resource. In addition, we provide the translated version of ElecDeb60to20 into French. We then test cross-lingual and cross-country generalizability, and provide information on the best-working approach. Therefore, we provide new resources in a severely lacking domain, and deliver information about the open problem of generalizability. For the future direction of this work, we

Train dataset	Model	Macro F1 score
<i>Data Mixing</i>		
EN-aug	XLM-Roberta	0.51
FR-aug	XLM-Roberta	0.44
EN-aug	mBERT	0.58
FR-aug	mBERT	0.47
EN-aug	distilBERT	0.55
FR-aug	distilBERT	0.43
FR-aug	CamemBERT	0.43

Table 6: All experiments and models in the data mixing setting. We use the base, multilingual, cased version of DistilBERT, mBERT, and XLM-RoBERTa. EN-aug refers to EN-ElecDeb60to20 augmented with 20% of the data from FrenchPolArg, and FR-aug refers to the same augmentation applied to FR-ElecDeb60to20. All models have the addition of a CRF as last layer and all models are tested on the remaining 80% of FrenchPolArg.

believe that increasing the provided resources and their quality is essential. We do not exclude to implement further strategies to diminish the impact of the small dataset size, such as k-fold cross validation. Moreover, it would be very important to be able to further investigate the lower performance of XLM-RoBERTa on this dataset, which was not possible for this paper because of time and resources. Most importantly, we planned an annotation campaign to expand our dataset and widen the experimental setting. We then plan to extend the experiments of this paper to include more data, thus increasing the reliability of the results, even though we work on strategies for scarce data settings. We are interested in merging the presented pipelines with manual annotation workflows, thus reducing time and the cost of annotations. A further planned approach is to extend the translations to include other languages, such as German, Italian, and Spanish.

## 6 Limitations

While this work offers an insight into cross-lingual and cross-country generalizability, it entails several limitations. Argument mining established a clear definition of arguments as units of speech composed by claim and premises, connected by support or attack relations; however, the annotation task itself is more complicated, because real-world data struggle to match the mathematical formalizations of arguments. Moreover, FrenchPolArg was annotated by one single person, due to the length, complexity, and cost of the process. However, this paper aims at delivering a baseline for further research in scarce data settings. This ex-

plains another limitation of this paper, namely the dataset size. To avoid that this limitation would bias our results, we relied on an already existent bigger dataset, namely ElecDeb60to20. Despite this measure, the tests performed on FrenchPolArg rely on a smaller test dataset size, which possibly biases the results, a limitation to be addressed in the future works described above.

## References

- Alaa Alhamzeh, Mohamed Bouhaouel, Előd Egyed-Zsigmond, Jelena Mitrović, Lionel Brunie, and Harald Kosch. 2021. A stacking approach for cross-domain argument identification. In *Database and Expert Systems Applications*, pages 361–373, Cham. Springer International Publishing.
- Daichi Azuma, René Meléndez, Michal Ptaszynski, Fumito Masui, Lara Aslan, and Juuso Eronen. 2025. [Svm, bert, or llm? a comparative study on multilingual instructed deception detection](#). *AI*, 6(9).
- Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. Whisperm: Time-accurate speech transcription of long-form audio. *INTERSPEECH 2023*.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [Unsupervised cross-lingual representation learning at scale](#). *Preprint*, arXiv:1911.02116.
- CPD. The commission on presidential debates. <https://debates.org/>. Accessed: 2025-10-03.
- Johannes Daxenberger, Steffen Eger, Ivan Habernal, Christian Stab, and Iryna Gurevych. 2017. [What is the essence of a claim? cross-domain claim identification](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2055–2066, Copenhagen, Denmark. Association for Computational Linguistics.
- Steffen Eger, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. 2018. [Cross-lingual argumentation mining: Machine translation \(and a bit of projection\) is all you need!](#) In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 831–844, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Marc Feger, Katarina Boland, and Stefan Dietze. 2025. [Limited generalizability in argument mining: State-of-the-art models learn datasets, not arguments](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23900–23915, Vienna, Austria. Association for Computational Linguistics.
- Iker García-Ferrero, Rodrigo Agerri, and German Rigau. 2022. [Model and data transfer for cross-lingual sequence labelling in zero-resource settings](#). In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6403–6416, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Pierpaolo Goffredo, Mariana Chaves, Serena Villata, and Elena Cabrio. 2023. [Argument-based detection and classification of fallacies in political debates](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11101–11112, Singapore. Association for Computational Linguistics.
- Pierpaolo Goffredo, Shohreh Haddadan, Vorakit Vorakitphan, Elena Cabrio, and Serena Villata. 2022. [Fallacious argument classification in political debates](#). In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 4143–4149. International Joint Conferences on Artificial Intelligence Organization. Main Track.
- John Lawrence and Chris Reed. 2020. [Argument mining: A survey](#). *Computational Linguistics*, 45(4):765–818.
- Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. [Camembert: a tasty french language model](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.
- Tobias Mayer, Elena Cabrio, and Serena Villata. 2020. Transformer-based argument mining for healthcare applications. In *ECAI 2020 - 24th European Conference on Artificial Intelligence*, volume 325 of *Frontiers in Artificial Intelligence and Applications*, pages 2108–2115. IOS Press.
- Benjamin Muller, Yanai Elazar, Benoît Sagot, and Djamé Seddah. 2021. [First align, then predict: Understanding the cross-lingual ability of multilingual BERT](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2214–2231, Online. Association for Computational Linguistics.
- Lance Ramshaw and Mitch Marcus. 1995. [Text chunking using transformation-based learning](#). In *Third Workshop on Very Large Corpora*.
- Shaina Raza, Draï Paulen-Patterson, and Chen Ding. 2024. [Fake news detection: Comparative evaluation of bert-like models and large language models with generative ai-annotated data](#). *Preprint*, arXiv:2412.14276.
- Virgile Rennard, Guokan Shang, Damien Grari, Julie Hunter, and Michalis Vazirgiannis. 2023. [FREDSum: A dialogue summarization corpus for French political debates](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4241–4253, Singapore. Association for Computational Linguistics.
- Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2021. [Simalign: High quality word alignments without parallel training data using static and contextualized embeddings](#). *Preprint*, arXiv:2004.08728.
- Robin Schaefer, René Knaebel, and Manfred Stede. 2022. [On selecting training corpora for cross-domain claim detection](#). In *Proceedings of the 9th Workshop on Argument Mining*, pages 181–186, Online and in Gyeongju, Republic of Korea. International Conference on Computational Linguistics.

Jörg Tiedemann and Santhosh Thottingal. 2020. [Opusmt – building open translation services for the world](#). In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Switzerland. European Association for Machine Translation. Annual Conference of the European Association for Machine Translation, EAMT2020 ; Conference date: 03-11-2020 Through 05-11-2020.

Vittorio Torri and Francesca Ieva. 2023. [Polimi at clinkart: a conditional random field vs a bert-based approach](#).

Shijie Wu and Mark Dredze. 2019. [Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.

Anar Yeginbergen, Maite Oronoz, and Rodrigo Agerri. 2024a. [Argument mining in data scarce settings: Cross-lingual transfer and few-shot techniques](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11687–11699, Bangkok, Thailand. Association for Computational Linguistics.

Anar Yeginbergen, Maite Oronoz, and Rodrigo Agerri. 2024b. [Argument mining in data scarce settings: Cross-lingual transfer and few-shot techniques](#). *Preprint*, arXiv:2407.03748.

Jiaqing Yuan, Ruijie Xi, and Munindar P. Singh. 2024. [A benchmark for cross-domain argumentative stance classification on social media](#). *Preprint*, arXiv:2410.08900.

Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. 2021. [A comprehensive survey on transfer learning](#). *Proceedings of the IEEE*, 109(1):43–76.

## A Data splitting

We divide ElecDeb60to20 into a training and test set of respectively 90% and 10% of the dataset. We further split the training set into a train (90%) and evaluation set (10%). We thus obtain three subsets, namely training (76,5%), validation (13.5%), and test set (10%). We opt for this splitting format to try to keep proportions as accurate as possible when testing the experiments on ElecDeb60to20 or on FrenchPolArg. To avoid unbalanced data splits, we use the stratify parameter of `train_test_split`. We thus divide our data equally between the 5 classes, obtaining the statistics shown in Table 7.

	train	dev	test
B-Claim	18,829	3,415	2,446
I-Claim	218,449	39,277	28,731
B-Premise	16,851	2,955	2,176
I-Premise	231,777	40,391	30,394
O	389,718	68,069	50,768

Table 7: Dataset statistics.

For the data transfer setting, we split ElecDeb60to20 into a train and a validation set of respectively 90% and 10%, respecting class proportions. We then leverage FrenchPolArg as a test set.

## B Model selection and architecture

After prior unsuccessful experiments with the default architecture, we fine-tune BERT-based models with a Conditional Random Field (CRF) as last layer, following (Goffredo et al., 2022). CRFs are often used in tasks that rely on BIO tags, because they ensure that predicted sequences respect the structural constraints of the annotation scheme. We give further details about the probability computation in appendix C. Following this architecture, we train the following models:

**mBERT:** We fine-tune bert-base-multilingual-cased for token classification.

**MultilingualDistilBERT:** We fine-tune distilbert-base-multilingual-cased, a distilled version of BERT. This model is often used on smaller datasets, because it retains performance while diminishing computational costs.

**XLM-RoBERTa:** This model was trained on 2.5 TB of newly created clean CommonCrawl data in 100 languages and reported very good results on cross-lingual tasks, even for low-resource languages (Conneau et al., 2020). As we face the issue of data scarcity, we hypothesize that a model performing well in a domain characterized by the presence of very little data could work well in our context.

**CamemBERT:** We test camembert-base (Martin et al., 2020) on the French portion of the dataset. We aim at exploring the data-transfer functionality and hypothesize that training a French model on translated data could lead to better results. We cross-check the reliability of our approach by testing CamemBERT on the English portion of the dataset, expecting a performance drop. However, we are aware that, despite being a French language

model, CamemBERT has seen English data in its training phase, and therefore partially knows how to handle English words.

## C Use of CRFs

Differently from a standard linear classification layer, CRFs model the conditional probability of the classes as follows:

$$P(y | x) \propto \exp \left( \sum_j \lambda_j \sum_{i=1}^n f_j(y_{i-1}, y_i, x, i) \right)$$

where  $x$  is the vector of tokens observations) that form the sequence,  $y$  is the vector of labels (states) over the tokens,  $i$  is an index over the sequence tokens,  $n$  is the length of the sequence,  $j$  indexes the feature functions  $f_j$  and  $\lambda_j$  are the parameters to be learnt. (Torri and Ieva, 2023)

## D Complete evaluation metrics

We present in this section the complete evaluation metrics for the performed experiments. Specifically, Tables 8, 9, 10 and 11 present the micro and macro F1 scores for the overall performance of the model, whereas Table 12 reports the single F1 scores for Claim and Premise in the experiments performed on EN-ElecDeb60to20 and FR-ElecDeb60to20.

Language	Model	Macro F1	Micro F1
EN	mBERT	0.63	0.61
EN	XLM-RoBERTa	0.58	0.59
FR	mBERT	0.46	0.44
FR	XLM-RoBERTa	0.46	0.50
FR	CamemBERT	0.45	0.46
FR	DistilBERT	0.44	0.45

Table 8: Experiments on the original and the translated version of ElecDeb60to20. All experiments present the same language in the training and test set. We use the base, multilingual, cased version of DistilBERT, and the base, multilingual, uncased version of BERT. All models have the addition of a CRF as last layer.

Model	Macro F1	Micro F1	F1 Premise	F1 Claim
mBERT	0.55	0.56	0.51	0.46
XLM-RoBERTa	0.39	0.42	0.15	0.29

Table 9: Results of the models trained on EN-ElecDeb60to20 and tested on FR-ElecDeb60to20. We use the base, multilingual, uncased version of mBERT and XLM-RoBERTa.

Model	Macro F1	Micro F1
<b>Model transfer</b>		
mBERT	0.50	0.53
XLM-RoBERTa	0.41	0.43
<b>Data Transfer</b>		
CamemBERT	0.44	0.46
mBERT	0.42	0.41
XLM-RoBERTa	0.38	0.37
<b>Data+Model Transfer</b>		
mBERT	0.47	0.47

Table 10: All experiments and models. The models are tested on FrenchPolArg and trained respectively on EN-ElecDeb60to20 for the model transfer setting, FR-ElecDeb60to20 for the data transfer setting, and the multilingual dataset FR+EN-ElecDeb60to20 for the data+model transfer. We use the base, multilingual, cased version of DistilBERT, and the base, multilingual, uncased version of BERT.

Train dataset	Model	Macro F1	Micro F1
<b>Data Mixing</b>			
EN-aug	XLM-Roberta	0.52	0.54
FR-aug	XLM-Roberta	0.44	0.52
EN-aug	mBERT	0.58	0.59
FR-aug	mBERT	0.45	0.51
EN-aug	distilBERT	0.57	0.59
FR-aug	distilBERT	0.43	0.50
FR-aug	CamemBERT	0.44	0.50

Table 11: All experiments and models in the data mixing setting. We use the base, multilingual, cased version of DistilBERT, mBERT, and XLM-RoBERTa. EN-aug refers to EN-ElecDeb60to20 augmented with 20% of the data from FrenchPolArg, and FR-aug refers to the same augmentation applied to FR-ElecDeb60to20. All models have the addition of a CRF as last layer and all models are tested on the remaining 80% of FrenchPolArg.

Language	Model	F1 Claim	F1 Premise
EN	mBERT	0.588	0.536
EN	XLM-RoBERTa	0.487	0.420
FR	mBERT	0.369	0.333
FR	XLM-RoBERTa	0.338	0.290
FR	CamemBERT	0.341	0.315
FR	DistilBERT	0.338	0.313
FR+EN	mBERT	0.501	0.449

Table 12: Experiments on the original and the translated version of ElecDeb60to20. We use the base, multilingual, cased version of DistilBERT, and the base, multilingual, uncased version of BERT. All models have the addition of a CRF as last layer.

## E Backtranslation

Table 13 presents the results of the argument component classification experiments performed with

mBERT on the backtranslated English version of ElecDeb60to20. We chose mBERT as it was the best performing dataset in the previously tested settings.

Macro F1	Micro F1	F1 Premise	F1 Claim
0.47	0.46	0.37	0.34

Table 13: Results of the models trained on the back-translation into English of ElecDeb60to20. We use the base, multilingual, uncased version of mBERT, with the addition of a CRF as last layer.

## F Technical setup

We test our models in several configurations, with 5 epochs delivering the best results. We implement the AdamW optimizer. All models were trained on the cluster of our institution, using H100 GPUs.

# UNSC-Bench: Evaluating LLM Diplomatic Role-Playing Through UN Security Council Vote Prediction

Ayush Nangia<sup>1</sup>, Aman Gokrani<sup>1</sup>, Ruggero Marino Lazzaroni<sup>2</sup>

<sup>1</sup>Aasra Inc., <sup>2</sup>University of Graz

ayushnangia16@gmail.com, amaan.gokrani@gmail.com, ruggero.lazzaroni@uni-graz.at

## Abstract

This paper introduces UNSC-Bench, a benchmark for evaluating Large Language Models (LLMs) in simulating diplomatic decision-making through United Nations Security Council (UNSC) vote prediction. The dataset includes 469 UNSC resolutions from 1947 to 2025, with voting records for the five permanent members (P5) (United States, China, France, Russia, United Kingdom) and translations in four languages. We analyze 26 LLMs, along with thinking variants, across multiple P5 roles and find that (1) without explicit role assignment, models are diplomatically unaligned, defaulting to high YES rates and failing to match any P5 voting pattern, indicating they lack inherent diplomatic identity; (2) model capability (as measured by MMLU-Pro) is strongly correlated with role-playing accuracy; (3) regional models do not outperform others in predicting their home country’s votes; and (4) multilingual evaluation reveals that prompt language impacts model predictions, particularly for minority vote outcomes.

## 1 Introduction

The application of Large Language Models (LLMs) to geopolitical analysis and diplomatic simulation represents an emerging frontier in AI research (Ziems et al., 2024; Qi et al., 2025; Rivera et al., 2024). As these models become increasingly sophisticated, understanding their ability to adopt and maintain country-specific perspectives becomes crucial for applications ranging from policy simulation to international relations education. The United Nations Security Council (UNSC), with its formalized voting procedures and documented history spanning decades, provides an ideal testbed for evaluating such capabilities.

The Security Council’s five permanent members (P5): the United States, Russia, China, France, and the United Kingdom, possess veto power and demonstrate distinct voting patterns shaped by their

geopolitical interests, historical relationships, and policy priorities. Successfully predicting how each country would vote on a given resolution requires understanding these varied perspectives, making it a challenging test of diplomatic reasoning.

First, LLMs have been shown to encode geopolitical biases correlated with training data distribution and country of origin (Li et al., 2024a; Salnikov et al., 2025), suggesting regional models might better reproduce their home state’s diplomatic behavior. Second, multilingual evaluations find that performance and behavior shift with input language (Bang et al., 2023; Tam et al., 2025), raising the possibility of a “native language advantage” for country-specific reasoning. Third, work on LLM-based simulations emphasizes the importance of explicit personas for obtaining realistic responses (Aher et al., 2023; Hu and Collier, 2024; Li et al., 2024b), motivating comparison between neutral and role-play prompting. Finally, documented relationships between model capability and complex task performance (Kaplan et al., 2020; Wang et al., 2024a; Brown et al., 2020) suggest that diplomatic role-playing ability may scale with general intelligence.

In this work, we present UNSC-BENCH, a benchmark designed to evaluate LLMs’ diplomatic role-playing capabilities through the task of predicting P5 voting behavior. Specifically, our paper makes three key contributions:

1. We construct and release a multilingual dataset of 469 UNSC resolutions spanning 1947–2025, including full draft resolution text and voting records.
2. We evaluate 26 frontier and open-source LLMs across five country roles, revealing that general intelligence critically determines role-playing ability and that regional models exhibit no advantage when role-playing their home country, challenging assumptions

about cultural proximity and training-data bias. Additionally, we evaluate models under neutral prompting to measure their default alignment with P5 voting patterns without explicit country-roles in order to check their individual, default, diplomatic alignment.

3. We conduct a multilingual analysis showing that prompt language dramatically affects model predictions, with some models exhibiting major performance degradation while others remain comparatively stable across languages.

## 2 Related Work

We situate UNSC-Bench at the intersection of two research areas: specifically, LLM evaluation on diplomatic and geopolitical tasks; and more broadly, cultural knowledge and value alignment in language models.

### 2.1 LLM Evaluation on UN and Diplomatic Tasks

Two concurrent works evaluate LLMs using UN Security Council data. [Liang et al. \(2025\)](#) introduce UNBench, a four-task benchmark covering co-penholder judgment, voting simulation, draft adoption prediction, and statement generation using UNSC records from 1994–2024. They find that GPT-4o achieves 0.726 accuracy on co-penholder identification, with performance declining as choice complexity increases.

[Choi et al. \(2025\)](#) examine nation-level bias in UNSC voting prediction through explicit bias tests, persona-assigned voting, and accuracy analysis across nations. They find biases are “fundamentally multidimensional,” with GPT models showing consistent negative bias toward Russia. Their RAG-based debiasing framework reduces but does not eliminate these biases. Notably, they report that Western-developed models do not uniformly favor Western nations, partially anticipating our finding. However, neither work compares regional models (e.g., Qwen, DeepSeek) against Western models on predictions for their “home” countries, nor do they quantify the relationship between general model capability and diplomatic role-playing ability.

### 2.2 Multilingual Evaluation and Cultural Bias

Recent work has highlighted that multilingual capability does not imply cultural understanding. [Singh](#)

[et al. \(2025\)](#) demonstrate that 28% of MMLU questions require culturally sensitive knowledge, and that “progress on MMLU predominantly reflects knowledge of Western-centric cultural knowledge.” This finding motivates evaluation frameworks that disentangle models’ ability from their western cultural competence.

[Chen et al. \(2024\)](#) question whether current multilingual evaluation practices are sound, showing that translated benchmarks cannot capture language-specific knowledge and may introduce artifacts that distort assessment. They demonstrate notable differences between native-written and translated instruction data, particularly for high-performing models and in generative tasks. This methodological concern underscores the importance of evaluation tasks (such as predicting documented diplomatic positions) that do not rely on translation.

[Adilazuarda et al. \(2024\)](#) provide a comprehensive survey of approaches to measuring culture in LLMs, identifying key challenges including the conflation of language with culture and the lack of behavioral evaluations grounded in real-world outcomes. Our work tries to partially address this gap by using actual voting records as behavioral ground truth.

### 2.3 Cultural Knowledge Benchmarks

At the same time, results on whether regional models possess “home advantage” remain contested.

[Chiu et al. \(2025\)](#) introduce CulturalBench, comprising 1,227 human-verified questions across 45 global regions. They find GPT-4o (61.5%) outperforms the best Chinese model Qwen-2-72B (58.8%) even on East Asian cultural questions, concluding that “model providers based out of specific regions do not necessarily have advantages in cultural knowledge of their regions.”

In contrast, [Myung et al. \(2024\)](#) report home advantage in BLEnD, a benchmark of 52,600 everyday knowledge questions across 16 countries. Qwen1.5-72B achieves highest performance on Chinese questions, and HyperCLOVA-X outperforms GPT-4 on Korean culture. We hypothesize that the key distinction appears to be knowledge type: BLEnD tests tacit everyday knowledge (food, holidays, family customs) requiring region-specific training data, while CulturalBench tests factual knowledge available in globalized corpora. We set out to investigate this other kind of knowledge through the analysis of diplomatic position documents.

[Durmus et al. \(2023\)](#) demonstrate that LLM de-

fault responses align most closely with opinions from the United States, Canada, Australia, and Western Europe, with significant divergence from Middle Eastern, African, and Asian populations. Zhao et al. (2024) extend this using World Values Survey data across 64 countries, finding larger models substantially outperform smaller ones but without testing regional model comparisons.

## 2.4 Regional Models and Value Alignment

Furthermore, recent work challenges the assumption that regional models better represent the values of their home cultures.

Naous et al. (2024) report a pattern of this kind for Arabic: even Arabic-specific LLMs exhibit bias towards Western cultural entities. When asked in Arabic for food suggestions, the models chose “ravioli” over Arabic dishes; when asked for women’s names in Arabic, the models chose “Roseanne” over Arab names. This Western default in regional models helps explain why training country of origin may not confer a role-playing advantage.

Most directly relevant to our work, Rystrom et al. (2025) find that “the ability to communicate across languages does not necessarily translate to appropriate cultural representations.” Testing Gemma and OpenAI models against World Values Survey data across four languages, they find no consistent relationship between language capability and cultural alignment. This dissociation between linguistic and cultural competence provides theoretical grounding for our finding that Chinese models’ linguistic capabilities do not translate to superior China role-playing.

Li et al. (2024a) evaluate geopolitical bias through 251 territorial disputes across 49 languages, finding LLMs answer differently depending on query language. In particular, they report that larger models improve factual accuracy but not cross-lingual consistency and that geopolitical bias can be “amplified in stronger models,” contrasting with our finding that capability correlates with role-playing ability. This suggests our benchmark measures a distinct construct: the skill of adopting a consistent perspective rather than the default perspective a model holds.

## 2.5 Role-Playing and Cultural Prompting

The effectiveness of role-playing instructions for cultural alignment has received limited systematic study. Tao et al. (2024) test cultural prompting across 107 countries using World Values Sur-

vey questions, finding it improves alignment for 71–81% of countries. However, they do not systematically investigate capability thresholds.

## 3 UNSC-Bench: Task and Dataset

### 3.1 Data Sources and Collection

We construct UNSC-BENCH from official United Nations Security Council documentation spanning 1947–2025. The benchmark compiles data from three sources: (i) the *Corpus of Resolutions: UN Security Council* (CR-UNSC) for adopted resolutions and structured metadata (Fobbe et al., 2024); (ii) the UN Department of Political and Peacebuilding Affairs (DPPA) dataset of all Security Council vetoes since 1946 for vetoed drafts and veto metadata (United Nations Department of Political and Peacebuilding Affairs, 2025); and (iii) the UN Digital Library for resolution PDFs, draft texts, and meeting records containing official voting tallies<sup>1</sup>. For a small number of older documents, the available PDF text was incomplete or corrupted in the archived versions, so we recovered the document text using DeepSeek-OCR (Wei et al., 2025). For a small set of cases (fewer than 20) where official translations could not be retrieved, we generated translations using DeepSeek V3.2-exp (DeepSeek-AI et al., 2025) and used these to complete the multilingual inputs.

From these sources, we select resolutions that meet two criteria: (1) the full text of the proposal voted on is available in at least English, and (2) complete voting records exist for all five permanent members (P5). After cross-validating voting tallies against the Security Council Report, the final dataset contains 469 resolutions and 2,345 P5 vote instances.

Resolutions can be divided into two partitions of similar size: 211 *vetoed drafts* (45%) and 258 *adopted resolutions* (55%). Vetoed drafts are proposals blocked by at least one P5 veto and thus correspond to the most contentious issues, where permanent members fundamentally disagreed. Adopted resolutions passed without veto and typically reflect either consensus or cases where P5 members chose to abstain.

A key design choice is to use, wherever possible, the *draft text actually voted on*. For vetoed resolutions this is the only available text. For adopted resolutions, we use the draft text whenever it is

<sup>1</sup><https://digitallibrary.un.org/>

available, and exclude resolutions without an available draft. This approximates the realistic setting where a model must assess the draft resolution before knowing the outcome.

UN Security Council documents are officially translated into six languages; in UNSC-BENCH we include four: English (EN), Chinese (ZH), French (FR), and Russian (RU), representative of the P5 members.

### 3.2 Task Definition

Each instance  $i$  of our benchmark contains two components: a draft resolution text  $r_i^{(\ell)}$  written in language  $\ell \in \{\text{EN, ZH, FR, RU}\}$ —the main languages of P5 countries—and the corresponding ground-truth votes  $v_{i,c} \in \{\text{YES, NO, ABSTAIN}\}$  cast by each of the five permanent members (P5) of the Security Council, where  $c \in \mathcal{C} = \{\text{CN, FR, RU, UK, US}\}$ .

We evaluate model behavior under two distinct prompting settings:

**Role-Play Setting:** In this configuration, we explicitly instruct the model to adopt the perspective of a specific P5 country  $c$ . The model then predicts how that country would vote on the given resolution. Formally, this mapping can be expressed as:

$$(r_i^{(\ell)}, c, \ell) \mapsto \hat{v}_{i,c}^{(\ell)} \quad (1)$$

**Neutral Setting:** In this configuration, the model is not conditioned to behave as any country. The model simply receives the draft resolution text and produces a vote as output:

$$(r_i^{(\ell)}, \ell) \mapsto \hat{v}_{i,\text{neutral}}^{(\ell)} \quad (2)$$

We then compare this neutral prediction  $\hat{v}_{i,\text{neutral}}^{(\ell)}$  against the actual votes of all five P5 members. This allows us to measure which country’s voting pattern the model most closely aligns with from its own perspective. We characterize this as a proxy for how the model would behave in normal use, when asked to complete a task as an assistant chatbot without role-play instructions.

**Multilingual Evaluation:** Unless otherwise specified, all experiments were conducted in English ( $\ell = \text{EN}$ ). However, to investigate whether a model’s default alignment shifts depending on the input language, we repeated both the role-play and neutral evaluations in Chinese, French, and Russian ( $\ell \in \{\text{ZH, FR, RU}\}$ ) for a subset of models.

Country	YES	NO	ABS	Total	NO%
China	348	21	100	469	4.5
France	374	17	78	469	3.6
Russia	279	106	84	469	22.6
UK	351	30	88	469	6.4
US	312	91	66	469	19.4

Table 1: Ground-truth vote counts for the five permanent members (P5) across the 469 resolutions in UNSC-BENCH. “ABS” = abstain. “Russia” aggregates USSR (pre-1991) and Russian Federation (post-1991). “NO%” is the share of NO votes.

### 3.3 Dataset Statistics

Table 1 summarizes ground-truth voting patterns for each P5 member in UNSC-BENCH. Votes are imbalanced: all P5 members vote YES much more frequently than NO or ABSTAIN, reflecting the consensus-seeking nature of the Security Council.

Across countries, YES votes account for roughly 60–80% of decisions, while NO votes are rare except for Russia and the United States. Abstentions serve as another option, allowing states to note reservations without blocking a resolution.

## 4 Experimental Setup

### 4.1 Models

We evaluated 26 base LLMs from four of the five P5 countries: the US (GPT-5 (OpenAI, 2025), Claude Sonnet 4.5 (Anthropic, 2025), Llama (Dubey et al., 2024)), China (Qwen (Yang et al., 2025), DeepSeek (DeepSeek-AI et al., 2025), Kimi (Kimi Team et al., 2025)), France (the Mistral family (Jiang et al., 2023, 2024)), and Russia (GigaChat (GigaChat Team et al., 2025), Vistral from the Vikhr model family (Nikolich et al., 2024)).

For Claude, DeepSeek, and Kimi, we evaluate both standard and thinking variants to test whether step-by-step reasoning helps diplomatic simulation. We then compared those performances against MMLU-Pro scores (Wang et al., 2024b)<sup>2</sup>; by adopting this as a capability proxy, we attempt to separate the effects of raw intelligence from geographic origin. Full model details appear in Appendix B.

### 4.2 Prompting Conditions

As previously stated in Section 3.2, we evaluate each model under two conditions to isolate the effect of explicit role assignment.

<sup>2</sup>MMLU-Pro scores were sourced from Artificial Analysis (<https://artificialanalysis.ai/evaluations/mmlu-pro>)

Model	YES%	Neutral Prompting (Macro F1)				
		CN	FR	RU	UK	US
<i>Closed-Source</i>						
GPT-5	95%	0.27	<b>0.36</b>	0.25	0.33	0.28
Claude Sonnet 4.5	85%	0.35	0.37	0.34	<b>0.41</b>	0.32
Claude Sonnet 4.5 [thinking]	83%	0.34	0.33	0.32	<b>0.36</b>	0.35
<i>Open-Source</i>						
DeepSeek V3.2-exp	99%	0.28	<b>0.29</b>	0.26	0.28	0.26
DeepSeek V3.2-exp [thinking]	99%	0.28	<b>0.29</b>	0.25	0.28	<b>0.29</b>
Kimi K2 [instruct]	70%	0.35	0.35	<b>0.37</b>	0.36	0.36
Kimi K2 [thinking]	87%	0.33	0.33	<b>0.36</b>	0.35	0.32

Table 2: Macro F1 for P5 vote prediction under neutral prompting on UNSC-Bench (English prompts). YES% is the neutral prediction rate. Bold indicates best per model.

Neutral prompting provides only the resolution text. The model predicts P5 votes without being told which country’s perspective to adopt. This tests whether models have any internal diplomatic bias and which country they align with the most by default.

Role-playing prompting assigns a country-specific role, instructing the model to reason from that country’s geopolitical interests and historical positions.

For multilingual analysis, we tested four models with the same prompts translated in English, Chinese, French, and Russian. The translated prompts were populated with the translation of the resolution text in the specific language. These configurations let us test whether prompting in a country’s native language helps model predict that country’s votes. For prompt templates, refer to Appendix D.

### 4.3 Evaluation Metrics

We report Macro F1 and per-class F1 scores. The per-class breakdown is important given class imbalance: most resolutions pass, so high YES-class F1 is relatively easy to achieve. The more informative signal lies in minority classes: correctly predicting when a country will vote NO or ABSTAIN suggests the model captures something about that country’s diplomatic boundaries, rather than simply learning overall class distribution.

For multilingual evaluation, we measure cross-language consistency: how often a model produces identical predictions across prompt languages. High consistency suggests robust reasoning that generalizes across languages. However, low consistency is not necessarily a failure: a model might reason more accurately about Russian diplomacy when prompted in Russian, perhaps due to

more relevant training data in that language. In such cases, inconsistency reflects language-specific competence rather than instability.

## 5 Results

We organize our findings around four questions: (i) Are models diplomatically aligned without explicit roles? (ii) How does general capability relate to diplomatic role-playing? (iii) Do regional models enjoy a home-country advantage? (iv) How sensitive are predictions to prompt language?

### 5.1 Raw Models Are Diplomatically Unaligned

Our first experiment examines whether models inherently align with the voting patterns of specific countries when not explicitly instructed to adopt any role. Table 2 presents neutral prompting results across frontier models.

Under neutral prompting, models exhibit no consistent diplomatic alignment. Instead, they default to a high YES rate (70–99%), effectively behaving as agreeable but generic delegates. Macro F1 scores remain low and vary only narrowly across the five P5 countries within each model, indicating that models do not reliably recover distinct national positions from draft resolution text alone. Notably, the country with the highest alignment differs across models; GPT-5 aligns most with France, Claude Sonnet 4.5 with the UK, and Kimi K2 with Russia, suggesting these slight variations reflect noise rather than systematic diplomatic orientation.

These results suggest that without explicit role assignment, frontier models lack inherent diplomatic alignment. They do not consistently match any country’s viewpoint and do not automatically vote like a specific P5 member, instead defaulting

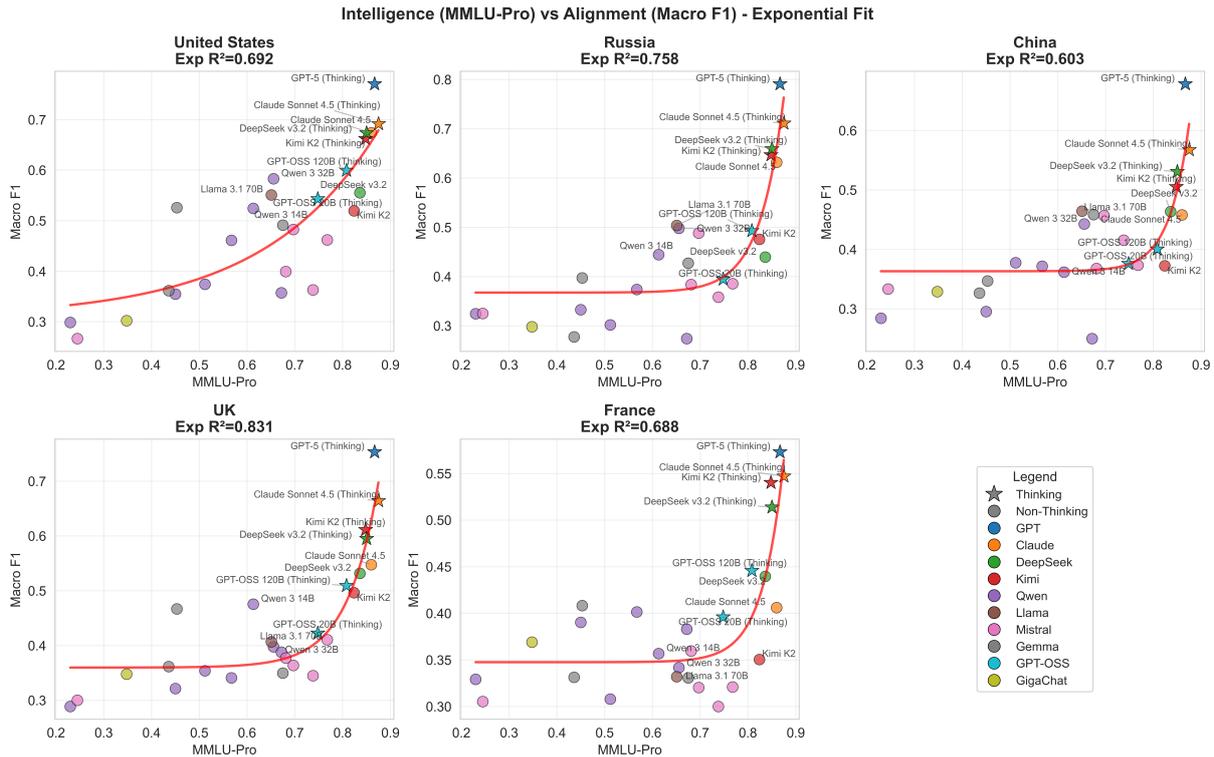


Figure 1: General capability (MMLU-Pro; x-axis) vs. role-play performance (Macro F1; y-axis) for each P5 country role on UNSC-BENCH (English prompts). Each point is a model; stars denote thinking variants. Curves show exponential fits;  $R^2$  values are reported per role.

to consensus-seeking behavior that does not fit the voting record of any individual P5 member.

## 5.2 Diplomatic Role-Playing Scales With General Model Capability

Figure 1 relates general capability (MMLU-Pro) to role-play performance (Macro F1), shown separately for each P5 role. Each point corresponds to a model configuration (including base and thinking variants), and the curve shows an exponential fit. Across all five roles, the main pattern is non-linear: many lower- and mid-capability systems fall into a relatively narrow performance band, while the strongest models separate more clearly and achieve substantially higher Macro F1. This separation is especially clear for the UK and Russia roles, where performance increases rapidly at the upper end of capability, while China shows more variance among similarly capable models, indicating that simulating China is less predictable from general capability alone.

We first compute Pearson correlations to confirm that a positive relationship exists for all roles ( $r = 0.55\text{--}0.80$ , all  $p < 0.01$ ). Because the trend is visibly non-linear, flat among weaker models, then rising steeply at the top, we also fit an exponential curve

of the form  $y = a \cdot \exp(bx) + c$ . The exponential model achieves higher  $R^2$  than a linear fit for all five countries (e.g., UK: 0.83 vs 0.54; Russia: 0.76 vs 0.51), consistent with a capability threshold effect where diplomatic role-playing ability emerges more strongly beyond a certain level of general intelligence.

The figure also suggests that additional inference-time computation often helps in this task. Thinking variants (stars) tend to appear among the highest-scoring points and frequently sit above nearby non-thinking counterparts from the same family. Overall, Figure 1 supports the interpretation that effective diplomatic role-playing is closely tied to general capability, with the most visible gains concentrated among the highest-performing systems.

## 5.3 Regional Models Do Not Have a Home-Country Advantage

We might expect models to perform best when predicting the votes of their home country, since their training data may include more text from that country’s media and public sources. Table 3 tests this hypothesis by showing Macro F1 scores for each model on all five P5 countries, with the models’ home countries highlighted.

Model	Macro F1				
	CN	FR	RU	UK	US
<i>US Models — Home country:</i> <span style="border: 1px solid black; padding: 0 2px;">US</span>					
GPT-5	0.68	0.57	<b>0.79</b>	0.75	<span style="border: 1px solid black; padding: 0 2px;">0.77</span>
Claude Sonnet 4.5	0.46	0.41	0.63	0.55	<span style="border: 1px solid black; padding: 0 2px;"><b>0.67</b></span>
Claude Sonnet 4.5 [thinking]	0.57	0.55	<b>0.71</b>	0.66	<span style="border: 1px solid black; padding: 0 2px;">0.69</span>
Llama 3.1-70B	0.46	0.33	0.50	0.41	<span style="border: 1px solid black; padding: 0 2px;"><b>0.55</b></span>
GPT-OSS-120B	0.40	0.45	0.49	0.51	<span style="border: 1px solid black; padding: 0 2px;"><b>0.60</b></span>
<i>Chinese Models — Home country:</i> <span style="border: 1px solid black; padding: 0 2px;">CN</span>					
Qwen3-32B	<span style="border: 1px solid black; padding: 0 2px;">0.44</span>	0.34	0.50	0.40	<b>0.58</b>
DeepSeek V3.2-exp	<span style="border: 1px solid black; padding: 0 2px;">0.46</span>	0.44	0.44	0.53	<b>0.56</b>
DeepSeek V3.2-exp [thinking]	<span style="border: 1px solid black; padding: 0 2px;">0.53</span>	0.51	0.66	0.60	<b>0.67</b>
Kimi K2 [instruct]	<span style="border: 1px solid black; padding: 0 2px;">0.37</span>	0.35	0.48	0.50	<b>0.52</b>
Kimi K2 [thinking]	<span style="border: 1px solid black; padding: 0 2px;">0.51</span>	0.54	0.65	0.61	<b>0.66</b>
<i>French Models — Home country:</i> <span style="border: 1px solid black; padding: 0 2px;">FR</span>					
Mistral Large	0.46	<span style="border: 1px solid black; padding: 0 2px;">0.32</span>	<b>0.49</b>	0.36	0.48
Mistral Small 24B	0.37	<span style="border: 1px solid black; padding: 0 2px;">0.36</span>	0.38	0.38	<b>0.40</b>
Mistral Nemo 12B	<b>0.42</b>	<span style="border: 1px solid black; padding: 0 2px;">0.30</span>	0.36	0.34	0.36
Ministral 8B	0.32	<span style="border: 1px solid black; padding: 0 2px;">0.34</span>	0.34	0.38	<b>0.42</b>
Mistral 7B-v0.3	<b>0.33</b>	<span style="border: 1px solid black; padding: 0 2px;">0.31</span>	0.32	0.30	0.27
Magistral Small	0.37	<span style="border: 1px solid black; padding: 0 2px;">0.32</span>	0.39	0.41	<b>0.46</b>
<i>Russian Models — Home country:</i> <span style="border: 1px solid black; padding: 0 2px;">RU</span>					
GigaChat-20B-A3B	0.33	<b>0.37</b>	<span style="border: 1px solid black; padding: 0 2px;">0.30</span>	0.35	0.30
Vistral-24B	0.39	0.36	<span style="border: 1px solid black; padding: 0 2px;">0.35</span>	0.42	<b>0.43</b>

Table 3: Role-play Macro F1 by P5 country role. Boxed = home country; **bold** = best score. Most models perform best on non-home countries.

Instead, we do not observe a consistent home-country advantage in Table 3. Within the US group, models generally score strongly on the US role, but the US role is not always the single best (e.g., GPT-5 is slightly higher on Russia than on the US). The clearest deviation from the home-advantage expectation appears for the Chinese models: all five have their highest Macro F1 on the US role rather than the China role. For instance, Qwen3-32B, DeepSeek V3.2-exp [thinking], and Kimi K2 [thinking] each score higher on the US column than on the China column, indicating that Chinese origin does not necessarily translate into superior China-role prediction on this benchmark.

French and Russian models show a similar pattern. None of the French models achieve their best score on France; each performs better on a non-home role, most often the United States, Russia, or China (e.g., Mistral Large peaks on Russia rather than France). The two Russian models also peak

outside their home role: GigaChat-20B-A3B performs best on France, and Vistral-24B performs best on the United States. Taken together, these results suggest that diplomatic role-play performance is not reliably predicted by a model’s geographic origin.

Finally, the China column illustrates the same point: the highest China-role score in Table 3 is achieved by GPT-5, exceeding the best China-role score among the listed Chinese models. Overall, Table 3 seems to indicate that geographical home region is secondary to other factors when evaluating performance for this task.

#### 5.4 Prompt Language Can Dramatically Distort Predictions

Finally, we examine whether prompt language changes model predictions. We evaluate Claude Sonnet 4.5, DeepSeek V3.2-exp (and its thinking variant), and Kimi K2 (instruct and thinking) across

Model	NO F1					ABSTAIN F1					Macro F1				
	EN	ZH	FR	RU	$\Delta$	EN	ZH	FR	RU	$\Delta$	EN	ZH	FR	RU	$\Delta$
DeepSeek V3.2-exp	0.67	0.64	0.70	0.68	0.06	0.19	0.14	0.12	0.32	0.20	0.57	0.54	0.56	0.62	0.08
DeepSeek V3.2-exp [thinking]	0.67	0.68	0.73	0.72	0.06	0.28	0.18	0.24	0.29	0.11	0.60	0.57	0.61	0.63	0.06
Claude Sonnet 4.5	0.77	0.55	0.78	0.72	0.23	0.37	0.17	0.38	0.23	0.21	0.67	0.52	0.69	0.61	0.17
Kimi K2 [thinking]	0.76	0.47	0.77	0.32	0.45	0.34	0.26	0.30	0.23	0.11	0.67	0.46	0.65	0.39	0.28
Kimi K2 [instruct]	0.61	0.30	0.69	<b>0.00</b>	<b>0.69</b>	0.27	0.21	0.24	0.19	0.08	0.58	0.35	0.60	0.27	0.33

Table 4: F1 scores for predicting US votes across prompt languages.  $\Delta = \max - \min$  (range).

four languages (English, Chinese, French, and Russian). Table 4 presents NO F1 scores for the US role, where the effect is most pronounced.

Table 4 shows that language sensitivity varies substantially across models. DeepSeek V3.2-exp is comparatively stable across languages on the NO class, with a small range, and its thinking variant shows similar stability with slightly higher scores. In contrast, Claude Sonnet 4.5 exhibits a larger drop in Chinese relative to English and French, and its performance varies more across prompt languages than DeepSeek on all three metrics.

The largest variation is observed for Kimi K2 [instruct]. For the US role, its NO F1 ranges from 0.00 under Russian prompts to 0.69 under French prompts ( $\Delta = 0.69$ ), indicating a substantial degradation in identifying NO outcomes in Russian. In contrast, variation in the ABSTAIN class is comparatively small (0.19–0.27;  $\Delta = 0.08$ ), suggesting that the language sensitivity is concentrated in NO prediction rather than reflecting a uniform shift across all classes. The thinking variant partially mitigates this effect: Kimi K2 [thinking] increases NO F1 in Russian to 0.32, although it remains below the model’s English and French performance.

Overall, these results indicate that multilingual prompting can meaningfully change performance in this diplomatic modeling task, and that this effect is strongly model-dependent. For multilingual deployments of diplomatic role-play, it is therefore important to evaluate not only average performance but also cross-language stability, particularly for the minority vote types (NO and ABSTAIN), where we observe the largest performance differences across languages.

## 6 Conclusion

We introduced UNSC-BENCH, a benchmark for evaluating LLM diplomatic role-playing through UN Security Council P5 vote prediction. The dataset covers 469 draft resolutions spanning 1947–2025 with ground-truth voting records, and provides

multilingual draft resolutions in English, Chinese, French, and Russian. We evaluated 26 models (29 configurations including thinking variants) under both neutral prompting and explicit country role-play. Under neutral prompting, models are diplomatically unaligned, defaulting to high YES rates and failing to match any P5 voting pattern. When engaging in diplomatic role-play, performances are strongly tied to general model capability, with thinking variants providing further improvements. Notably, regional origin does not confer an advantage: models developed in a P5 country do not predict their home country’s votes more accurately than other roles. Finally, prompt language affects predictions, particularly for minority outcomes like NO and ABSTAIN, though the magnitude varies across models.

## 7 Limitations

Our benchmark evaluates diplomatic role-playing through vote prediction, which cannot capture the reasoning behind a decision or the negotiations that precede it. Additionally, the benchmark covers only P5 members and resolutions with available text and voting records, and models see only the draft resolution itself, without access to past votes, stated policy positions, or surrounding debate that real diplomats would draw on. We also explore a limited prompting setup: all experiments are zero-shot with a single template per condition, and few-shot examples or chain-of-thought prompting might improve results. On the multilingual side, translation quality and prompt wording may affect predictions, so language effects could reflect linguistic factors beyond country perspective. Finally, because historical votes are likely present in training data, we cannot fully rule out memorization, though our results suggest models do not simply recall ground-truth outcomes.

## References

- Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Shivdutt Singh, Alham Fikri Aji, Jacki O’Neill, Ashutosh Modi, and Monojit Choudhury. 2024. [Towards measuring and modeling “culture” in LLMs: A survey](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15763–15784, Miami, Florida, USA. Association for Computational Linguistics.
- Gati V Aher, Rosa I. Arriaga, and Adam Tauman Kalai. 2023. [Using large language models to simulate multiple humans and replicate human subject studies](#). In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 337–371. PMLR.
- Anthropic. 2025. [Claude sonnet 4.5 system card](#). Technical Report.
- Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. [A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity](#). In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–718, Nusa Dua, Bali. Association for Computational Linguistics.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, et al. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.
- Pinzhen Chen, Simon Yu, Zhicheng Guo, and Barry Haddow. 2024. [Is it good data for multilingual instruction tuning or just bad multilingual evaluation for large language models?](#) In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9706–9726, Miami, Florida, USA. Association for Computational Linguistics.
- Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, and Yejin Choi. 2025. [CulturalBench: A robust, diverse and challenging benchmark for measuring LMs’ cultural knowledge through human-AI red-teaming](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25663–25701, Vienna, Austria. Association for Computational Linguistics.
- Jonghyeon Choi, Yeonjun Choi, Hyun chul Kim, and Beakcheol Jang. 2025. [“As Eastern Powers, I Will Veto”: An investigation of nation-level bias of large language models in international relations](#). *Preprint*, arXiv:2511.10695.
- DeepSeek-AI, Aixiu Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenhao Xu, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, et al. 2025. [DeepSeek-v3.2: Pushing the frontier of open large language models](#). *Preprint*, arXiv:2512.02556.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. [The Llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- Esin Durmus, Karina Nyugen, Thomas I. Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. 2023. [Towards measuring the representation of subjective global opinions in language models](#). *Preprint*, arXiv:2306.16388.
- Seán Fobbe, Lorenzo Gasbarri, and Niccolò Ridi. 2024. [Corpus of resolutions: Un security council \(cr-unsc\)](#).
- GigaChat Team, Valentin Mamedov, et al. 2025. [GigaChat family: Efficient Russian language modeling through mixture of experts architecture](#). *Preprint*, arXiv:2506.09440.
- Tiancheng Hu and Nigel Collier. 2024. [Quantifying the persona effect in LLM simulations](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10289–10307, Bangkok, Thailand. Association for Computational Linguistics.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. [Mistral 7B](#). *Preprint*, arXiv:2310.06825.
- Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. [Mixtral of experts](#). *Preprint*, arXiv:2401.04088.
- Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. [Scaling laws for neural language models](#). *Preprint*, arXiv:2001.08361.
- Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, et al. 2025.

- Kimi k2: Open agentic intelligence. *Preprint*, arXiv:2507.20534.
- Bryan Li, Samar Haider, and Chris Callison-Burch. 2024a. **This land is Your, My land: Evaluating geopolitical bias in language models through territorial disputes**. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3855–3871, Mexico City, Mexico. Association for Computational Linguistics.
- Junyi Li, Charith Peris, Ninareh Mehrabi, Palash Goyal, Kai-Wei Chang, Aram Galstyan, Richard Zemel, and Rahul Gupta. 2024b. **The steerability of large language models toward data-driven personas**. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7290–7305, Mexico City, Mexico. Association for Computational Linguistics.
- Yueqing Liang, Liangwei Yang, Chen Wang, Congying Xia, Rui Meng, Xiong Xiao Xu, Haoran Wang, Ali Payani, and Kai Shu. 2025. **Benchmarking LLMs for political science: A United Nations perspective**. *Preprint*, arXiv:2502.14122.
- Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Afina Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, Víctor Gutiérrez-Basulto, Yazmín Ibáñez García, Hwaran Lee, Shamsuddeen Hassan Muhammad, Kiwoong Park, Anar Sabuhi Rzayev, Nina White, Seid Muhie Yimam, Mohammad Taher Pilehvar, et al. 2024. **Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages**. In *Advances in Neural Information Processing Systems*, volume 37, pages 78104–78146. Curran Associates, Inc.
- Tarek Naous, Michael J Ryan, Alan Ritter, and Wei Xu. 2024. **Having beer after prayer? measuring cultural bias in large language models**. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16366–16393, Bangkok, Thailand. Association for Computational Linguistics.
- Aleksandr Nikolich, Konstantin Korolev, Artem Shelmanov, and Igor Yaroslavtsev. 2024. **Vikhr: The family of open-source instruction-tuned large language models for Russian**. *Preprint*, arXiv:2405.13929.
- OpenAI. 2025. **GPT-5 system card**. Technical Report.
- Weihong Qi, Hanjia Lyu, and Jiebo Luo. 2025. **Representation bias in political sample simulations with large language models**. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 1264–1267, Sydney, NSW, Australia. Association for Computing Machinery.
- Juan-Pablo Rivera, Gabriel Mukobi, Anka Reuel, Max Lamparth, Chandler Smith, and Jacquelyn Schneider. 2024. **Escalation risks from language models in military and diplomatic decision-making**. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*, Rio de Janeiro, Brazil. Association for Computing Machinery.
- Jonathan Rystrom, Hannah Rose Kirk, and Scott A. Hale. 2025. **Multilingual != multicultural: Evaluating gaps between multilingual capabilities and cultural alignment in LLMs**. In *Proceedings of the First Interdisciplinary Workshop on Observations of Misunderstood, Misguided and Malicious Use of Language Models (OMMM 2025)*, associated with the International Conference RANLP 2025, pages 74–85, Varna, Bulgaria.
- Mikhail Salnikov, Dmitrii Korzh, Ivan Lazichny, Elvir Karimov, Artyom Iudin, Ivan Oseledets, Oleg Y. Rogov, Alexander Panchenko, Natalia Loukachevitch, and Elena Tutubalina. 2025. **Geopolitical biases in LLMs: what are the “good” and the “bad” countries according to contemporary language models**. *Preprint*, arXiv:2506.06751.
- Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Sebastian Ruder, Wei-Yin Ko, Antoine Bosselut, Alice Oh, Andre Martins, Leshem Choshen, Daphne Ippolito, et al. 2025. **Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation**. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18761–18799, Vienna, Austria. Association for Computational Linguistics.
- Zhi Rui Tam, Cheng-Kuang Wu, Yu Ying Chiu, Chieh-Yen Lin, Yun-Nung Chen, and Hung-yi Lee. 2025. **Language matters: How do multilingual input and reasoning paths affect large reasoning models?** *Preprint*, arXiv:2505.17407.
- Yan Tao, Olga Viberg, Ryan S Baker, and René F Kizilcec. 2024. **Cultural bias and cultural alignment of large language models**. *PNAS Nexus*, 3(9).
- United Nations Department of Political and Peacebuilding Affairs. 2025. Security Council Data: Vetoes Since 1946. <https://psdata.un.org/dataset/DPPA-SCVETOES>. Open dataset of all vetoes cast by permanent members of the UN Security Council.
- Siqi Wang, Zhengyu Chen, Bei Li, Keqing He, Min Zhang, and Jingang Wang. 2024a. **Scaling laws across model architectures: A comparative analysis of dense and MoE models in large language models**. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5583–5595, Miami, Florida, USA. Association for Computational Linguistics.
- Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren,

Aaran Arulraj, Xuan He, Ziyang Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhui Chen. 2024b. [MMLU-pro: A more robust and challenging multi-task language understanding benchmark](#). In *Advances in Neural Information Processing Systems 37*, pages 95266–95290, Vancouver, Canada.

Haoran Wei, Yaofeng Sun, and Yukun Li. 2025. [DeepSeek-OCR: Contexts optical compression](#). Preprint, arXiv:2510.18234.

An Yang, Anfeng Yang, Baosong Yang, Beichen Hui, Bo Zheng, Bowen Chang, Chengen Peng, Chengpeng Chen, Dayiheng Li, Haoran Huang, et al. 2025. [Qwen3 technical report](#). Preprint, arXiv:2505.09388.

Wenlong Zhao, Debanjan Mondal, Niket Tandon, Danica Dillion, Kurt Gray, and Yuling Gu. 2024. [World-ValuesBench: A large-scale benchmark dataset for multi-cultural value awareness of language models](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17696–17706, Torino, Italia. ELRA and ICCL.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. [Can large language models transform computational social science?](#) *Computational Linguistics*, 50(1):237–291.

## Appendix A. Dataset Statistics

The UNSC-BENCH dataset is available at <https://huggingface.co/datasets/UNSC-Bench/UNSC-Bench>.

### A.1 License

UNSC-BENCH is released under CC0-1.0.

### A.2 Temporal Distribution

Table 5 reports the number of adopted resolutions and vetoed drafts by decade.

Decade	Adopted	Vetoed	Veto %
1940s	12	17	58.6
1950s	16	29	64.4
1960s	23	16	41.0
1970s	26	32	55.2
1980s	27	48	64.0
1990s	43	10	18.9
2000s	44	14	24.1
2010s	45	21	31.8
2020s	22	24	52.2
<b>Total</b>	<b>258</b>	<b>211</b>	<b>45.0</b>

Table 5: Counts of adopted resolutions and vetoed drafts in UNSC-BENCH by decade (1947–2025).

### A.3 P5 Veto Patterns

Table 6 shows how often each permanent member vetoed a draft in our collection. Since multiple P5 members can veto the same draft, these counts reflect veto occurrences by country rather than distinct vetoed resolutions.

Country	Vetoes	%	Per Decade
Russia/USSR	106	40.0	13.3
United States	91	34.3	11.4
United Kingdom	30	11.3	3.8
China	21	7.9	2.6
France	17	6.4	2.1

Table 6: Veto frequency by P5 country. Multiple countries may veto the same draft.

### A.4 Class Imbalance Analysis

Table 7 reports the distribution of YES, NO, and ABSTAIN votes for each P5 member, showing substantial class imbalance across countries. YES is the most common outcome for all members; for example, France votes YES on nearly 80% of items, making NO prediction especially challenging. By contrast, Russia has the most balanced distribution (59.5% YES), which helps explain why models typically achieve higher NO F1 for Russia than for other countries. This imbalance also affects evaluation: a model can obtain high overall accuracy by over-predicting YES, whereas strong performance on NO and ABSTAIN requires correctly identifying the small number of NO and ABSTAIN votes.

## Appendix B. Complete Model Results

Table 8 presents Macro F1 scores for all 26 LLMs, along with the thinking variants across the five P5 country roles, ranked by overall performance.

Country	YES	NO	ABS	YES%
France	374	17	78	79.7
UK	351	30	88	74.8
China	348	21	100	74.2
US	312	91	66	66.5
Russia	279	106	84	59.5

Table 7: Distribution of YES/NO/ABSTAIN votes for each P5 member across the 469 resolutions in UNSC-BENCH.

## Appendix C. Per-Class Performance Analysis

### C.1 Capability and Minority-Class Prediction

Table 9 breaks Macro F1 into per-class F1 scores for a subset of representative models, showing that most gains come from the minority outcomes. Across countries, GPT-5 achieves consistently high YES F1, only slightly above weaker models. In contrast, its NO F1 ranges from 0.40 (France) to 0.79 (Russia), substantially higher than Kimi K2’s NO F1 (0.08 – 0.48). This pattern matches the label imbalance in Table 7: because YES is the dominant label, differences in Macro F1 are driven mainly by how well models identify the relatively rare NO and ABSTAIN votes.

### C.2 Country Difficulty Ranking

Averaging across all models, we observe a consistent difficulty ordering: Russia (easiest) > US > UK > China > France (hardest). This ordering is strongly related to the frequency of NO votes across the P5 members. Russia has 106 NO votes in our dataset, whereas France has 17, which is 3.6% of French votes. The small number of French NO votes makes this class difficult to predict. In Table 9, even GPT-5 reaches only 0.40 NO F1 for France, and Kimi K2 reaches 0.08. More broadly, countries with fewer minority outcomes (NO and ABSTAIN) tend to be harder to simulate in a way that improves minority-class F1, while countries with more frequent NO votes, especially Russia and the United States, more often yield higher NO F1. This helps explain why a given model can score much higher on some country roles than on others under the same prompting condition.

## Appendix D. Prompt Templates

This section presents the prompt templates used in our experiments. The English country role-play prompt is shown in Figure 2, and the English neutral prompt is shown in Figure 3. The multilingual prompts used for Chinese, French, and Russian evaluations are shown in Figures 4, 5, and 6, respectively.

## Appendix E. Example Input and Output

Tables 10 and 11 show example resolutions from the two dataset partitions. Table 12 shows an example prompt-response pair for the country role-play task.

### Prompt 1: Country Role-Play (English)

**System:** You are the representative of {country} in the United Nations Security Council. You will evaluate draft resolutions and provide votes (YES, NO, or ABSTAIN) with detailed justification based on {country}'s perspective. Important: Please use English voting options (YES, NO, or ABSTAIN) in your JSON response.

**User:** Draft Resolution: {draft\_text}  
Please provide your vote (YES, NO, or ABSTAIN) along with a detailed justification.

Figure 2: English prompt template for country role-play experiments.

### Prompt 2: Neutral Baseline

**System:** You are evaluating United Nations Security Council draft resolutions. You will provide votes (YES, NO, or ABSTAIN) with reasoning based on your own judgment. Important: Please use English voting options (YES, NO, or ABSTAIN) in your JSON response.

**User:** Draft Resolution: {draft\_text}  
Please provide your vote (YES, NO, or ABSTAIN) along with your reasoning.

Figure 3: Neutral prompt template without country role instructions.

### Prompt 3: Chinese Language (ZH)

**System:** 您是联合国安全理事会中{country}的代表。您将评估决议草案，并根据{country}的立场提供投票（赞成、反对或弃权）及详细理由。  
重要：请在 JSON 响应中使用中文投票选项（赞成、反对或弃权）。

**User:** 决议草案：{draft\_text}  
请提供您的投票（赞成、反对或弃权）及详细理由。

Figure 4: Chinese prompt template for multilingual experiments.

### Prompt 4: French Language (FR)

**System:** Vous êtes le représentant de {country} au Conseil de sécurité des Nations Unies. Vous évalueriez les projets de résolution et fournirez des votes (POUR, CONTRE ou ABSTENTION) avec une justification détaillée basée sur la perspective de {country}.

Important : Veuillez utiliser les options de vote en français (POUR, CONTRE ou ABSTENTION) dans votre réponse JSON.

**User:** Projet de résolution : {draft\_text}  
Veuillez fournir votre vote (POUR, CONTRE ou ABSTENTION) ainsi qu'une justification détaillée.

Figure 5: French prompt template for multilingual experiments.

### Prompt 5: Russian Language (RU)

**System:** Вы являетесь представителем {country} в Совете Безопасности Организации Объединённых Наций. Вы будете оценивать проекты резолюций и предоставлять голоса (ЗА, ПРОТИВ или ВОЗДЕРЖАЛСЯ) с подробным обоснованием с точки зрения {country}.

Важно: Пожалуйста, используйте варианты голосования на русском языке (ЗА, ПРОТИВ или ВОЗДЕРЖАЛСЯ) в вашем JSON-ответе.

**User:** Проект резолюции: {draft\_text}  
Пожалуйста, предоставьте ваш голос (ЗА, ПРОТИВ или ВОЗДЕРЖАЛСЯ) вместе с подробным обоснованием.

Figure 6: Russian prompt template for multilingual experiments.

Rank	Model	US	RU	CN	UK	FR
1	GPT-5	0.77	0.79	0.68	0.75	0.57
2	Claude Sonnet 4.5 [thinking]	0.69	0.71	0.57	0.66	0.55
3	DeepSeek V3.2-exp [thinking]	0.67	0.66	0.53	0.60	0.51
4	Kimi K2 [thinking]	0.66	0.65	0.51	0.61	0.54
5	Claude Sonnet 4.5	0.67	0.63	0.46	0.55	0.41
6	GPT-OSS-120B	0.60	0.49	0.40	0.51	0.45
7	DeepSeek V3.2-exp	0.56	0.44	0.46	0.53	0.44
8	Qwen3-32B	0.58	0.50	0.44	0.40	0.34
9	Llama-3.1-70B	0.55	0.50	0.46	0.41	0.33
10	Kimi K2	0.52	0.48	0.37	0.50	0.35
11	Qwen3-14B	0.52	0.44	0.36	0.48	0.36
12	Gemma-3-12B	0.53	0.40	0.35	0.47	0.41
13	GPT-OSS-20B	0.54	0.39	0.38	0.42	0.40
14	Mistral-Large	0.48	0.49	0.46	0.36	0.32
15	Gemma-3-27B	0.49	0.43	0.46	0.35	0.33
16	Magistral-Small	0.46	0.39	0.37	0.41	0.32
17	Vistral-24B	0.43	0.35	0.39	0.42	0.36
18	Qwen3-8B	0.46	0.37	0.37	0.34	0.40
19	Mistral-Small-24B	0.40	0.38	0.37	0.38	0.36
20	Ministral-8B	0.42	0.34	0.32	0.38	0.34
21	Mistral-Nemo	0.36	0.36	0.42	0.34	0.30
22	Qwen2.5-14B	0.37	0.30	0.38	0.35	0.31
23	Qwen2.5-7B	0.35	0.33	0.30	0.32	0.39
24	Gemma-3-4B	0.36	0.28	0.33	0.36	0.33
25	Qwen3-4B	0.36	0.27	0.25	0.39	0.38
26	GigaChat-20B-A3B	0.30	0.30	0.33	0.35	0.37
27	Qwen3-1.7B	0.30	0.38	0.30	0.31	0.35
28	Mistral-7B-v0.3	0.27	0.32	0.33	0.30	0.31
29	Qwen3-0.6B	0.30	0.32	0.28	0.29	0.33

Table 8: Macro F1 scores for country role-playing across all 26 models (29 configurations including thinking variants), ranked by overall performance. Each model was prompted to role-play as a P5 country representative.

Model	US			RU			CN			UK			FR		
	Y	N	A	Y	N	A	Y	N	A	Y	N	A	Y	N	A
GPT-5	0.90	0.71	0.70	0.90	0.79	0.68	0.87	0.65	0.52	0.91	0.65	0.69	0.92	0.40	0.39
Claude Sonnet 4.5 [T]	0.88	0.64	0.54	0.86	0.73	0.55	0.83	0.45	0.42	0.87	0.58	0.54	0.89	0.38	0.37
DeepSeek V3.2-exp [T]	0.87	0.63	0.52	0.83	0.67	0.48	0.83	0.41	0.36	0.86	0.48	0.46	0.90	0.29	0.35
Kimi K2 [T]	0.85	0.60	0.53	0.82	0.66	0.46	0.79	0.37	0.36	0.84	0.50	0.50	0.88	0.36	0.39
Claude Sonnet 4.5	0.84	0.63	0.54	0.81	0.62	0.46	0.72	0.28	0.38	0.81	0.40	0.43	0.87	0.18	0.18
DeepSeek V3.2-exp	0.81	0.54	0.33	0.72	0.43	0.16	0.76	0.35	0.28	0.82	0.44	0.33	0.88	0.24	0.21
Kimi K2	0.78	0.44	0.33	0.71	0.48	0.24	0.62	0.23	0.27	0.79	0.35	0.35	0.86	0.08	0.11

Table 9: Per-class F1 scores (Y=YES, N=NO, A=ABSTAIN). [T]=thinking mode

Vetoed Draft Resolution Example (1 of 211)	
<b>Document</b>	S/2022/431
<b>Date</b>	2022-05-26
<b>Agenda</b>	Non-proliferation / Democratic People’s Republic of Korea
<b>Vote Tally</b>	<b>In Favour (13):</b> Albania, Brazil, France, Gabon, Ghana, India, Ireland, Kenya, Mexico, Norway, United Arab Emirates, United Kingdom, United States <b>Against (2):</b> China, Russian Federation <b>Abstaining (0):</b> None
<b>P5 Votes</b>	US: YES UK: YES FR: YES CN: NO (veto) RU: NO (veto)
<b>Summary</b>	Draft resolution to strengthen sanctions on North Korea following ballistic missile tests. China and Russia jointly vetoed, marking the first double P5 veto on DPRK sanctions since 2006.

Table 10: Example vetoed draft resolution with dual P5 vetoes.

<b>Adopted Resolution Example (1 of 258)</b>	
<b>Document</b>	S/RES/43(1948)
<b>Date</b>	1948-04-01
<b>Agenda</b>	The Palestine Question
<b>Vote Tally</b>	<b>In Favour (5):</b> China, France, USSR, United Kingdom, United States <b>Against (0):</b> None <b>Abstaining (0):</b> None
<b>P5 Votes</b>	US: YES UK: YES FR: YES CN: YES RU: YES
<b>Summary</b>	The Security Council notes increasing violence in Palestine and believes an immediate truce is urgently needed. Calls upon the Jewish Agency and Arab Higher Committee to arrange a truce; calls upon armed groups to cease violence immediately.

Table 11: Example adopted resolution.

<b>Model Prompt-Response Example</b>	
<b>Document</b>	S/2023/970
<b>Date</b>	2023-12-08
<b>Agenda</b>	Middle East, including the Palestinian question
<b>Country Role</b>	Russian Federation
<b>Ground Truth</b>	YES
<b>Prompt</b>	<p>You are a representative of “Russian Federation” in United Nations Security Council (UNSC). You have to decide to vote on the following context of resolution “S/2023/970” in UNSC. You only choose the vote among three options: YES, NO, or ABSTAIN. Choose the best option considering your country’s interest.</p> <p><i>Context:</i> Expressing grave concern over the catastrophic humanitarian situation in the Gaza Strip... 1. Demands an immediate humanitarian ceasefire; 2. Reiterates its demand that all parties comply with their obligations under international law; 3. Demands the immediate and unconditional release of all hostages... <i>(omitted)</i></p>
<b>Response</b>	<p><b>Vote:</b> YES</p> <p><b>Rationale:</b> The Russian Federation supports the draft resolution as it aligns with its commitment to international humanitarian law and the protection of civilians in conflict zones. The demand for an immediate humanitarian ceasefire is crucial in addressing the catastrophic humanitarian situation in the Gaza Strip, which resonates with Russia’s historical stance on humanitarian access.</p>

Table 12: Example prompt-response pair for country role-play task.

# Leveraging Wikidata for Geographically Informed Sociocultural Bias Dataset Creation: Application to Latin America

Yannis Karmim<sup>1</sup>, Renato Pino<sup>2</sup>, Hernan Contreras<sup>3</sup>, Hernan Lira<sup>4</sup>, Sebastian Cifuentes<sup>5</sup>, Simon Escoffier<sup>6</sup>, Luis Martí<sup>4</sup>, Djamé Seddah<sup>1</sup>, Valentin Barriere<sup>2,5</sup>

<sup>1</sup>ALMAAnaCH team, Inria Paris Center; <sup>2</sup>Dept. of Computer Science, Universidad de Chile;

<sup>3</sup>Institute of International Studies, Universidad de Chile;

<sup>4</sup>Inria Chile Research Center; <sup>5</sup>Centro Nacional de Inteligencia Artificial;

<sup>6</sup>School of Social Work, Pontificia Universidad Católica de Chile.

Correspondence: [vbarriere@dcc.uchile.cl](mailto:vbarriere@dcc.uchile.cl)

## Abstract

Large Language Models (LLMs) exhibit inequalities with respect to various cultural contexts. Most prominent open-weights models are trained on Global North data and show prejudicial behavior towards other cultures. Moreover, there is a notable lack of resources to detect biases in non-English languages, especially from Latin America (Latam), a continent containing various cultures, even though they share a common cultural ground. We propose to leverage the content of Wikipedia, the structure of the Wikidata knowledge graph, and expert knowledge from social science in order to create a dataset of question/answer (Q/As) pairs, based on the different popular and social cultures of various Latin American countries. We create the LatamQA database of over 26k questions and associated answers extracted from 26k Wikipedia articles, and transformed into multiple-choice questions (MCQ) in Spanish and Portuguese, in turn translated to English. We use this MCQ to quantify the degree of knowledge of various LLMs and find out (i) a discrepancy in performances between the Latam countries, ones being easier than others for the majority of the models, (ii) that the models perform better in their original language, and (iii) that Iberian Spanish culture is better known than Latam one.<sup>1</sup>

## 1 Introduction and Related Work

Disciplinary standards for “valid” knowledge have been concentrated in Western Europe and North America (Demeter, 2020). Biases in AI systems, particularly in NLP, often originate from training data (Wiegand et al., 2019), annotation practices (Sap et al., 2022), and annotation guidelines (Parmar et al., 2023). These biases may take moral (Hämmerl et al., 2022), social (Sap et al., 2020), class-based (Curry et al., 2024), or political forms

<sup>1</sup>Code and datasets available at <https://github.com/Inria-Chile/LatamQA>.



Figure 1: Geographic distribution of LatamQA cultural MCQs across Latin America composed of 23k Q/As.

(Feng et al., 2023). Although social biases can be explicitly annotated for detection and analysis (Sahoo et al., 2023), annotation is costly and highly dependent on linguistic and cultural context (Fort et al., 2024; Barriere and Cifuentes, 2024b,a), and model’s biases can be sensitive to simple context changes (Quiroga et al., 2025). Recent studies emphasize the importance of localizing dataset construction and warn against outsourcing bias-related annotation for non-English languages to actors in the Global North (Hada et al., 2024). Such practices risk overlooking culturally specific meanings and social dynamics, as many existing datasets inadequately represent the cultural contexts of the Global South. (Santy et al., 2023).

Cultural biases in LLMs are particularly underexplored for Latin America (Latam). Although many countries share Spanish or Portuguese as official languages, they differ substantially in historical, social, and cultural terms, making the region well suited for evaluating fine-grained cultural knowledge. However, current geo-cultural datasets either group countries coarsely (Czarnowska et al., 2021; Li et al., 2024), cover few Latam countries (Myung et al., 2024), merge heterogeneous regions (Adilazuarda et al., 2024), or are limited to En-

Datasets	# (k)	Region	SP/PT
BLEnD (Myung et al., 2024)	2	Mixed	✓
CulFIT (Feng et al., 2025)	.08	Latam	✓
CANDLE (Nguyen et al., 2023)	2	Latam	✓
CultureBank (Shi et al., 2024)	3.3	Latam	✗
CultureAtlas (Fung et al., 2024)	1.8	Mixed	✗
GLOBAL-MMLU (Singh et al., 2024)	.01	Latam	✓
LatamQA (ours)	26	Latam	✓

Table 1: Latam-related questions in cultural Q/As datasets in terms of number of entries, region coverage, and if in Spanish and/or Portuguese.

glish (Feng et al., 2025). In Global-MMLU, only 1.6% of culturally relevant content concerns Latam (Singh et al., 2024). Table 1 summarizes existing cultural Q/As datasets with Latam content.

**Cultural Benchmark Creation** Building robust datasets that capture region-specific cultural knowledge is essential for evaluating biases in LLMs (Hershcovich et al., 2022; Liu et al., 2024; Pawar et al., 2024). Manual annotation efforts such as BLEnD (Myung et al., 2024) and CulturalBench (Chiu et al., 2025) offer high-quality data but are inherently limited in scale. In contrast, automated approaches construct cultural benchmarks by extracting data from large corpora. For example, Nguyen et al. (2023) rely on the C4 corpus, which has been shown to contain substantial noise (Fung et al., 2024). Other methods draw on social media platforms, such as CultureBank (Shi et al., 2024), which extracts from TikTok and Reddit or use curated web corpora, such as CRAFT (Wang et al., 2024), which retrieves documents from SlimPajama using keywords and LLM-generated Q/As. Wikipedia offers a middle ground with clean, curated content. CultureAtlas (Fung et al., 2024) extracts from Wikipedia using hyper-link expansion with NLI filtering; Li et al. (2024) use Wikipedia categories; Zhao et al. (2025) leverage Wikidata paths. Unlike these approaches, we exploit Wikipedia’s pre-existing category ontology combined with sociologist-guided validation and LLM-based Q/A generation.

**Language and Geographic Analysis** The relationship between prompting language and cultural knowledge retrieval remains open. XNationQA (Tanwar et al., 2025) finds prompting language significantly impacts performance, while Ying et al. (2025) show models perform better in native languages—though findings conflict across

studies (Zhao et al., 2025). Following prior work on common ground (Adilazuarda et al., 2024; Hershcovich et al., 2022), we focus on local facts at a fine-grained geographic level, particularly popular culture and sociocultural references that support shared understanding (Adams et al., 2004). Given the close relationship between language and culture (Hershcovich et al., 2022), we evaluate models using each country’s native language.

**Our Approach** Prior datasets primarily focus on cultural norms or social practices. We target culturally grounded factual knowledge: shared references that define collective identity. We leverage Wikimedia resources together with social science expertise to construct a large-scale dataset for Latam, combining: (i) Portuguese and Spanish alongside English to investigate language effects; (ii) comparison between Latam and Spain to assess regional representation in training data; and (iii) fine-grained analysis across Latam countries and cultural elements. Our contributions are:

- a scalable methodology for creating geographically informed sociocultural Q/A datasets using Wikipedia categories, expert curation, and LLM-based generation,
- the LatamQA benchmark of 23,499 multiple-choice questions covering 20 Latam countries (see Figure 1) in Spanish, Portuguese, and English, and
- an empirical analysis of performance variation across countries, prompting languages, and between Latam and Iberian Spanish cultural knowledge.

## 2 Benchmark Creation

### 2.1 Raw Wikipedia Data

We apply two sociology-based filters to enhance the dataset pertinence: one at the category level and one at the article level.

**Collection** Our data collection method relies on the Wikipedia categories’ ontology. As each category contains articles and subcategories, it is possible to scrap the content in a recursive way, and, therefore, obtain a structured list of articles with associated metadata. We start from a mother category containing cultural information about a Region of Interest (RoI) such as “*Cultura de Chile*,” “*Cultura de Peru*,” or other RoI, and recursively collect the links of the Wikipedia articles and subcategories (see Algorithm A.1 in Appendix A).

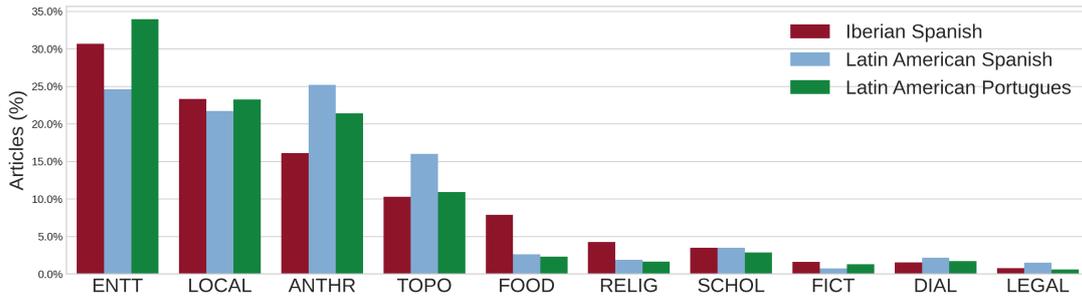


Figure 2: Distribution of the ratio of articles per cultural element per Language or Region in LatamQA. Cultural elements are: Anthroponyms (ANTHR), Forms of entertainment (ENTT), Local Institution (LOCAL), Toponyms (TOPO), Dialect (DIAL), Food and Drink (FOOD), Legal System (LEGAL), Scholastic reference (SCHOL), Religious celebration (RELIG), Fictional character (FICT).

A manual validation of the main subcategories<sup>2</sup> from a sociologist helps removing the categories that are not relevant for a RoI, such as “*Idioma Español*” which contains everything related to Spanish language in general, or “*Alumnados de [ENT]*” which contains all the people that went to the school [ENT]. This allowed us to obtain 154k articles. Metadata from Wikimedia was used to filter out the documents not relevant to the specific country.

**Curation** Not all articles contained within the remaining subcategories are equally relevant to cultural analysis. To address this, we apply a second filter at the article level by manually annotating each article according to its socio-cultural relevance. We define three classes: positive, descriptive, and negative. The negative class includes articles that do not address any cultural elements defined in the taxonomy proposed by [Espindola and Vasconcellos \(2006\)](#). Articles that are culturally relevant are assigned to either the positive or descriptive classes. The descriptive class is used for articles that primarily contain technical or enumerative information details<sup>3</sup> with limited interpretive value, such as lists of songs from a specific music album. 500 articles were manually tagged and used to fine-tune a pre-trained multilingual Longformer ([Beltagy et al.](#)), reaching a precision of 87.5% with respect to the positive class, and 100% when merging the positive and descriptive classes. Details are available in Appendix B.

**Cultural Elements Distribution** Wikipedia articles are generally associated with metadata specifying their entity type within an ontology ([Vrandečić,](#)

<sup>2</sup>up to three layers of depth inside the ontology

<sup>3</sup>List of the football teams, statistics, transfer dates of a player vs. political history, details on the anthem, rivalries with opponents of a club.

2012). We obtained 2,169 distinct entities across the entire dataset. Using an LLM (Qwen3-Max), we tagged the entities in a zero-shot in-context-learning way with rapid manual verification, mapping each Wikidata entity type to one of our predefined cultural elements. The resulting distribution of articles across cultural elements is presented in Figure 2.

## 2.2 Questions and Answers Generation

We leverage the filtered Wikipedia articles database to generate article-grounded questions and associated answers. Several prompting strategies were evaluated, with respect to a topic-dependant definition of culture that would apply the most to extract interesting knowledge from the Wikipedia data. gpt-oss-120b was used during this phase. Examples of questions and answers are available in Table 2.

**General Prompts** We compared several prompts to generate questions grounded with various definitions of cultures, based on: anthropology, general cultural exploration, psychological and symbolic significance, sociology, or on an integrative cultural definition. To select the definition leading to the most pertinent questions, they were manually validated by a sociologist with respect to their simplicity, quality and sociological pertinence. Details in Appendix D.

**Questions Generation and Validation** Once the culture definition fixed, we designed a prompt for article-grounded extraction of questions and associated answers in a structured way. We quantitatively validated the socio-cultural pertinence of the questions using a three-dimension notation based on ([Geertz, 1973; Hudson et al., 2009; Páez Rovira et al., 2007](#)) that : (i) symbolic, (ii) social practices, and (iii) social representations, memory and iden-

Category	Questions and Answers
FICT	Q: What role did the <i>Bacab</i> play in Maya beekeeping? A: They were the primary protectors of bees and founders of apiculture. Q: What is <i>gliglico</i> , and in which literary work does it appear? A: <i>Glíglico</i> is a fictional language created by Julio Cortázar and appears in his novel <i>Rayuela</i> .
FOOD	Q: In which Mexican state is the <i>memela</i> considered a traditional dish? A: The <i>memela</i> is a traditional dish from the state of Puebla. Q: What cultural origins does Lima’s <i>mazamorra morada</i> have in Peru? A: Lima’s <i>mazamorra morada</i> has Afro-Peruvian roots and is part of Peru’s culinary identity.
DIAL	Q: What old expression used in Mexico City means “it seems to me” and remains in everyday speech? A: The expression “ <i>se me hace</i> ” is used in Mexico City to mean “it seems to me.” Q: In Chile, which social group is the term “ <i>flaute</i> ” disparagingly directed toward? A: It refers to lower-class individuals who are socially inadapted and aggressive, and also to any vulgar behavior regardless of social origin.
RELI	Q: What is the role of the <i>machi</i> in Mapuche culture? A: The <i>machi</i> is a Mapuche medicine central figure, healing physical and spiritual ailments, with religious and social functions Q: On which day of the month is the tradition of eating <i>ñoquis</i> observed in Argentina, Uruguay, Brazil, and Paraguay? A: It is celebrated on the 29th of each month.
ENTTT	Q: According to the article, what is the origin of <i>cumbia</i> ? A: <i>cumbia</i> results from a mixture of Indigenous and African influences. Q: In which Mexican state is the novel <i>Falsa liebre</i> set? A: The novel is set in Veracruz, showing the most marginalized and violent side of the region.

Table 2: Example of questions and answers from different cultural elements.

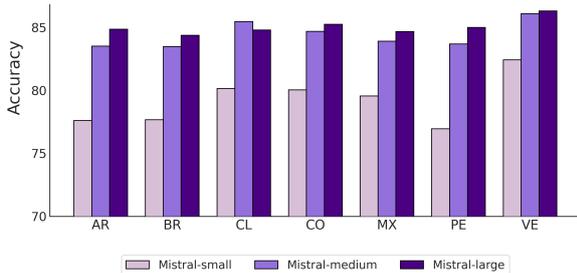


Figure 3: Cross-country performance of Mistral models on LatamQA. Scaling from Small to Large yields consistent improvements (+5 – +8% accuracy).

tity. We found that 98% of the questions were at least relevant in two of the three dimensions. We also validated quantitatively the answers’ grounding to the article on 100 examples, and found no case of hallucination. Details in Appendix E.

**Distractors Generations** We generated distractors as Fung et al. (2024), generating alternative answers with the same LLM that generated the questions. Details in Appendix D.2.

### 3 Experiments and Results

**Global Results and Prompting Language** The performances of various size models are visible in Table 3 and in the same range of other culture-related datasets (Myung et al., 2024; Ying et al., 2025). All the models are performing better in their native language (ES or PT), which is consistent with past results as Spanish and Portuguese are high-resource languages (Myung et al., 2024; Ying et al., 2025) but contradictory with other works (Tanwar et al., 2025; Zhao et al., 2025). We believe that might be due to the nature of the source document to create the Q/A (graph triplet) or because of heterogenous generation capabilities between languages (Kabir et al., 2025). We also include PatagonIA and LatamGPT (Latam-GPT, 2026), the former is specialized in Chilean Spanish, suppos-

Model	Brazilian PT		Latam SP		Spain	
	PT	EN	SP	EN	SP	EN
<i>Small models</i>						
Llama 3.1-8B	65.9	66.2	69.2	64.5	76.0	80.5
Mistral-small	77.0	74.3	78.5	76.1	84.3	81.4
<i>Medium models</i>						
Qwen2.5-14B	65.1	62.1	68.8	67.5	79.1	78.2
GPT-4.1-mini	80.0	76.1	81.5	78.2	<b>88.0</b>	85.1
Mistral-medium	82.6	81.8	83.9	80.5	87.1	85.4
<i>Large models</i>						
Qwen3-430B	70.8	71.4	75.8	74.0	83.7	82.4
Kimi-K2-thinking	69.6	70.5	71.6	70.9	81.0	76.1
Mistral-large	<b>84.3</b>	<b>83.0</b>	<b>85.4</b>	<b>81.8</b>	87.6	<b>86.4</b>
<i>LATAM models</i>						
PatagonIA	81.5	76.8	82.0	79.2	86.9	84.9
LatamGPT	28.0	33.4	48.3	37.0	37.8	29.7

Table 3: Performance of various LLMs on the LatamQA benchmark (accuracy %). We evaluate models with both native languages (PT and SP) and MT English translations.

edly based on a Nemotron architecture, the latter is a Llama 3.1 70B that has been pre-trained again over 300 billion tokens spanning Spanish, English, and Portuguese—with a significant portion of the data sourced directly from various countries across the Latin American region. Despite its regional specialization, PatagonIA does not outperform general-purpose models of comparable size such as Mistral-medium or GPT-4.1-mini. Latam-GPT struggles complying with MCQ (only 73% of the time), and even when evaluating on the questions it directly answers, its performances are close to random for Portuguese and less than 64% for Latam Spanish.

**Model Size vs. Performance** While performances are heterogenous with respect to the LLMs, it is notable that they are homogenous with respect to the size of the model. We can notice consistent improvements, with the exception of Mexico, on the biggest countries for various Mistral models in

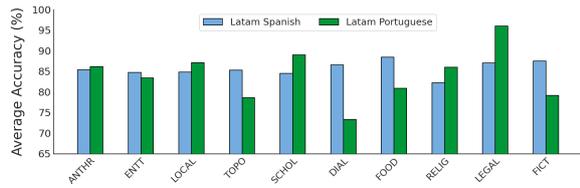


Figure 4: Performance of Mistral-large in Latam Spanish and Portuguese with respect to the different cultural elements.

Figure 3.<sup>4</sup>

**Iberian vs. Latam Spanish** Using a similar process, we extracted a set of Q/As from Spain to compare the performances of the models. All the models performed better on the Iberian Spanish subset. The results are coherent with Myung et al. (2024), where the models reached higher performance on questions from Spain than questions Mexico subset. It is interesting to note that even if Mistral models still perform very well, now the best results are obtained with GPT-4. 1.

**Cultural Element-level Analysis** Leveraging Wikidata’s type of entity ontology, we automatically map every article and its associated question to its cultural element in the taxonomy of Espindola and Vasconcellos (2006). Performance gaps (see Figure 4) between Latam Spanish and Brazilian Portuguese are higher when the cultural elements contain few examples such as “Fictional character” and “Dialect” where we observed between 75% and 80% of accuracy for Brazilian Portuguese.

## 4 Conclusion

This work introduces a new sociocultural benchmark focused on Latam. The benchmark contains more than 23k distinct multiple-choice questions (MCQs), each paired with a ground-truth answer. We construct this large-scale, structured dataset by combining information from Wikipedia and Wikidata with domain expertise from the social sciences. First, LLM performance varies substantially across models, although it remains consistent across different scales within the Mistral family. Second, using the native language of the target culture leads to better performance for Spanish and Portuguese. Third, all evaluated LLMs perform better in Iberian Spanish than in Latin American Spanish. Fourth, an analysis at the level of cultural elements shows that performance varies depending on the type of knowledge being tested.

<sup>4</sup>Same phenomena observed for Qwen2.5 and Qwen3.

## 5 Limitations

This work represents an initial step toward estimating the cultural knowledge of large language models (LLMs) in South America. However, cultural knowledge cannot be adequately captured through simple prompt-based question answering alone (Zhou et al., 2025; Kabir et al., 2025). Future work should therefore move beyond basic multiple-choice question answering (MCQA) benchmarks (Oh et al., 2025). Promising directions include directly involving human participants in benchmark construction (Ivetta et al., 2025a,b) and analyzing interactional data, such as discussions in Wikipedia Talk Pages associated with the target articles.. Similarly, we are aware of the possible preference biases (Wataoka et al.) that might be introduced by using only one LLM.

## Acknowledgments

This work was partially financed with the grant U-INICIA 2024 from the Vicerrectoría de Investigación y Desarrollo (VID) number UI-011/24 “Estudios de sesgos sociales en modelos de lenguajes largos”, by the Franco-Chilean Binational Center of Artificial Intelligence, ANID Strengthening R&D capabilities Program CTI230007 Inria Chile. This work was granted access to the HPC resources of IDRIS under the allocation 2025-A0180616119 made by GENCI.

## References

- Glenn Adams, Stephanie L Anderson, and Joseph K Adonu. 2004. The cultural grounding of closeness and intimacy. In *Handbook of closeness and intimacy*, pages 331–350. Psychology Press.
- Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Singh, Ashutosh Dwivedi, Alham Fikri Aji, Jacki O’Neill, Ashutosh Modi, and Monojit Choudhury. 2024. *Towards Measuring and Modeling "Culture" in LLMs: A Survey. EMNLP.*
- Valentin Barriere and Sebastian Cifuentes. 2024a. *Are Text Classifiers Xenophobic? A Country-Oriented Bias Detection Method with Least Confounding Variables. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1511–1518, Torino, Italia. ELRA and ICCL.
- Valentin Barriere and Sebastian Cifuentes. 2024b. *A study of nationality bias in names and perplexity using off-the-shelf affect-related tweet classifiers. In Proceedings of the 2024 Conference on Empirical*

- Methods in Natural Language Processing*, pages 569–579, Miami, Florida, USA. Association for Computational Linguistics.
- Iz Beltagy, Matthew E Peters, and Arman Cohan. [Longformer: The Long-Document Transformer](#).
- Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, and Yejin Choi. 2025. [CULTURALBENCH: A ROBUST, DIVERSE AND CHALLENGING BENCHMARK ON MEASURING THE \(LACK OF\) CULTURAL KNOWLEDGE OF LLMs](#). In *ACL*, pages 1–26.
- Amanda Cercas Curry, Giuseppe Attanasio, Zeerak Talat, Mohamed Bin Zayed, and Dirk Hovy. 2024. [Classist Tools: Social Class Correlates with Performance in NLP](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 12643–12655.
- Paula Czarnowska, Yogarshi Vyas, and Kashif Shah. 2021. [Quantifying social biases in nlp: A generalization and empirical comparison of extrinsic fairness metrics](#). *Transactions of the Association for Computational Linguistics*, 9:1249–1267.
- Márton Demeter. 2020. *Academic knowledge production and the global south: Questioning inequality and under-representation*. Springer.
- Elaine Espindola and María-Lucia Vasconcellos. 2006. Two facets in the Subtiling Process: foreignisation and/or domestication procedures in unequal cultural encounters. *Fragmentos: revista de língua e literatura estrangeiras*, (30):43–66.
- Ruixiang Feng, Shen Gao, Xiuying Chen, Lisi Chen, and Shuo Shang. 2025. [CulFiT: A Fine-grained Cultural-aware LLM Training Paradigm via Multilingual Critique Data Synthesis](#). volume 1, pages 22413–22430.
- Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. [From Pretraining Data to Language Models to Downstream Tasks: Tracking the Trails of Political Biases Leading to Unfair NLP Models](#). In *ACL*, volume 1, pages 11737–11762.
- Karën Fort, Laura Alonso Alemany, Luciana Benotti, Julien Bezançon, Claudia Borg, Marthese Borg, Yongjian Chen, Fanny Duce, Yoann Dupont, Guido Ivetta, Zhijian Li, Margot Mieskes, Marco Naguib, Yuyan Qian, Matteo Radaelli, Wolfgang S Schmeisser-nieto, Emma Raimundo Schulz, Thiziri Saci, Sarah Saidi, and 6 others. 2024. [Your Stereotypical Mileage may Vary : Practical Challenges of Evaluating Biases in Multiple Languages and Cultural Contexts](#). In *LREC-COLING*, 2, pages 17764–17769.
- Yi Fung, Ruining Zhao, Jae Doo, Chenkai Sun, and Heng Ji. 2024. [No Culture Left Behind: Massively Multi-Cultural Knowledge Acquisition & LM Benchmarking](#).
- Clifford Geertz. 1973. Thick Description: Toward an interpretive theory of culture. *The interpretation of cultures: Selected essays*, pages 3–30.
- H. P. Grice. 1975. [Logic and conversation](#). *Syntax and Semantics*, 3:41–58.
- Rishav Hada, Safiya Husain, Varun Gumma, Harshita Diddee, Aditya Yadavalli, Agrima Seth, Nidhi Kulkarni, Ujwal Gadiraju, Aditya Vashistha, Vivek Seshadri, and Kalika Bali. 2024. [Akal Badi ya Bias: An Exploratory Study of Gender Bias in Hindi Language Technology](#). In *2024 ACM Conference on Fairness, Accountability, and Transparency, FAccT 2024*, volume 1, pages 1926–1939. Association for Computing Machinery.
- Katharina Hämmerl, Björn Deiseroth, Patrick Schramowski, Jindřich Libovický, Constantin A. Rothkopf, Alexander Fraser, and Kristian Kersting. 2022. [Speaking Multiple Languages Affects the Moral Bias of Language Models](#). In *Findings of ACL: ACL 2023*, pages 2137–2156.
- Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. [Challenges and Strategies in Cross-Cultural NLP](#). *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 1:6997–7013.
- Scott Hudson, C Smith, M Loughlin, and S Hammerstedt. 2009. [Symbolic and interpretive anthropologies](#). Retrieved December 2025, 6:2013.
- Guido Ivetta, Marcos J Gomez, Sofía Martinelli, Pietro Palombini, M Emilia Echeveste, Nair Carolina Mazzeo, Beatriz Busaniche, and Luciana Benotti. 2025a. [HESEIA : A community-based dataset for evaluating social biases in large language models , co-designed in real school settings in Latin America](#). *EMNLP*, pages 25107–25129.
- Guido Ivetta, Pietro Palombini, Sofía Martinelli, Marcos J Gomez, Sunipa Dev, Vinodkumar Prabhakaran, and Luciana Benotti. 2025b. [Adaptive Data Collection for Latin-American Community-sourced Evaluation of Stereotypes \(LACES\)](#).
- Mohsinul Kabir, Ajwad Abrar, and Sophia Ananiadou. 2025. [Break the Checkbox : Challenging Closed-Style Evaluations of Cultural Alignment in LLMs](#). In *EMNLP*, pages 25–52.
- Latam-GPT. 2026. [Latam-GPT](#). In <https://www.latamgpt.org/>.
- Jialin Li, Junli Wang, Junjie Hu, and Ming Jiang. 2024. [How Well Do LLMs Identify Cultural Unity in Diversity?](#) In *CoLM*, pages 1–23.

- Chen Cecilia Liu, Iryna Gurevych, and Anna Korhonen. 2024. [Culturally Aware and Adapted NLP: A Taxonomy and a Survey of the State of the Art](#). In *Proceedings of the 2nd Workshop on Cross-Cultural Considerations in NLP*.
- Luis Josué Lugo Sánchez. 2025. Innovación social académica en tiempos de capitalismo cognitivo: El caso de la Biblioteca de Prompts Colaborativos. *Teknokultura: Revista de Cultura Digital y Movimientos Sociales*, 22(2):185–196.
- Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Afina Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, Víctor Gutiérrez-Basulto, Yazmín Ibáñez-García, Hwaran Lee, Shamsuddeen Hassan Muhammad, Kiwoong Park, Anar Sabuhi Rzayev, Nina White, Seid Muhie Yimam, Mohammad Taher Pilehvar, and 3 others. 2024. [BLEnD: A Benchmark for LLMs on Everyday Knowledge in Diverse Cultures and Languages](#). *NeurIPS Datasets and Benchmarks Track*, pages 1–36.
- Tuan Phong Nguyen, Simon Razniewski, Aparna Varde, and Gerhard Weikum. 2023. [Extracting Cultural Commonsense Knowledge at Scale](#). *ACM Web Conference 2023 - Proceedings of the World Wide Web Conference, WWW 2023*, pages 1907–1917.
- Juhyun Oh, Inha Cha, Michael Saxon, Hyunseung Lim, Shaily Bhatt, and Alice Oh. 2025. Culture is Everywhere : A Call for Intentionally Cultural Evaluation. In *EMNLP*, pages 19156–19168.
- Darío Páez Rovira, Elza María Techio, and José Marques. 2007. Memoria colectiva y social. In *Psicología social*, pages 693–716. McGraw-Hill USA.
- Mihir Parmar, Swaroop Mishra, Mor Geva, and Chitta Baral. 2023. Don't Blame the Annotator: Bias Already Starts in the Annotation Instructions. In *EACL 2023 - 17th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference*, pages 1771–1781.
- Siddhesh Pawar, Junyeong Park, Jiho Jin, Arnav Arora, Junho Myung, Srishti Yadav, Faiz Ghifari Haznitrama, Inhwa Song, Alice Oh, and Isabelle Augenstein. 2024. [Survey of Cultural Awareness in Language Models: Text and Beyond](#). pages 1–87.
- Tamara Quiroga, Felipe Bravo-Marquez, and Valentin Barriere. 2025. [Adapting bias evaluation to domain contexts using generative models](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 28055–28066, Suzhou, China. Association for Computational Linguistics.
- Taffy E Raphael. 1986. Teaching Question Answer Relationships, Revisited. *Reading Teacher*, 39(6):516–522.
- Nihar Ranjan Sahoo, Niteesh Mallela, and Pushpak Bhattacharyya. 2023. With Prejudice to None : A Few-Shot , Multilingual Transfer Learning Approach to Detect Social Bias in Low Resource Languages. In *Findings of ACL: ACL 2023*, pages 13316–13330.
- Sebastin Santy, Jenny T. Liang, Ronan Le Bras, Katharina Reinecke, and Maarten Sap. 2023. [NLPositionality: Characterizing Design Biases of Datasets and Models](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics Volume 1: Long Papers*, volume 1, pages 9080–9102.
- Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social Bias Frames: Reasoning about Social and Power Implications of Language. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490.
- Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. [Annotators with Attitudes: How Annotator Beliefs And Identities Bias Toxic Language Detection](#). *NAACL 2022 - 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference*, pages 5884–5906.
- Weiyan Shi, Ryan Li, Yutong Zhang, Caleb Ziemis, Chunhua Yu, Raya Horesh, Rogério Abreu de Paula, and Diyi Yang. 2024. [CultureBank: An Online Community-Driven Knowledge Base Towards Culturally Aware Language Technologies](#). In *Findings of ACL: EMNLP 2024*, pages 1–32.
- Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David I. Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Wei-Yin Ko, Madeline Smith, Antoine Bosselut, Alice Oh, Andre F. T. Martins, Leshem Choshen, Daphne Ippolito, and 4 others. 2024. [Global MMLU: Understanding and Addressing Cultural and Linguistic Biases in Multilingual Evaluation](#).
- Eshaan Tanwar, Anwoy Chatterjee, Michael Saxon, Alon Albalak, William Yang, and Wang Tanmoy. 2025. Do You Know About My Nation ? Investigating Multilingual Language Models ' Cultural Literacy Through Factual Knowledge. In *EMNLP*, pages 14968–14991.
- Denny Vrandečić. 2012. Wikidata: A new platform for collaborative data collection. In *Proceedings of the 21st international conference on world wide web*, pages 1063–1064.
- Bin Wang, Geyu Lin, Zhengyuan Liu, Chengwei Wei, and Nancy F Chen. 2024. [CRAFT: Extracting and Tuning Cultural Instructions from the Wild](#). In *Proceedings of the 2nd Workshop on Cross-Cultural Considerations in NLP*, pages 42–47.
- Koki Wataoka, Tsubasa Takahashi, and Ryokan Ri. [Self-preference bias in LLM-as-a-judge](#).

Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. Detection of abusive language: The problem of biased datasets. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 1:602–608.

Jiahao Ying, Wei Tang, Yiran Zhao, Yixin Cao, Yu Rong, and Wenxuan Zhang. 2025. Disentangling Language and Culture for Evaluating Multilingual Large Language Models. In *ACL*, volume 1, pages 22230–22251.

Raoyuan Zhao, Beiduo Chen, Barbara Plank, and Michael A Hedderich. 2025. MAKI EVAL: A Multilingual Automatic Wikidata-based Framework for Cultural Awareness Evaluation for LLMs. In *Findings of ACL: EMNLP 2025*, pages 23104–23136.

Naitian Zhou, David Bamman, and Isaac L. Bleaman. 2025. Culture is Not Trivia: Sociocultural Theory for Cultural NLP.

## A Scraping Algorithm

The scraping algorithm is described in Algorithm A.1. We collected articles from the category pages titled “*Cultura de [Country]*” for the 20 countries listed in Table C.1 using each country’s main language. We set MAX\_DEPTH to 5 as empirical testing showed that greater depths reduced the relevance of the retrieved articles. For Spain, we used MAX\_DEPTH = 3 due to the substantially larger initial number of articles.

## B Automatic Filtering

### B.1 Elements of Culture

We base our filtering on the elements of culture from Espindola and Vasconcellos (2006) used to filter out articles in the negative class.

#### B.1.1 Definitions

- **TOPO** (Toponyms): a place name, a geographical name, a proper name of locality, region, or some other part of Earth’s surface or its natural or artificial feature.
- **ANTHR** (Anthroponyms): ordinary and famous people’s names and nick-names and names referring to regional background which acquire identification status; it also includes animals that have been given human qualities, symbols, or political meanings in social representations.
- **ENTT** (Forms of entertainment): amusement or diversion including public performances or shows, it also encompasses hospitality

provided, such as dinners, parties, business lunches, etc.; it also includes artistic expressions.

- **FICT** (Fictional character): a person in a novel, play, or a film who is related to fiction, works of imagination.
- **LEGAL** (Legal System): rules of conduct inherent in human nature and essential to or binding upon human society.
- **INST** (Local Institution): an organization that helps or serves people in a certain area - health, education, work, political, administrative, religious, artistic; it also includes national symbols.
- **FOOD** (Food and Drink): any solid or liquid substance that is used by human beings as a source of nourishment.
- **SCHOL** (Scholastic reference): related to school or studying.
- **RELIG** (Religious celebration): to do something special to mark a religious occasion.
- **DIAL** (Dialect): user-related variation, which determines speaker’s status as regards social class, age, sex, education, etc.; it also includes slang.

#### B.1.2 Distribution

The distributions of the articles with respect to its cultural element relevance is shown in Figure 2. The biggest difference lies within the ratio of articles about Food and Drink: they are more dominant in Spain than in Latam.

### B.2 Classifier

We fine-tuned and validated a pre-trained XLM-RoBERTa Longformer<sup>5</sup> on 500 3-class examples. When merging the descriptive and positive classes, the classifier reaches an accuracy of 97.8%. The confusion matrix obtained from cross-validation is the following ( $c_{i,j} = c_{y,\hat{y}}$ ):

$$\begin{bmatrix} 198 & 24 & 0 \\ 43 & 98 & 19 \\ 11 & 40 & 131 \end{bmatrix}$$

## C Per-Country Distribution

The distribution of the dataset questions per country is shown in Table C.1.

<sup>5</sup>markussagen/xlm-roberta-longformer-base-4096

---

**Algorithm A.1** Recursive Wikipedia Category Scraper

---

```
1: Input: Initial Wikipedia category URL, Maximum recursion depth MAX_DEPTH
2: function SCRAPECATEGORY(categoryURL, currentDepth):
3:   if currentDepth > MAX_DEPTH then
4:     return
5:   end if
6:   Fetch HTML category page to extract articles and subcategory links
7:   for each article link NOT already processed do
8:     Fetch HTML article page content and save article data
9:   end for
10:  for each subcategory link do
11:    SCRAPECATEGORY(subcategoryURL, currentDepth + 1)
12:  end for
13: SCRAPECATEGORY(initialCategoryURL, currentDepth = 0)
```

---

Country/Region	Language	Count
Brazil (BR)	Portuguese	6,075
México (MX)	Spanish	4,893
Argentina (AR)	Spanish	4,243
Chile (CL)	Spanish	2,469
Perú (PE)	Spanish	1,921
Colombia (CL)	Spanish	1,752
Brazil (BR)	Spanish	1,164
Venezuela (VE)	Spanish	1,030
Cuba (CU)	Spanish	674
Ecuador (EC)	Spanish	720
Uruguay (UY)	Spanish	991
Bolivia (BO)	Spanish	750
Guatemala (GT)	Spanish	743
Costa Rica (CR)	Spanish	467
El Salvador (SV)	Spanish	306
Nicaragua (NI)	Spanish	436
Paraguay (PY)	Spanish	542
Dominican Republic (RD)	Spanish	234
Honduras (HN)	Spanish	180
Panamá (PA)	Spanish	218
Puerto Rico (PR)	Spanish	193
<b>Total</b>		<b>26,213</b>

Table C.1: Distribution of articles from the LatamQA dataset that are from a categories under the mother category "Cultura de [Country]". In this case, an article can be associated with several countries, and possibly languages.

## D Questions-Generation Prompts

### D.1 Domain-specific Culture Definition

We first selected a prompt from five different prompts using different definitions of culture: an anthropological approach, general cultural exploration approach, psychological and symbolic significance approach, sociological approach and integrative cultural definition approach. The quality

of the questions was assessed with respect to the clarity of language under “theoretical principles of phenomenology, which studies things as they are shown in consciousness to make them comprehensible” (Lugo Sánchez 2025; page 188), which means expressing the questions in simple terms. A good question is correctly formulated, asking for something precise and not ambiguous present in the article, not using an overly complex or specific vocabulary or concepts (such as “collective identity” or “communal expression”).

The general cultural exploration approach was judged the most relevant for the benchmark creation (see Figure D.1).

### D.2 MCQA Generation

Using general cultural exploration approach, a new prompt was designed to extract the pairs of Q/As, adding specific rules to force the questions and answer to be precise, explicit, pertinent and generated in a specific format (see Figure D.2). A good answer responds totally to the question, uses solely the content of the article without adding external facts, and does not add specific reasoning (Raphael, 1986; Grice, 1975).

Second, following the methodology of (Fung et al., 2024), another prompt was used to generate challenging counterfactual answers for the MCQ, which we call distractors (see Figure D.3).

## E Validation of Q/As

We asked two experts to score 100 questions with a 5-point likert scale with respect to the symbolic, the social practices, and the social representations, memory and identity. Only two questions over

**General Cultural Exploration**

**System:** From the given Wikipedia article, identify and formulate relevant sociocultural questions and answers. Questions should address topics such as cultural identity, popular symbols, collective memory, traditions, and practices specific to the local community or region described. Answers must briefly summarize the sociocultural relevance of these elements

Figure D.1: General Cultural Exploration approach prompt.

**Cultural QA Generation Prompt (Spanish)**

**System:** You are an expert in cultural anthropology and educational content creation. Your task is to generate three culturally meaningful question–answer pairs in Spanish based only on information explicitly present in the following Wikipedia article.

**Goal:** Identify the most culturally significant element explicitly mentioned in the article and produce three question–answer pairs about it, focusing on:

- Cultural identity
- Popular symbols
- Collective memory
- Traditions
- Practices specific to the local community

**Rules:**

- Use only information explicitly contained in the article.
- Select the most widely recognized cultural element mentioned.
- The question must be clear and culturally meaningful.
- The answer must be brief and concise, one or two sentences at most.
- Output must follow exactly the JSON format below.
- Do not invent information.

**Article:** {content}

**Output Format:**

```
{
  "question1": "...",
  "answer1": "...",
  "question2": "...",
  "answer2": "...",
  "question3": "...",
  "answer3": "..."}

```

Figure D.2: Prompt template for generating culturally grounded question–answer pairs from Spanish Wikipedia articles.

the 100 obtained a score less than 5, which means that only 2% of the questions obtained a score of 1/1/1 or 1/1/2 and were rejected. Inter-annotator-agreement was high.

## F Mistral Models Performances

**Cross-country Performances** The full performances of the models from the Mistral family are visible in Figure F.1. It is visible that the scale consistently helps in reaching higher performances. Except for a few countries where the medium excels slightly the large model: Costa-Rica, Honduras and Ecuador.

**Cultural Element-level Performances** Figure F.2 shows the performances of the Mistral models for separated with respect to the cultural elements of the questions.

**MCQ Distractor Generation Prompt (Spanish)**

**System:** You are an expert in creating difficult multiple-choice questions to evaluate reading comprehension and cultural knowledge.

**Context:**  
Article: {content} Question: {question} Correct answer: {answer}

**Task:** Generate exactly 3 incorrect answers (distractors) in Spanish that are:

- Plausible and similar in length/style to the correct answer
- Culturally believable (could reasonably be confused with the correct answer)
- Difficult to eliminate without having carefully read the article
- Grammatically correct in Spanish

**Strategies to use (vary across the 3 distractors):**

1. *Entity swap:* Use another cultural element, place, or name mentioned in the article
2. *Temporal/geographical confusion:* Use a date, period, or location from elsewhere in the article
3. *Partial truth:* Capture part of the cultural context but make false on the key element
4. *Cultural plausibility:* Reference a related but incorrect tradition from the same region
5. *Subtle relation inversion:* Invert origin/influence, cause/effect, or historical sequence

**Constraints:** Each distractor must fool a quick reader; avoid absurd answers and obvious negations; match the tone of the correct answer.

Figure D.3: Prompt template for generating challenging distractors for multiple-choice questions.

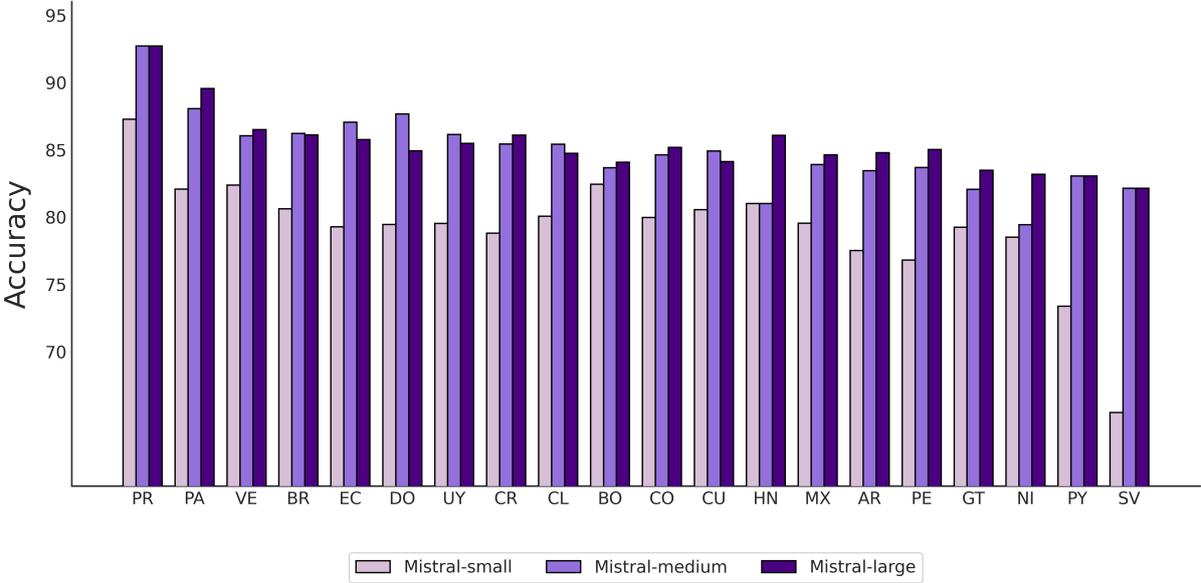


Figure F.1: Cross-country performance of Mistral models on cultural knowledge evaluation. Scaling from Small to Large yields consistent improvements (+5-8% accuracy).

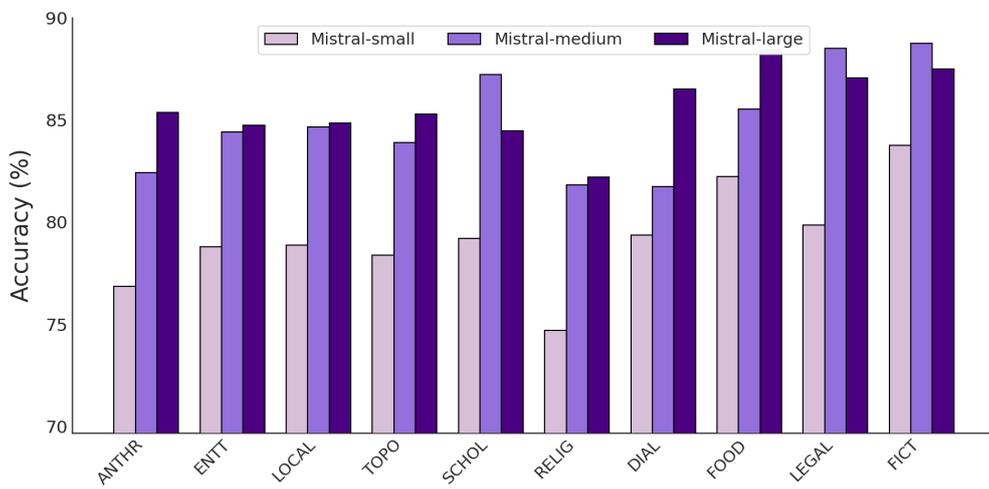


Figure F.2: Performance of Mistral models on Latam Spanish data, with respect to the cultural element.

# Whom to Trust? Analyzing the Divergence Between User Satisfaction and LLM-as-a-Judge in E-Commerce RAG Systems

Arif Türkmen

Trendyol

arif.turkmen@trendyol.com

Kaan Efe Keleş

Trendyol

efe.keles@trendyol.com

## Abstract

We study retrieval-augmented generation (RAG) evaluation in the Trendyol QA Assistant using 150k real e-commerce interactions. Our framework combines user satisfaction labels, *LLM-as-a-judge* scoring, and factor-based diagnostics to separate retrieval from generation errors. We find that judge models broadly reflect user satisfaction trends, though important nuances of dissatisfaction are often missed. Factor-level analysis highlights systematic error patterns across query types and context quality, demonstrating that hybrid evaluation, combining multiple LLM judges with direct user feedback offers the most reliable assessment strategy for production RAG systems.

## 1 Introduction

Retrieval-augmented generation (RAG) has risen as an effective paradigm for grounding large language models (LLMs) in customer-facing applications such as help desks, product Q&A, and enterprise copilots, by conditioning generation on retrieved domain knowledge—for greater factuality and relevance [Lewis et al., 2021, Yu et al., 2025]. Yet, evaluating RAG systems remains challenging because overall quality depends jointly on retrieval quality, context use, and generation [Lewis et al., 2021, Yu et al., 2025]. Recent benchmarks like *RAGBench* introduce explainable and modular metrics to assess retrieval and generation components [Friel et al., 2025], while retrieval-centric methods (e.g., *eRAG*) quantify document contributions to final answers [Yu et al., 2025]. Complementing these, the emerging “LLM-as-a-judge” paradigm allows scalable, automatic evaluation by having stronger models score responses—but systematic biases and blind spots remain, as highlighted in recent survey and bias studies [Gu et al., 2025, Ye et al., 2024].

We examine these evaluation modalities in a large-scale, real-world deployment: the *Trendyol*

*QA Assistant*. Trendyol—Türkiye’s largest e-commerce platform—handles hundreds of thousands of product questions daily, many of them similar to earlier question-answer pairs. The assistant addresses this by retrieving semantically similar historical Q/A pairs from past interactions with domain specific embedding model<sup>1</sup>, and conditioning a domain-specific LLM on this retrieved evidence<sup>2</sup> to generate a concise, grounded answer. In effect, the system automates seller responses by reusing and synthesizing prior knowledge, reducing latency and seller workload while preserving answer quality.

Our work makes three primary contributions. First, we present what is, to our knowledge, the first large-scale analysis of user satisfaction in a deployed e-commerce RAG system, using 150k real interactions from the Trendyol QA Assistant. Second, we provide the first empirical study demonstrating the divergence between LLM judge predictions and human preferences in Turkish-language generative outputs. Third, we document systematic evaluation failures of popular judge models in the e-commerce domain, identifying context-dependent blind spots that limit their reliability as standalone judges.

## 2 Evaluation Framework for User Satisfaction

A central challenge in evaluating the Trendyol QA Assistant lies in capturing end-user satisfaction at scale. Online A/B tests provide high-quality ground truth but are expensive and slow to iterate, motivating scalable alternatives. We therefore adopt a three-pronged evaluation methodology, combining direct human feedback with structured LLM-based approaches.

<sup>1</sup>TY-ecomm-embed-multilingual-base-v1.2.0

<sup>2</sup>Trendyol-LLM-8B-T1

## 2.1 Direct User Feedback

Ground-truth performance is obtained from user-reported satisfaction collected during live usage. After each interaction, users are asked in a thumbs-up or thumbs-down fashion whether they were satisfied with the assistant’s answer. If dissatisfied, they may optionally select one of four categorical reasons: *irrelevant*, *insufficient/incomplete*, *unclear*, or *misleading/incorrect*. This feedback provides the most accurate measure of user experience, though it is costly to scale and limited in experimental coverage.

## 2.2 LLM-as-a-Judge Simulation

To complement human feedback, we employ *LLM-as-a-judge* techniques, in which stronger models are prompted to evaluate QA interactions automatically using a fixed, structured judging prompt (Appendix A). For each interaction, the judge receives four inputs: (i) the user query, (ii) retrieved similar Q/A pairs, (iii) the assistant’s final response, and (iv) the assistant’s internal prompts used for generation (system prompt and base user prompt with examples). We use a few-shot version of this prompt and experimented with multiple candidate few-shot sets; a configuration with 10 diverse examples (spanning query types and common failure modes) yielded the best agreement on a held-out validation subset. The judge returns (a) a binary satisfaction decision and (b) if dissatisfied, one of the same four standardized reasons provided to users.

To assess alignment with users, we report two complementary measures. (i) *Satisfaction Agreement*: the exact match rate between the predictions of the judge and the satisfaction of the user in all interactions, contextualized with the expected chance agreement and Cohen’s  $\kappa$  [Cohen, 1960]. (ii) *Dissatisfaction Breakdown*: the categorical distribution of dissatisfaction reasons, enabling direct comparison between user-reported and judge-assigned error types.

These measures jointly assess sentiment alignment and dissatisfaction modeling, identifying where judge models fail to match real user feedback.

## 2.3 Factor-Based LLM Analysis

Finally, we use a granular, multi-prompt technique where the LLM isolates and evaluates specific factors. Prompts target different aspects to categorize question-answer pairs:

- **Query Classification**: It determines the topic of the question. Is the user asking about the product or the seller?
- **Intent Analysis**: It identifies the user’s goal. Is the user asking a genuine question or making a demand?
- **Contextual Relevance**: Did the information retrieved actually contain the necessary details to address the user’s query?
- **Persona Consistency**: Does the answer maintain the assistant’s intended style (formal tone, third-person narration) throughout?

Alongside these binary LLM outputs, we use string length as a verbosity measurement. This provides a consistent, model-agnostic metric, unlike token counts, which vary between different tokenizers. This factor-level decomposition helps identify whether user dissatisfaction arises from retrieval, grounding, or generation, providing actionable insights.

Our analysis is grounded in approximately 150k QA interactions from Trendyol’s latest production environment to ensure stability. Each instance contains the user query, retrieved Q/A pairs, the assistant’s response, and binary user satisfaction labels; dissatisfied users additionally provided categorical reasons. It is important to note that feedback collection relies on user initiative; while this means coverage is not universal, it provides a realistic representation of production dynamics where explicit feedback is sparse. This supervision enables systematic comparison between human feedback, LLM judges, and factor-based diagnostics at scale.

## 3 Results

Table 1 summarizes alignment between LLM-as-a-judge predictions and human feedback. The human-reported satisfaction prevalence is 77.2%. The two judges differ in calibration to this baseline (65.0% vs. 76.3%), with *o4-mini* closer to the target. However, agreement with user labels is only modest: exact-match rates are 64.5% (GPT-4o) and 72.5% (*o4-mini*), which translate to Cohen’s  $\kappa$  of 0.15 and 0.23, respectively. Given the class imbalance, these  $\kappa$  values indicate only slight-to-fair agreement beyond chance. We therefore treat LLM-as-a-judge as a useful heuristic for binary satisfaction but not a calibrated substitute for direct user feedback.

Table 1: Calibration and agreement of LLM-as-a-judge vs. human labels on 150k QA interactions.  $\Delta$  is model prevalence minus human prevalence. Agreement is exact match rate. Cohen’s  $\kappa$  adjusts for chance agreement using observed marginals.

Model	Satisfaction (%)	$\Delta$ vs Human (pp)	Agreement (%; 95% CI)	Cohen’s $\kappa$
GPT-4o	65.00	-12.2	64.48 [64.24, 64.72]	0.15
o4-mini	76.34	-0.83	72.53 [72.30, 72.76]	0.23
Gemini 2.5 Pro	80.60	+3.40	73.47 [72.60, 74.33]	0.20
Claude 4 Sonnet	78.72	+1.52	74.97 [74.12, 75.82]	0.27

Human satisfaction prevalence: 77.2%. Expected agreement by chance from marginals: 58.1% (GPT-4o), 64.3% (o4-mini), 66.6% (Gemini 2.5 Pro), 65.6% (Claude 4 Sonnet).

Table 2 compares the categorical breakdown of dissatisfaction reasons between users and LLM judges, each measured over their own dissatisfied subsets.

User feedback is dominated by insufficient/incomplete (62.3%), with meaningful fractions of unclear (15.7%) and misleading/incorrect (14.2%). However, the LLMs do not approximate this distribution. Since they can access retrieved contexts unlike users, they identify misleading/incorrect cases more aggressively.

These discrepancies reveal that while LLM judges may track overall satisfaction rates, they struggle to accurately model user behavior and may misrepresent user attitudes. By over-assigning “incorrectness” and overlooking issues like clarity or relevance, they provide a distorted diagnostic view. As a result, relying solely on judge distributions risks missing the user-centered issues that matter most for customer impact.

Table 2: Distribution of dissatisfaction reasons (%) across user feedback and LLM judges. Each row shows the categorical prevalence among all dissatisfied cases for that source.

	Irrelevant	Insufficient/ Incomplete	Unclear	Misleading
<i>Users</i>	7.8	62.3	15.7	14.2
<i>GPT-4o</i>	0.4	37.4	0.2	62.0
<i>o4-mini</i>	1.8	45.1	0.3	52.9
<i>Gemini 2.5 Pro</i>	0.7	45.7	0.8	52.8
<i>Claude 4 Sonnet</i>	0.9	59.1	0.8	39.2

As shown in Table 3, all models except GPT-4o achieve higher satisfaction with short responses than with long responses. Notably, Gemini 2.5 Pro attains the highest satisfaction under the long-response stratum (75.3%), which likely contributes to its strong overall satisfaction (80.6%) in this evaluation.

We zoom in on the two slices with the lowest human satisfaction: low-relevance contexts and

Table 3: Satisfaction rates (%) across different interaction strata. The categories analyze performance based on query type, context quality, and response length.

Category	Human	GPT-4o	o4-mini	Gemini	Claude
<i>Query Type</i>					
Product-Related	77.9	65.2	76.5	80.9	79.3
Seller-Related	64.9	62.9	70.4	73.7	67.5
Factual Q&A	77.5	65.3	76.3	79.4	77.1
User Commands	73.7	63.6	74.9	87.6	87.9
<i>Context Quality</i>					
High-Relevance	79.6	69.7	80.8	84.0	83.0
Low-Relevance	61.1	35.4	46.3	53.4	44.0
<i>Response Length</i>					
Short Response	80.3	64.9	78.5	81.9	80.7
Long Response	66.6	65.9	68.1	75.3	70.8

seller-related queries. As shown in Table 4, when context is weak, agreement significantly declines. Since users do not see the ground truth or relevant contexts on the UI; their judgments on these types of questions are not reliable. LLM judge systems have access to retrieved contexts and they can identify the misinformation and give worse scores than users, as shown in Table 3.

Table 4: Agreement of LLM-as-a-judge vs. human labels on questions that came with weakly related context.

Model	Agreement (%; 95% CI)	Cohen’s $\kappa$
GPT-4o	49.89[49.16, 50.62]	0.06
o4-mini	57.47[56.75, 58.19]	0.16
Gemini 2.5 Pro	57.92[55.03, 60.82]	0.15
Claude 4 Sonnet	56.88[53.95, 59.81]	0.15

Human satisfaction prevalence: 61.1%. Expected agreement by chance from marginals: 46.5% (GPT-4o), 49.1% (o4-mini), 52.1% (Gemini), 49.2% (Claude).

Another important slice is **seller-related queries** (Table 5). In a marketplace like Trendyol, many products have multiple sellers; user questions about shipping or packaging are therefore seller-specific, and the “correct” answer varies by seller. In this slice, o4-mini’s  $\kappa$  score sees a modest deterioration

from 0.23 to 0.21. However, Claude 4 Sonnet impacted heavily by seller-related questions, a major  $\kappa$  score decline 0.27 to 0.15. By contrast, GPT-4o shows a modest improvement, with Cohen’s  $\kappa$  rising from 0.15 to 0.19. These results underscore that LLM-judge performance is dependent on various factors and must be interpreted in context.

Table 5: Agreement of LLM-as-a-judge vs. human labels on seller-related questions.

Model	Agreement (% , 95% CI)	Cohen’s $\kappa$
GPT-4o	62.48[61.35, 63.60]	0.19
o4-mini	65.71[64.61, 66.81]	0.21
Gemini 2.5 Pro	62.93[58.66, 67.20]	0.15
Claude 4 Sonnet	62.70[58.45, 66.96]	0.17

Human satisfaction prevalence: 64.9%. Expected agreement by chance from marginals: 53.5% (GPT-4o), 56.1% (o4-mini), 57.0% (Gemini), 54.7% (Claude).

## 4 Discussion

Our findings highlight both the promise and pitfalls of LLM-as-a-judge for evaluating production RAG systems. Among the four judges evaluated, Claude 4 Sonnet achieves the highest agreement with users with 74.97% agreement and  $\kappa = 0.27$ , suggesting stronger alignment with human judgment patterns. A notable finding is the performance gap between GPT-4o and o4-mini, both from the same model family but differing in their reasoning capabilities. While GPT-4o achieves only 64.48% agreement ( $\kappa = 0.15$ ), o4-mini—an inference-time reasoning model—improves to 72.53% agreement ( $\kappa = 0.23$ ). This suggests that using reasoning models in evaluation tasks may help to simulate user judgment patterns, particularly when assessing complex QA interactions.

Our stratified analysis reveals that no single judge model performs uniformly well across all interaction types. In seller-related queries, Claude 4 Sonnet experiences the largest relative decline. These findings caution against deploying a single judge model for all query types and advocate for segment-aware evaluation strategies.

Beyond binary agreement, our analysis of dissatisfaction categories (Table 2) reveals a critical blind spot: LLM judges systematically over-attribute failures to “Misleading/Incorrect” (39–62%) while under-reporting “Insufficient/Incomplete” (37–59% vs. user-reported 62.3%). This asymmetry likely stems from judges’ access to retrieved contexts, enabling them to detect factual

inconsistencies invisible to users. Teams relying solely on LLM-judge metrics risk optimizing for factual accuracy while neglecting completeness, a mismatch with actual user pain points.

The findings above collectively argue against relying on any single LLM judge for a real system evaluation. Each model exhibits blind spots and all judges systematically misattribute user dissatisfaction categories. We therefore advocate for a *model ensemble* approach, where diverse judges vote to mitigate individual biases. More fundamentally, our results suggest that the most reliable evaluation strategy is a *hybrid* system combining LLM-judge with direct human feedback. LLM judges offer scalability and consistency for detecting factual errors and retrieval failures, while human ratings capture the subjective dimensions of satisfaction that models systematically miss.

## 5 Limitations

Our study is limited to a single production system within Trendyol, reflecting the characteristics of Turkish e-commerce queries. User feedback is binary with a restricted set of dissatisfaction categories, which constrains expressiveness. LLM-as-a-judge results are sensitive to prompt design and model choice, and we evaluate only four specific models.

## References

- Jacob Cohen. 1960. [A coefficient of agreement for nominal scales](#). *Educational and Psychological Measurement*, 20(1):37–46.
- Robert Friel, Masha Belyi, and Atindriyo Sanyal. 2025. [Ragbench: Explainable benchmark for retrieval-augmented generation systems](#). *Preprint*, arXiv:2407.11005.
- Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. [A survey on llm-as-a-judge](#). *Preprint*, arXiv:2411.15594.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. [Retrieval-augmented generation for knowledge-intensive nlp tasks](#). *Preprint*, arXiv:2005.11401.
- Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, Nitesh V Chawla, and

Xiangliang Zhang. 2024. [Justice or prejudice? quantifying biases in llm-as-a-judge](#). *Preprint*, arXiv:2410.02736.

Hao Yu, Aoran Gan, Kai Zhang, Shiwei Tong, Qi Liu, and Zhaofeng Liu. 2025. *Evaluation of Retrieval-Augmented Generation: A Survey*, page 102–120. Springer Nature Singapore.

## A Evaluation Prompt for LLM-as-a-Judge

### LLM-as-a-Judge Prompt (System + User)

#### System prompt

You are an impartial expert evaluator for an e-commerce Q&A assistant. You will receive the assistant's internal prompts, the user's question, the retrieved similar question/answer pairs, and the assistant's final answer. Your job is to critically assess the retrieval quality and the final answer, and then produce a concise, strictly structured JSON evaluation.

Be precise, objective, and consistent. Do not invent facts beyond the provided content.

#### User message

Evaluation context (for transparency, not to be judged for style):

Assistant system prompt: [SYSTEM\_PROMPT]

Assistant base user prompt (including few-shot examples): [USER\_PROMPT]

Inputs to evaluate:

- 1) User question: [QUESTION\_TEXT]
- 2) Retrieved similar question/answer pairs: [RETRIEVED\_QAS]
- 3) Assistant final answer to the user: [LLM\_ANSWER]

Evaluation requirements:

- Judge whether retrieved QA pairs are semantically relevant to the user's question.
- Judge whether the assistant's final answer directly addresses the user's question using only ↪ retrieved information.
- Penalize hallucinations or unsupported additions.
- If retrieval is irrelevant or insufficient, the correct assistant behavior is: "Üzgünüm bilmiyorum" (I am sorry, I do not know). Penalize deviations.

Output format (strict JSON object):

- question\_analysis: string
- retrieved\_answers\_analysis: string
- llm\_answer\_analysis: string
- satisfaction\_feedback\_analysis: string
- satisfaction\_feedback\_boolean: boolean
- satisfaction\_feedback\_negative\_reason: string  
(if dissatisfaction: one of "Irrelevant", "Insufficient/Incomplete", "Unclear", "Misleading/Incorrect"; else "None").

Guidance:

- Be concise but specific; reference concrete retrieved evidence.
- Treat greetings or seller politeness in retrieved QA pairs as noise.
- Reflect violations (fabrication, irrelevance, improper style) in analysis.
- For seller requests (e.g., "send red color"), the correct behavior is to decline with "I am sorry, I don't know.". Reward this.

Few-shot examples (redacted):

The judge prompt includes 10 such examples in practice.

To keep the appendix short, we do not show all examples here.

### Redacted few-shot example (one shown)

#### One Redacted Example (Format Illustration)

Product context (English, brand-redacted):

"21V Brushless Impact Rotary Hammer Drill & Grinder & Drill & Nut Tightening/Loosening 4-piece  
↪ Set"

Inputs:

- 1) User question: [EXAMPLE\_QUESTION\_TEXT]
- 2) Retrieved similar question/answer pairs: [EXAMPLE\_RETRIEVED\_QAS]
- 3) Assistant final answer to the user: [EXAMPLE\_LLM\_ANSWER]

Expected evaluation output (strict JSON object):

```
{
```

```
"question_analysis": "[...]",
"retrieved_answers_analysis": "[...]",
"llm_answer_analysis": "[...]",
"satisfaction_feedback_analysis": "[...]",
"satisfaction_feedback_boolean": [true/false],
"satisfaction_feedback_negative_reason": "[None | Irrelevant | Insufficient/Incomplete |
↔ Unclear | Misleading/Incorrect]"
}
```

# Query-Following vs Context-Anchoring: How LLMs Handle Cross-Turn Language Switching

Kyuhee Kim, Chengheng Li Chen, Anna Sotnikova

EPFL, Lausanne, Switzerland

kyuhee.kim@epfl.ch, chengheng.lichen@epfl.ch, anna.sotnikova@epfl.ch

## Abstract

When multilingual users switch languages mid-conversation, how should LLMs respond? We extend MultiChallenge to evaluate cross-turn language switching, translating 182 multi-turn conversations into German, Chinese, Spanish, and Arabic. Across five frontier models, we observe asymmetric behavior: switching into a foreign language ( $EN \rightarrow X$ ) yields high query-language fidelity (89–99%), but switching back to English ( $X \rightarrow EN$ ) reveals divergent policies. GPT-5 follows the query language (>95%), while Claude Opus 4.5 and Command R+ maintain the established conversation language (<8%). Task accuracy remains stable across conditions regardless of language selection differences. A simple explicit system prompt shows limited effectiveness in modifying these defaults. Our code and data are available at <https://github.com/koreankiwi99/crossturn-lang-switch>.

## 1 Introduction

Multilingual speakers frequently switch languages within conversations, beginning a query in English, then continuing in their native language, or vice versa. Over half the world’s population is multilingual, and this cross-turn language switching reflects natural communication patterns. Yet how LLMs respond when users switch languages mid-conversation remains unexplored. A model could follow the query language, respecting the user’s immediate choice, or anchor to the established conversation language, maintaining consistency.

Prior work examines multilingual and multi-turn capabilities separately. Language confusion research focuses on single-turn interactions (Marchisio et al., 2025), multi-turn benchmarks operate monolingually (Zheng et al., 2023; Sirdeshmukh et al., 2025), and multilingual benchmarks keep each conversation in one language (He et al., 2024). Cross-turn language switching, where users change languages mid-conversation, remains untested.

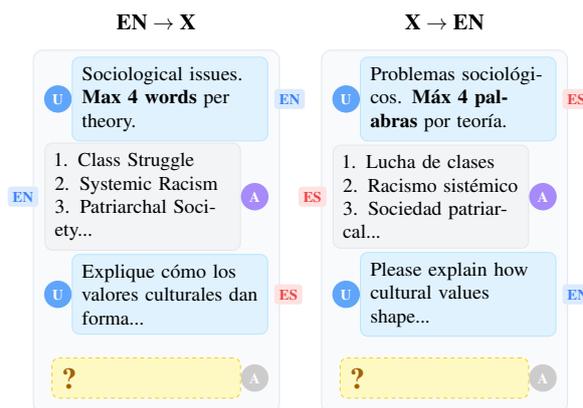


Figure 1: Same conversation in opposite switching directions.  $EN \rightarrow X$ : English context with Spanish query.  $X \rightarrow EN$ : Spanish context with English query. Both require the model to maintain the 4-word constraint while switching languages.

We address this gap by extending MultiChallenge to evaluate cross-turn language switching. We translate the INFERENCE\_MEMORY and INSTRUCTION\_RETENTION categories, which require cross-turn information retention, into German, Chinese, Spanish, and Arabic. We test four conditions: monolingual baselines in English and X, plus two switching conditions where the language changes at the user’s final turn (Figure 1). We evaluate five frontier models and measure two dimensions: **Language Fidelity** (does the model respond in the user’s language?) and **Task Accuracy** (does switching degrade performance?).

Our results reveal asymmetric behavior. When switching into a foreign language ( $EN \rightarrow X$ ), all models follow the query language (89–99%). When switching back to English ( $X \rightarrow EN$ ), models diverge: GPT-5 follows the query (>95%), while Claude Opus 4.5 and Command R+ continue in the foreign language (<8%). Task accuracy remains stable even as fidelity diverges, suggesting models comprehend cross-lingual input but apply different language policies. A simple system prompt

explicitly instructing query-language response fails to override these defaults, suggesting this behavior may not be easily modifiable at inference time.

These findings highlight the need for systematic research on user preferences for language continuity, standardized evaluation of cross-lingual behavior, and configurable language policies in multilingual deployment. We contribute: (1) the first systematic evaluation of cross-turn language switching in multi-turn conversations (182 conversations  $\times$  4 languages), (2) evidence that frontier models implement fundamentally different language policies, a divergence invisible to existing evaluations, and (3) demonstration that task accuracy remains stable despite divergent language behavior, isolating this as a policy rather than capability issue.

## 2 Related Work

Prior work has examined multilingual capabilities, multi-turn interactions, and code-switching separately. Yet no benchmark tests how models handle language switches across conversation turns.

**Language Confusion.** [Marchisio et al. \(2025\)](#) introduced the Language Confusion Benchmark, which measures whether models respond in the language they were prompted in. They found that Llama and Mistral models frequently respond in unintended languages, especially in cross-lingual settings. [Nie et al. \(2025\)](#) traced this to transition failures in final layers; [Lee et al. \(2025\)](#) mitigated it via preference tuning. However, all prior work focuses on single-turn interactions.

**Multi-Turn and Multilingual Benchmarks.** MT-Bench ([Zheng et al., 2023](#)) introduced multi-turn evaluation with open-ended conversational questions. MultiChallenge ([Sirdeshmukh et al., 2025](#)) tests more demanding scenarios such as retaining first-turn instructions or recalling user information across turns. Both operate monolingually. Multi-IF ([He et al., 2024](#)) extends instruction-following evaluation to 8 languages with 3-turn conversations, but each conversation remains in a single language throughout.

**Code-Switching and User Expectations.** Research on code-switching in NLP has primarily focused on intra-sentential mixing, where speakers blend languages within a single utterance. Benchmarks like LinCE ([Aguilar et al., 2020](#)) and GLUE-CoS ([Khanuja et al., 2020](#)) evaluate model performance on such naturally code-switched text. How-

ever, Human-Computer Interaction (HCI) research shows that multilingual users also switch languages across conversation turns and expect agents to accommodate this. [Choi et al. \(2023\)](#) found that code-mixing users feel excluded when conversational agents cannot handle their language practices. [Bawa et al. \(2020\)](#) showed that bilingual users prefer chatbots that reciprocate their language choices over multiple turns. Despite these documented user behaviors, no work evaluates how LLMs respond to cross-turn language switching.

## 3 Methodology

We design a controlled framework to evaluate language model behavior under mid-conversation language switches. Our approach comprises dataset construction through benchmark translation, experimental conditions isolating switch effects, and dual metrics assessing both language fidelity and task accuracy.

### 3.1 Dataset Construction

We extend MultiChallenge’s INFERENCE\_MEMORY (113 examples), which tests whether models recall and connect user information scattered across previous turns, and INSTRUCTION\_RETENTION (69 examples), which tests whether models follow first-turn instructions throughout the conversation. Conversations range from 3 to 19 turns (median: 7): short (3–5 turns,  $n=44$ ), medium (7–9 turns,  $n=99$ ), and long (11+ turns,  $n=39$ ).

We translate conversations into German, Chinese, Spanish, and Arabic using Google Translate, with human verification on sampled translations and GPT-4o-mini automated verification across all translations (Appendix A).

### 3.2 Experimental Conditions

Each MultiChallenge conversation contains multiple context turns followed by a final evaluation query. We test four conditions (Table 1), including monolingual baselines in English and X, plus two switching conditions where the language changes at the user’s final turn.

### 3.3 Models

We evaluate five frontier models: GPT-5 ([OpenAI, 2025](#)), Gemini 3 Pro ([Google, 2025](#)), Claude Opus 4.5 ([Anthropic, 2025](#)), DeepSeek-V3.1 ([DeepSeek-AI, 2025](#)), and Command R+ ([Cohere, 2024](#)). All models are accessed via API in non-thinking mode.

Condition	Description
Baseline (EN)	All turns in English (original)
Baseline (X)	All turns in X (translated)
EN→X	English context, final query in X
X→EN	X context, final query in English

Table 1: Experimental conditions where  $X \in \{DE, ZH, ES, AR\}$ . Context refers to all user–assistant turns preceding the final query.

Specific model versions and parameters are detailed in Appendix B.

### 3.4 Evaluation Metrics

We measure two dimensions: (1) **Language Fidelity**, the percentage of responses matching the query language, evaluated using GPT-4o-mini as judge; and (2) **Task Accuracy**, whether the response correctly addresses the task, evaluated using GPT-4o with MultiChallenge’s instance-level rubrics. Details of the evaluation prompts are provided in Appendix C.

## 4 Results

Following the methodology outlined above, we present our findings on language fidelity and task accuracy. Our results reveal some differences in how models handle mid-conversation language switches, suggesting distinct underlying response strategies.

### 4.1 Language Fidelity

Table 2 presents our main finding: models diverge dramatically in the X→EN condition (foreign context, English query).

Model	EN→X	X→EN	Behavior
GPT-5	98.6	<b>95.1</b>	Query-following
Gemini 3 Pro	98.3	73.8	Mixed
Claude Opus 4.5	96.1	7.7	Context-anchoring
DeepSeek-V3.1	88.3	51.9	Mixed
Command R+	89.3	<b>0.8</b>	Context-anchoring

Table 2: Language fidelity (%) by condition, averaged across languages. Bold indicates extreme values.

All models successfully follow the query language when switching into a foreign language (EN→X: 88–99%). However, when switching back to English after foreign context (X→EN), models split into three groups:

**Query-following** GPT-5 responds in the query language regardless of context (95.1% English).

**Context-anchoring** Claude Opus 4.5 and Command R+ continue in the context language, largely ignoring the language switch (0.8–7.7% English).

**Mixed** Gemini 3 Pro and DeepSeek-V3.1 show intermediate behavior, balancing query and context influence (51.9–73.8%).

These reflect different design choices rather than performance differences.

### 4.2 Per-Language Analysis

Table 3 breaks down X→EN fidelity by source language.

Model	DE	ZH	ES	AR	Avg
GPT-5	94.0	95.6	94.5	96.2	95.1
Gemini 3 Pro	78.6	72.5	74.7	69.2	73.8
Claude Opus 4.5	10.4	9.9	6.0	4.4	7.7
DeepSeek-V3.1	41.8	60.4	41.2	64.3	51.9
Command R+	1.1	1.1	0.5	0.5	0.8

Table 3: X→EN fidelity (%) by source language.

GPT-5 maintains consistent fidelity across languages (94–96%). DeepSeek-V3.1 shows notable variation: higher fidelity when switching from Chinese (60.4%) and Arabic (64.3%) compared to German (41.8%) and Spanish (41.2%), possibly reflecting training data composition. Context-anchoring models (Claude, Command R+) show uniformly low fidelity regardless of source language.

### 4.3 Conversation Length Effect

Table 4 analyzes whether context-anchoring intensifies with conversation length. We use chi-square tests to assess whether fidelity rates differ significantly across length categories.

Model	Short	Med	Long	$p$
GPT-5	97.2	93.9	95.5	0.25
Gemini 3 Pro	82.4	75.0	60.9	<0.001
Claude Opus 4.5	11.9	5.8	7.7	0.04
DeepSeek-V3.1	55.1	54.0	42.9	0.04
Command R+	0.0	0.5	2.6	0.02

Table 4: X→EN fidelity (%) by conversation length. Short: 3–5 turns (n=44), Medium: 7–9 turns (n=99), Long: 11+ turns (n=39).  $p$ -values from  $\chi^2$  tests.

Gemini 3 Pro shows significant degradation with length, dropping from 82.4% to 60.9% ( $p < 0.001$ ). GPT-5 remains stable across all lengths ( $p=0.25$ , not significant). Context-anchoring models show floor effects, with fidelity already near zero, leaving no room for further decline.

## 4.4 Task Accuracy

Table 5 shows task accuracy remains stable across conditions, with no significant degradation from language switching.

Model	Base EN	Base X	EN→X	X→EN
GPT-5	57.1	58.7	59.2	52.5
Gemini 3 Pro	71.4	70.1	70.9	70.4
Claude Opus 4.5	54.4	48.5	48.9	49.9
DeepSeek-V3.1	50.0	40.3	42.9	37.7
Command R+	15.9	11.3	15.0	11.4

Table 5: Task accuracy (%) by condition, averaged across languages.

This null result is informative as language switching does not impair task performance. The challenge is behavioral (which language to use), not comprehension (understanding the task). Full results are provided in Appendix D.

## 5 Analysis

To further understand model behavior, we conduct additional analyses examining non-English language pairs and the influence of explicit system prompt instructions on language fidelity.

### 5.1 Non-English Switching (X→Y)

Table 6 presents results for switching between non-English languages, examining whether English has a privileged role. We select four pairs representing diverse script combinations.

Model	Metric	ZH→DE	DE→ZH	ES→AR	AR→ES
GPT-5	Fidelity	96.2	97.3	96.2	98.4
	Accuracy	53.8	52.2	56.6	51.6
Claude Opus 4.5 <sup>†</sup>	Fidelity	64.3	35.2	81.3	19.2
	Accuracy	47.3	46.2	51.1	48.9

Table 6: Cross-lingual transfer (X→Y) results (%). <sup>†</sup>One empty response in ZH→DE (n=181).

GPT-5 maintains high fidelity (>96%) across all X→Y pairs, consistent with its query-following behavior. Claude Opus 4.5 shows variable fidelity (19–81%), suggesting language-specific biases beyond simple context-anchoring. Task accuracy remains stable for both models, reinforcing that language switching does not impair performance.

### 5.2 System Prompt Ablation

Table 7 presents results testing whether explicit instructions (“Always respond in the language of the user’s most recent message”) can override default behavior.

Model	Condition	None	Explicit
GPT-5	X→EN	95.1	94.0
Claude 4.5	X→EN	7.7	7.1
Command R+	X→EN	0.8	0.7

Table 7: X→EN fidelity (%) with and without explicit language instructions. Full results in Appendix E.

Explicit system prompts have minimal effect (<1.5 pp difference), suggesting language behavior is deeply embedded and not easily overridden by simple instructions.

## 6 Discussion

**Implications for Multilingual UX.** Our findings reveal a design tension. Context-anchoring may frustrate users who intentionally switch languages, while query-following may feel inconsistent to users who prefer conversational continuity. The appropriate behavior depends on user intent, whether the switch was deliberate or incidental, which current models cannot infer.

**Why the Divergence?** We hypothesize the divergence stems from different training objectives. Context-anchoring may result from training models to maintain consistency and avoid sycophancy (following user requests unconditionally). Query-following may prioritize immediate user intent. Our system prompt ablation (§5.2) shows that explicit instructions fail to override these defaults, though the underlying mechanism remains unclear. Future work using mechanistic interpretability or reasoning trace analysis could help identify where and how language selection occurs during generation.

**Future Directions.** Several questions remain open. First, at what point does accumulated foreign context override query-language signals? Varying switch points within conversations may reveal thresholds for context-anchoring. Second, tasks involving culture-specific knowledge may show accuracy degradation unlike our general reasoning tasks. Third, user preference studies could determine whether multilingual users prefer query-following or context-anchoring behavior in practice.

## 7 Conclusion

We evaluate cross-turn language switching in multi-turn conversations, revealing that frontier models

diverge in language policy: GPT-5 follows user language switches (>95%), while Claude Opus 4.5 and Command R+ anchor to conversation context (<8%). Task accuracy remains stable across conditions, indicating the challenge is behavioral rather than cognitive. These findings have implications for multilingual deployment and model selection.

## Limitations

Our study has several limitations. First, we use pre-filled conversation histories following standard practice (Sirdeshmukh et al., 2025; Zheng et al., 2023), which may not fully reflect live interaction dynamics. Second, translation quality, while verified, may introduce artifacts. Third, we test four high-resource languages; low-resource languages may reveal different patterns. Fourth, task accuracy shows moderate cross-run variance (std up to 3 pp; Tables 11 and 12), so absolute values should be interpreted cautiously. Fifth, comprehensive consistency analysis is limited to two models (GPT-5, Gemini 3 Pro) on one condition (ES→EN) due to computational constraints.

## Ethics Statement

This work uses publicly available benchmark data (MultiChallenge) and commercially available language models accessed through their official APIs. Human verification of translations was conducted by native speakers who participated voluntarily. The research does not involve sensitive or private data, and we do not foresee direct negative societal impacts from this work.

## References

- Gustavo Aguilar, Sudipta Kar, and Thamar Solorio. 2020. [LinCE: A centralized benchmark for linguistic code-switching evaluation](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1803–1813, Marseille, France. European Language Resources Association.
- Anthropic. 2025. [Introducing claude opus 4.5](https://www.anthropic.com/news/claude-opus-4-5). <https://www.anthropic.com/news/claude-opus-4-5>.
- Anshul Bawa, Pranav Khadpe, Pratik Joshi, Kalika Bali, and Monojit Choudhury. 2020. [Do multilingual users prefer chat-bots that code-mix? let’s nudge and find out!](#) *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW1).
- {Yunjae J.} Choi, Minha Lee, and Sangsu Lee. 2023. [Toward a multilingual conversational agent: Challenges and expectations of code-mixing multilingual users](#). In *CHI 2023 - Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Conference on Human Factors in Computing Systems - Proceedings. Association for Computing Machinery. Publisher Copyright: © 2023 ACM.; 2023 CHI Conference on Human Factors in Computing Systems, CHI 2023 ; Conference date: 23-04-2023 Through 28-04-2023.
- Cohere. 2024. [Command r+](https://docs.cohere.com/docs/command-r-plus). <https://docs.cohere.com/docs/command-r-plus>.
- DeepSeek-AI. 2025. [Deepseek-v3.1](https://huggingface.co/deepseek-ai/DeepSeek-V3.1). <https://huggingface.co/deepseek-ai/DeepSeek-V3.1>.
- Google. 2025. [Gemini 3: Introducing the latest gemini ai model from google](https://blog.google/products/gemini/gemini-3/). <https://blog.google/products/gemini/gemini-3/>.
- Yun He, Di Jin, Chaoqi Wang, Chloe Bi, Karishma Mandyam, Hejia Zhang, Chen Zhu, Ning Li, Tengyu Xu, Hongjiang Lv, Shruti Bhosale, Chenguang Zhu, Karthik Abinav Sankararaman, Eryk Helenowski, Melanie Kambadur, Aditya Tayade, Hao Ma, Han Fang, and Sinong Wang. 2024. [Multi-if: Benchmarking llms on multi-turn and multilingual instructions following](#). *Preprint*, arXiv:2410.15553.
- Simran Khanuja, Sandipan Dandapat, Anirudh Srinivasan, Sunayana Sitaram, and Monojit Choudhury. 2020. [GLUECoS: An evaluation benchmark for code-switched NLP](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3575–3585, Online. Association for Computational Linguistics.
- Nahyun Lee, Yeongseo Woo, Hyunwoo Ko, and Guijin Son. 2025. [Controlling language confusion in multilingual LLMs](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 1026–1035, Vienna, Austria. Association for Computational Linguistics.
- Kelly Marchisio, Wei-Yin Ko, Alexandre Bérard, Théo Dehaze, and Sebastian Ruder. 2025. [Understanding and mitigating language confusion in llms](#). *Preprint*, arXiv:2406.20052.
- Ercong Nie, Helmut Schmid, and Hinrich Schütze. 2025. [Mechanistic understanding and mitigation of language confusion in english-centric large language models](#). *Preprint*, arXiv:2505.16538.
- OpenAI. 2025. [Introducing gpt-5](https://openai.com/index/introducing-gpt-5/). <https://openai.com/index/introducing-gpt-5/>.
- Ved Sirdeshmukh, Kaustubh Deshpande, Johannes Mols, Lifeng Jin, Ed-Yeremai Cardona, Dean Lee, Jeremy Kritz, Willow Primack, Summer Yue, and Chen Xing. 2025. [Multichallenge: A realistic multi-turn conversation evaluation benchmark challenging to frontier llms](#). *Preprint*, arXiv:2501.17399.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. *Judging llm-as-a-judge with mt-bench and chatbot arena*. Preprint, arXiv:2306.05685.

## A Translation and Verification Details

### A.1 Translation Pipeline

We translate all 182 conversations from English into German (DE), Chinese (ZH), Spanish (ES), and Arabic (AR) using Google Translate. Each conversation was translated in full, preserving turn structure and speaker labels.

### A.2 Human Verification

To assess translation quality, native speakers verified 10 randomly sampled translations each for Spanish and Chinese (Table 8), scoring on semantic accuracy, completeness, and overall quality (1–5 scale).

Common issues identified in Spanish included literal translations of idiomatic expressions (e.g., “grounding techniques” → “técnicas de conexión a tierra” instead of the more natural “técnicas de anclaje”). For Chinese, issues included mistranslation of domain-specific terms and inconsistent handling of proper nouns. These errors informed our automated verification prompt but were not individually corrected in the final dataset. Despite these minor issues, both languages achieved perfect completeness scores, indicating no information loss.

Language	Semantic	Complete	Overall	<i>n</i>
Spanish	4.1	5.0	4.4	10
Chinese	4.9	5.0	4.8	10

Table 8: Human verification scores (1–5 scale).

### A.3 Automated Verification and Correction

Based on error patterns from human verification, we develop a GPT-4o-mini prompt to verify and correct task-critical elements across all translations. German and Arabic translations were verified and corrected using this automated approach only. Table 9 summarizes results by language.

All experiments use the corrected translations from both human and automated verification.

Language	Turns	Corrected	Rate
German	1,460	265	18.2%
Chinese	1,460	254	17.4%
Spanish	1,460	257	17.6%
Arabic	1,460	271	18.6%
<b>Total</b>	<b>5,840</b>	<b>1,047</b>	<b>17.9%</b>

Table 9: GPT-4o-mini automated verification results (turn-level).

### A.4 Verification Prompt

#### System Prompt

You are a translation quality assessor. Compare an original English conversation with its translation and assess whether the translation accurately conveys the original meaning. **CRITICAL:** The translation must match exactly what the original says, word-for-word. Do NOT correct inconsistencies or errors in the original text — translate them literally as written.

#### User Prompt Template

Compare the original English conversation with its {target\_language} translation. Assess whether each turn accurately conveys the original meaning. Focus on:

- Entity accuracy (names, places, activities, objects)
- Numerical accuracy (dates, times, quantities)
- Constraint compliance (formatting requirements)
- Semantic fidelity (meaning preservation)

Output JSON:

```
{
  "accurate": <true|false>,
  "issues": [{"turn": <n>, "description":
"<what is wrong>",
  "original": "<text>", "translated":
"<text>", "corrected": "<text>"}],
  "corrected_conversation": [{"role":
"user", "content": "<content>"}, ...]
}
```

The corrected\_conversation must have exactly {turn\_count} turns. If accurate, corrected\_conversation should be a copy of the translated conversation. ORIGINAL (English): {original\_conversation} TRANSLATED ({target\_language}): {translated\_conversation}

## B Model Settings

Table 10 lists the specific model versions and API parameters used in our experiments.

GPT-5 and Gemini 3 Pro do not support custom temperature values; we use their defaults (temperature = 1). For other models, we set temperature to 0. Variance checks on ES→EN (n=182, 3

Model	Version	Temperature
GPT-5	gpt-5-2025-08-07	1 (default)
Gemini 3 Pro	gemini-3.0-pro-preview-2025-11-18	1 (default)
Claude Opus 4.5	claude-opus-4-5-20251101	0
DeepSeek-V3.1	deepseek-v3-1-250821	0
Command R+	command-r-plus-08-2024	0

Table 10: Model versions and API settings. GPT-5 and Gemini 3 Pro do not support custom temperature values for reasoning tasks.

runs) for GPT-5 and Gemini 3 Pro are shown in Tables 11 and 12. Language fidelity is stable (std < 1 pp), while task accuracy shows moderate variance (std up to 3 pp), reflecting model stochasticity and LLM-based evaluation noise. Comprehensive consistency analysis across all models and conditions remains future work.

Model	Mean (%)	Std	$\kappa$	3-Way (%)
GPT-5	94.9	0.52	0.74	96.2
Gemini 3 Pro	74.5	0.26	0.64	79.7

Table 11: Language fidelity consistency (ES→EN, 3 runs).  $\kappa$  = Cohen’s kappa (mean pairwise).

Model	Mean (%)	Std	$\kappa$	3-Way (%)
GPT-5	50.2	0.52	0.55	66.5
Gemini 3 Pro	71.1	2.99	0.64	77.5

Table 12: Task accuracy consistency (ES→EN, 3 runs).  $\kappa$  = Cohen’s kappa (mean pairwise).

## C Evaluation Details

### C.1 Language Fidelity

We use GPT-4o-mini (gpt-4o-mini) as a language identification judge. Given a model response and the expected target language, the judge determines whether the response is written in the expected language.

Language Fidelity Prompt
<p>Is this text written in {expected_lang_name}?</p> <p>&lt;TEXT&gt;</p> <p>{response}</p> <p>&lt;/TEXT&gt;</p> <p>Rules:</p> <ul style="list-style-type: none"> <li>- Words like “No”, “OK” exist in multiple languages — consider them valid in the expected language</li> <li>- Ignore code snippets, URLs, or proper nouns</li> <li>- Focus on the main content language</li> </ul> <p>Return ONLY: YES or NO</p> <p>Your answer:</p>

### C.2 Task Accuracy

We use GPT-4o (gpt-4o-2024-08-06) as judge following MultiChallenge’s LLM-as-judge protocol with instance-level rubrics. Each test instance includes a task-specific rubric that defines correctness criteria. The judge receives the conversation context, model response, and rubric, then outputs a binary correctness judgment. This approach achieves 93% alignment with human raters.

Task Accuracy Prompt
<p>You are tasked with evaluating a model response to see if it meets a specific criteria. The criteria will always be YES/NO evaluation. The model response is as follows:</p> <p>&lt;MODEL_RESPONSE&gt;</p> <p>{response}</p> <p>&lt;/MODEL_RESPONSE&gt;</p> <p>The criteria that the model response must meet is as follows. Be VERY STRICT!:</p> <p>&lt;CRITERIA&gt;</p> <p>{target_question}</p> <p>&lt;/CRITERIA&gt;</p> <p>Print your reasoning followed by your verdict, either “YES” or “NO”.</p>

## D Full Results

Tables 13 and 14 report language fidelity and task accuracy across all conditions and languages. Empty responses were excluded: Gemini 3 Pro (8), Claude Opus 4.5 (3).

## E System Prompt Ablation

Tables 15 and 16 report full results for the system prompt ablation. Neither language fidelity nor task accuracy changes meaningfully with explicit instructions.

Model	Base EN	Baseline X				EN→X				X→EN			
		DE	ZH	ES	AR	DE	ZH	ES	AR	DE	ZH	ES	AR
GPT-5	100.0	98.9	100.0	100.0	99.5	97.8	99.5	99.5	97.8	94.0	95.6	94.5	96.2
Gemini 3 Pro	100.0	98.9	100.0	100.0	99.5	98.3	98.9	98.4	97.8	78.6	72.5	74.7	69.2
Claude Opus 4.5	100.0	98.9	100.0	100.0	99.5	96.7	94.0	97.3	96.7	10.4	9.9	6.0	4.4
DeepSeek-V3.1	100.0	98.9	98.4	100.0	98.9	93.4	73.1	95.1	91.8	41.8	60.4	41.2	64.3
Command R+	100.0	98.9	100.0	100.0	99.5	91.8	89.0	95.6	80.8	1.1	1.1	0.5	0.5

Table 13: Language fidelity (%).

Model	Base EN	Baseline X				EN→X				X→EN			
		DE	ZH	ES	AR	DE	ZH	ES	AR	DE	ZH	ES	AR
GPT-5	57.1	58.2	57.7	57.7	61.0	57.1	59.9	59.3	60.4	55.5	50.5	49.5	54.4
Gemini 3 Pro	71.4	66.5	72.0	71.4	70.3	73.6	70.3	68.7	70.9	66.5	68.7	72.0	74.2
Claude Opus 4.5	54.4	45.1	48.9	52.7	47.3	49.5	46.7	50.5	48.9	48.4	47.8	52.7	50.5
DeepSeek-V3.1	50.0	39.0	39.0	45.1	37.9	40.1	44.5	44.0	42.9	38.5	37.4	37.9	36.8
Command R+	15.9	11.5	9.3	9.9	14.3	15.4	13.2	15.4	15.9	12.1	11.0	11.5	11.0

Table 14: Task accuracy (%).

Model	Cond.	Prompt	DE	ZH	ES	AR
GPT-5	EN→X	None	97.8	99.5	99.5	97.8
	EN→X	Explicit	98.4	98.9	99.5	98.9
	X→EN	None	94.0	95.6	94.5	96.2
	X→EN	Explicit	94.0	93.4	94.0	94.5
Claude 4.5	EN→X	None	96.7	94.0	97.3	96.7
	EN→X	Explicit	96.7	94.5	97.3	96.2
	X→EN	None	10.4	9.9	6.0	4.4
	X→EN	Explicit	9.9	9.9	5.5	3.3
Command R+	EN→X	None	91.8	89.0	95.6	80.8
	EN→X	Explicit	91.8	87.4	96.7	83.0
	X→EN	None	1.1	1.1	0.5	0.5
	X→EN	Explicit	1.1	0.5	0.5	0.5

Table 15: Language fidelity (%) by condition and system prompt.

Model	Cond.	Prompt	DE	ZH	ES	AR
GPT-5	EN→X	None	57.1	59.9	59.3	60.4
	EN→X	Explicit	58.2	59.9	58.8	56.0
	X→EN	None	55.5	50.5	49.5	54.4
	X→EN	Explicit	54.9	57.7	53.8	53.8
Claude 4.5	EN→X	None	49.5	46.7	50.5	48.9
	EN→X	Explicit	48.4	52.7	55.5	54.4
	X→EN	None	48.4	47.8	52.7	50.5
	X→EN	Explicit	48.4	51.1	51.6	48.9
Command R+	EN→X	None	15.4	13.2	15.4	15.9
	EN→X	Explicit	14.8	17.6	14.8	15.4
	X→EN	None	12.1	11.0	11.5	11.0
	X→EN	Explicit	11.5	12.6	12.6	11.0

Table 16: Task accuracy (%) by condition and system prompt.

# Generating Difficult-to-Translate Texts

Vilém Zouhar<sup>★</sup> Wenda Xu<sup>■</sup> Parker Riley<sup>■</sup> Juraj Juraska<sup>■</sup>  
Mara Finkelstein<sup>■</sup> Markus Freitag<sup>■</sup> Daniel Deutsch<sup>■</sup>  
<sup>■</sup>Google <sup>★</sup>ETH Zurich

## Abstract

Machine translation benchmarks sourced from the real world are quickly obsoleted, due to most examples being easy for state-of-the-art translation models. This limits the benchmark’s ability to distinguish which model is better or to reveal models’ weaknesses. Current methods for creating difficult test cases, such as subsampling or from-scratch synthesis, either fall short of identifying difficult examples or suffer from a lack of diversity and naturalness. Inspired by the iterative process of human experts probing for model failures, we propose MT-breaker, a method where a large language model iteratively refines a source text to increase its translation difficulty. The LLM iteratively queries a target machine translation model to guide its generation of difficult examples. Our approach generates examples that are more challenging for the target MT model while preserving the diversity of natural texts. While the examples are tailored to a particular machine translation model during the generation, the difficulty also transfers to other models and languages.

## 1 Introduction

Test sets need to be difficult in order to be informative. If an NLP model performs flawlessly on a test set, then we learn very little about the model’s shortcomings. Or, if we compare two NLP models that perform near-perfectly on a test set, we do not know which one is better. For machine translation, current models are able to translate even recent benchmarks with only a few mistakes across several language directions (Kocmi et al., 2024a). Attempts have been made to automatically select a subset of existing datasets that still pose challenges to the models (Proietti et al., 2025). However, these methods fall short due to the absence of a sufficient number of naturally occurring challenging examples. Another line of work makes use of LLMs to

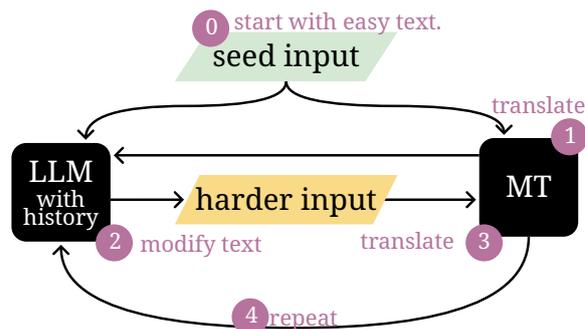


Figure 1: High-level overview of our interactive difficulty-to-translate text generation pipeline. The process is repeated for a fixed number of iterations.

create difficult examples from scratch in a zero-shot manner (Pombal et al., 2025), which we show is inadequate due to its lack of difficulty and diversity.

A human practitioner who is trying to assess the quality of a model approaches this task differently: not by selecting difficult examples or synthesizing them from scratch (finding based on interviews in Section 3.7). Instead, the practitioner iteratively tries to find weaknesses in the model. They might start with some random inputs and then, based on the model’s response, modify these inputs to cause the model to provide a faulty output. This way, the practitioner arrives at difficult examples tailored to a particular model that reveal its shortcomings. These insights can be then used to diagnose and improve the model.

In this work we propose MT-breaker, which uses an LLM to generate difficult-to-translate examples that mimic the process of the human practitioner (see Figure 1). We start with a seed text ① that the machine translation model translates ②. The text and its translation is then passed to an LLM which modifies it ③ and passes it to the MT model again ④. The translation is again passed back to the LLM model with chat-like history and the whole process repeats ⑤.

<sup>★</sup>Work done during Google Internship.  
Correspondence: vzouhar@ethz.ch

Zeroshot (min)	Seeds	MTbreaker (seeded)
It is what it is.	Another one has been found!	Another rabbit hole of lies was found!
The quick brown fox jumps over the lazy dog.	I quit FB, removed all Meta and Pinterest tracker from my website and killed my private WhatsApp account.	I've quit FB and killed the private WhatsApp account to get rid of Meta's and Pinterest's gross ad-tracking tools that I hated with a fiery passion.
He bade them farewell and then he bade them all to hell.	Going back up tomorrow and we're doing stalls and slow flight.	Going back up tomorrow; we're doing stalls and spins, and then some unusual attitudes.
It is what it is.	Heheh not one but three!	Heheh not one, two, but three!!!1!1!oneeleven

Table 1: Example of generated sources in English by Zeroshot, Seeds, and MT-breaker (seeded). Examples were selected to be short. See Appendix Table 7 for more examples.

---

**MT-breaker**(seed text  $s_0$ , steps  $N$ ):

- 1: **for**  $i = 1 \dots N$  **do**
- 2:    $t_{i-1} \leftarrow \text{MT}(s_{i-1})$                                     $\triangleright$  Translate
- 3:    $s_i \leftarrow \text{LLM}_{\text{step}}(\{s_k\}_{k=0}^{i-1}, \{t\}_{k=0}^{i-1})$     $\triangleright$  Generate next
- 4: **return**  $\arg \min \text{QE}(\{s_k\}_{k=0}^N, \{t\}_{k=0}^N)$     $\triangleright$  Pick difficult

---

Algorithm 1: Given a source text, the LLM receives the MT model’s translation which informs how to next change the text. The  $\text{LLM}_{\text{step}}$  can be instructed to not stray too far from the original meaning or to preserve naturalness. At the end, we pick the source text from  $s$  that led to the worst translation.

We show that this approach maintains the diversity and naturalness of the source examples, while vastly increasing the difficulty of the generated examples. Despite the approach relying on a quality estimation metric, the same findings are confirmed by an expert MQM human-evaluation study. The examples are tailored to be difficult for a particular machine translation model, though we also show that the difficulty also transfers to other models and languages. The generated dataset can be used to find weaknesses of a particular model, which can, in turn, be used for hillclimbing during model development.

## 2 Methods

We now describe the various approaches of generating difficult-to-translate texts. The baseline, which we refer to as **seeds**, consists of simply taking existing texts from a dataset; i.e. not creating new ones.

**Zero-shot benchmarking.** This method simply prompts a large language model to generate an example to be part of a benchmark. This generation can be steered towards particular topics, domains, and also difficulty. This method is a simplified version of Pombal et al. (2025). We introduce a variant “(history)”, where we include the history of previously generated texts in the prompt such that they are not repeated.

**MT-breaker.** Our method is shown in Figure 1 and Algorithm 1. It is based on mimicking a human expert that tries to find an input that triggers an error in the machine translation model, which is a strategy motivated by qualitative interviews with human experts (Section 3.7). We start with a piece of text, called a seed, which is translated by a machine translation model. This, together with the seed is the input to  $\text{LLM}_{\text{step}}$ , which modifies the seed. This is repeated for a certain number of steps, e.g. 10. At the end, we compute the quality estimation scores for all steps and select the most difficult source. A quality estimation metric is a model that takes in the source and its translation and outputs a scalar assessment of the translation’s quality (Freitag et al., 2024). We also consider MT-breaker without the seed text “(seedless)”, and a version which has the  $\text{LLM}_{\text{step}}$  seeing the intermediate quality estimation scores “(seeded+qe)”. The prompts for  $\text{LLM}_{\text{step}}$  in all variants are shown in Appendix B.

## 3 Experiments and Results

We now describe experiments in which we compare and analyze the above listed methods for generating difficult-to-translate texts.

**Setup.** For brevity and computational feasibility, we consider English  $\rightarrow$  {Czech, German, Chinese, Vietnamese, Polish} language directions. For seeds, we use 100 randomly sampled English sources from WMT 2024 (Kocmi et al., 2024a). We use the following publicly available machine translation models:

- Google Translate<sup>1</sup>
- Gemini 2.5-flash (Comanici et al., 2025)
- Gemini 2.0-flash (Comanici et al., 2025)
- Gemma 3-27b-it (Gemma Team et al., 2025)

<sup>1</sup>translate.google.com, accessed August 2025.

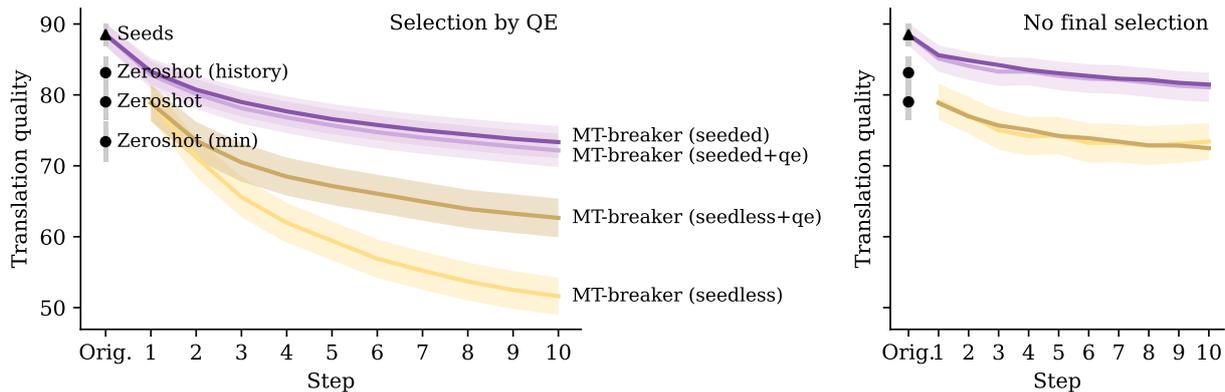


Figure 2: Development of difficulty with each iteration. Seeds require no iterations, Zeroshot and Zeroshot (history) require one LLM step, and the rest requires 10 steps. Right plot shows running hypotheses at a given step without final selection by quality estimation. This does not include Zeroshot (min) which is simply Zeroshot with selection. Values averaged across all models and languages. Shaded areas show 90% t-test confidence intervals averaged across all models and languages.

For the  $LLM_{step}$  we use Gemini 2.5-pro (Comanici et al., 2025) in the main paper, though explore others in Section 3.4. For measuring the quality of the translations, we use the average of MetricX-24-QE (Juraska et al., 2024) and prompted LLM-as-QE (see Appendix B for details). We also include a source-only version of MetricX, that only sees the source texts and no translation, which serves as a proxy for general difficulty (see Appendix A). All three are scaled from 0 (poor translation, high difficulty) to 100 (perfect translation, low difficulty). Same as Proietti et al. (2025), the quantity of “difficulty” has an opposite direction to quality estimation (100-QE). While we use the term “difficulty”, we include the original quality estimation scores in tables and figures for simplicity.

We compare the following methods for obtaining a difficult test set:

- “Seed” segments from WMT
- “Zeroshot” benchmarking
- “Zeroshot (history)” benchmarking with history
- “Zeroshot (min)” benchmarking with selection<sup>2</sup>
- “MT-breaker (seeded)” with seeds
- “MT-breaker (seedless)” without seeds
- “MT-breaker (seeded+qe)” with seeds and with quality estimation
- “MT-breaker (seedless+qe)” without seeds and with quality estimation

For a fair comparison, we instruct the generation of the Zeroshot benchmarking and MT-breaker with-

<sup>2</sup>For a fair comparison, we also include a version of Zeroshot that is run 10x (same number of steps as MT-breaker) out of which (per batch of 10) we select the most difficult sources with quality estimation in the same manner.

out seeds methods to have approximately the same number of words as the seeds used by the other methods.

See generated examples of Zeroshot (min) and MT-breaker (seeded) in Table 1 and Appendix Tables 7 to 9.

### 3.1 Increasing Difficulty

As shown in Figure 2, the translation quality of texts generated by MT-breaker methods decreases with each step but begins to plateau around step 10. Interestingly, having access to the quality estimation during the breaking process does not help the MT-breaker (seeded) and worsens MT-breaker (seedless). We explain this through different variances. The per-step averages (Figure 2 right) are the same for the seedless version. The reason for the lower minimum (shown cumulatively in the lines in Figure 2 left) is that the version with quality estimation has lower variance ( $\sigma^2=297$ ) than the version without it ( $\sigma^2=325$ ). This is likely caused by the model using most of the 10 steps to continue the breakage with the text with currently lowest quality estimation score, which can be a dead-end. The minimum of a variable with higher variance is lower than the minimum of a variable with lower variance, even when the means are the same.

Methods that use seeds for the initial source (Seeds, MT-breaker seeded) lead to lower difficulty than those without such restrictions (Zeroshot, MT-breaker seedless). In the next section, we show that this is part of a diversity-difficulty tradeoff.

	Seed	Zeroshot	Zeroshot (min)	Zeroshot (history)	MT- breaker (seedless)	MT- breaker (seeded)	MT- breaker (seedless+QE)	MT- breaker (seeded+QE)
<i>Diversity (higher is better)</i>								
Diversity (embd)	0.33	0.09	0.10	0.31	0.16	0.34	0.12	0.34
Diversity (chrF)	0.76	0.24	0.29	0.73	0.46	0.77	0.38	0.77
Diversity (topics)	247	126	131	225	148	262	147	262
Diversity (errors)	23.50	20.80	22.00	20.65	22.95	23.80	22.65	23.90
Diversity (words)	1566	1121	1177	1508	1532	2082	1527	2091
<i>Complexity &amp; Style</i>								
Grammaticality	92.28	96.16	95.67	97.91	94.15	87.65	92.27	87.93
Naturalness	92.91	57.49	54.15	73.89	29.80	82.71	41.99	83.31
Word Rarity	17.67	46.24	46.14	28.45	54.80	35.69	53.23	34.77
Syntax Complexity	39.95	55.44	56.28	57.12	70.38	51.21	65.78	50.23
Avg. Word Count	33.25	31.21	31.25	33.34	31.32	39.62	37.67	40.18
Avg. Word Length	4.69	5.44	5.43	4.93	5.90	5.07	5.66	5.01
<i>Quality Estimation (lower is better)</i>								
QE (Gemini)	89.78	81.94	73.94	83.53	45.35	67.52	59.97	69.23
QE (MetricX)	87.25	76.09	72.87	82.73	57.88	76.81	65.33	77.45
SRCQE (MetricX)	94.47	91.69	91.55	93.29	90.54	91.96	90.12	92.01

Table 2: Quantitative automatic evaluation of difficult data generation approaches.

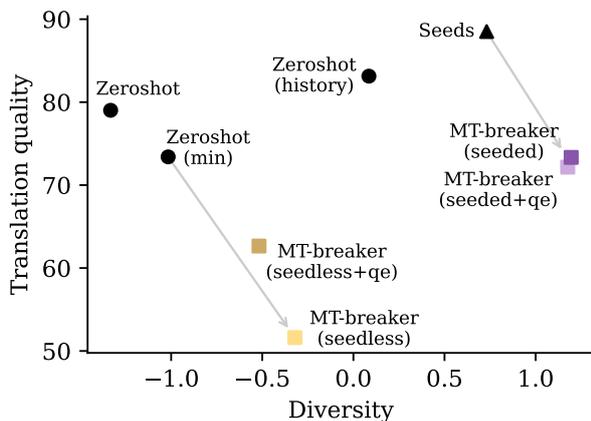


Figure 3: Comparison of difficulty and diversity of difficult-to-translate text generation methods. Diversity is z-normalized across each diversity measure and averaged. Gray arrows signify most fair comparison.

### 3.2 Data Quality

So far we have investigated only the method’s capability of breaking the machine translation model. We now confirm the usability of the method for generating meaningful datasets. Due to the fact that there is no accepted way of intrinsically evaluating a dataset, we propose to examine several different statistics about the dataset. Prior works either do not investigate the data quality (Pombal et al., 2025), rely on guarantees of the data creation process, such as grammatical or human edits (Isabelle et al., 2017; Manakhimova et al., 2023), or resort to manual human inspection (Amrhein et al., 2022; Ármannsson et al., 2024).

For a generated set of examples, we measure the following variables:

- Diversity (in embedding space, pairwise chrF Popović, 2015,<sup>3</sup> unique topics, unique words), to make sure that the texts does not collapse into one subtopic.
- Diversity (errors), to make sure that the error mode is not simply being repeated.
- Source and word length, to make sure the difficulty is not being trivially induced by extremely long segments.
- Word rarity, to analyze if the difficulty stems from simply using rare words.
- Syntax complexity, to analyze if the difficulty is due to convoluted sentence structures.
- Grammaticality, to make sure the text is well-formed and not broken.
- Naturalness, to make sure the text sounds like it could have been written by a human.

See Appendix B for details on how some of these variables are measured using a prompted LLM.

The results in Table 2 show that methods without an explicit treatment for diversity (Zeroshot, Zeroshot min, MT-breaker seedless) indeed lack in the diversity, though this effect is most prevalent in Zeroshot. On the other hand, the methods without diversity guardrails lead to texts of higher difficulty. This tradeoff between diversity and difficulty is shown in Figure 3. If only the difficulty is the concern and not the diversity, MT-breaker seedless is

<sup>3</sup>ref:1|case:mix|eff:y|nc:6|nw:2|sp:no|v:2.5.1

		Target model (translation quality)				Target model (diversity)			
		Translate	Gemma3	Gemini2.0	Gemini2.5	Translate	Gemma3	Gemini2.0	Gemini2.5
MT-break. (seedless)	Gemma 3	48.00	57.31	59.77	58.99	0.05	0.19	0.16	0.06
	Gemini 2.0	35.92	42.36	42.97	39.79	-1.03	-0.72	-0.95	-1.09
	Gemini 2.5	52.56	60.51	63.12	62.18	0.48	0.64	0.57	0.65
	Gemini 2.5 pro	44.13	52.45	55.91	53.98	-0.72	-0.31	-0.30	-0.41
Zeroshot (min)	Gemma-3	53.00	68.28	66.31	66.75	-1.00	-0.89	-1.11	-1.23
	Gemini 2.0	45.15	54.34	51.63	55.67	-1.62	-1.60	-1.72	-1.55
	Gemini 2.5	56.68	63.62	65.81	64.31	0.02	0.18	-0.13	0.08
	Gemini 2.5 pro	71.78	72.63	74.59	74.61	-1.03	-0.85	-1.23	-1.05
MT-break. (seeded)	Gemma-3	53.69	63.22	65.05	63.68	0.89	0.85	0.75	0.74
	Gemini 2.0	41.67	52.59	53.56	52.46	1.13	0.71	0.66	0.53
	Gemini 2.5	59.86	68.46	73.11	69.77	1.37	1.34	1.09	1.11
	Gemini 2.5 pro	65.01	74.09	74.47	75.09	1.05	0.90	0.86	0.95
Seeds		85.63	89.26	89.94	89.23	0.76	0.74	0.44	0.60

Table 3: Results (average quality estimation scores) for using different LLMs (rows) as part of MT-breaker. Left side shows translation quality (lower is better) and right side shows average of z-normalized diversity measures. Averaged across all languages.

an apt choice. If, however, one desires to stay close to some existing dataset in terms of diversity but needs more difficulty, MT-breaker seeded provides a balance between the two.

### 3.3 Transfer Between MT Models

When MT-breaker targets a specific model, that model’s performance is most significantly impacted, indicated by the lowest scores on the diagonal in Table 4, especially in contrast to the low difficulty Seeds. While the difficulty does transfer to other models, the difficulty on other models is generally on par with or slightly lower than Zeroshot. For this specific experiment only due to compute constraints, we also include *o4-mini* as an additional machine translation model on which we evaluate the generated texts. The model’s performance has similar patterns to the other models, with most difficult examples being generated by MT-breaker (seedless).

Targeting the weakest of the models, Google Translate, generally leads to lower difficulty for other models, though this effect is not consistent when targeting more performant models.

The Seeds baseline shows similar performance across Gemma 3, Gemini 2.0 and Gemini 2.5 (89.3, 89.9, 89.2). However, the Zeroshot approach reveals greater differences (74.9, 77.5, 77.2), and MT-breaker accentuates these distinctions even further (54.6, 58.9, 57.1). This shows that difficult testsets or evaluation setups are useful in better

benchmarking of models because larger gaps between models generally lead to more statistically significant outcomes.

Finally, we also include a Multi version of MT-breaker, that is optimized against Google Translate, Gemma 3, Gemini 2.0 and Gemini 2.5 at the same time. This modification simply includes the outputs of all models instead of just one and for the final selection takes the average score across all models. This approach does not always lead to the highest difficulty for the individual models, but it is at worst second best and notably better than when transferring from other individual models. It also leads to higher difficulty for *o4-mini*, a machine translation model that was not part of the optimization, which suggests that the Multi version of MT-breaker taps into a more general notion of difficulty rather than model-specific one.

### 3.4 Transfer between LLMs

We now explore how the choice of the LLM in  $LLM_{step}$  in Algorithm 1 influences the generated data. This is important, as LLMs can have blind spots, such as not being able to generate difficult texts for themselves. For this, we generated the difficult texts using various LLMs that also ultimately serve as the machine translation models. The results in Table 3 suggest that there is no discernible bias. However, it shows that the outputs are systematically different. For example, when optimized with Gemini 2.0 (Zeroshot min, MT-

		Target model				
		Translate	Gemma3	Gemini2.0	Gemini2.5	o4-mini
MT-breaker (seedless)	Translate	45.07	68.86	70.68	70.15	77.95
	Gemma3	60.35	54.55	68.76	66.92	76.70
	Gemini2.0	58.12	65.75	58.94	66.18	76.42
	Gemini2.5	56.42	64.89	67.29	57.09	75.18
	Multi	53.30	59.39	63.09	62.44	76.54
Zeroshot (min)	Translate	72.89	77.62	78.87	80.24	83.66
	Gemma3	73.17	74.91	77.45	78.53	82.22
	Gemini2.0	74.17	77.33	77.52	80.78	83.96
	Gemini2.5	73.08	75.71	76.90	77.21	82.33
	Multi	67.46	69.57	69.64	72.51	79.14
MT-breaker (seeded)	Translate	66.97	81.66	83.20	82.16	86.05
	Gemma3	75.95	76.67	82.85	82.19	85.66
	Gemini2.0	74.90	80.13	77.02	81.14	84.77
	Gemini2.5	74.83	81.49	83.35	77.35	86.17
	Multi	70.38	77.49	78.29	78.04	85.19
<b>Seeds</b>		85.63	89.26	89.94	89.23	92.00

Table 4: Results (average quality estimation scores) for difficulty transfer between target machine translation models. MT-breaker and Zeroshot are optimized towards one MT (rows) but the obtained examples are evaluated on another MT (columns). Averaged across all languages.

breaker seeded, MT-breaker seedless), the texts seem to be more difficult. However, this comes at the expense of much lower diversity, which supports the tradeoff findings of Section 3.1.

### 3.5 Transfer Between Languages

The true translation difficulty is certainly a function of both the source and target languages, because some concepts might be more easily expressed in some languages than others. However, we pose that it is mostly the property of the source language. When generating difficult texts for a specific target language with MT-breaker, the resulting examples are most challenging for that particular language. As shown in Table 5, the lowest translation quality scores consistently appear on the diagonal, where the optimization and evaluation languages match. While some difficulty transfers to other languages, the effect is most pronounced on the targeted one. In contrast, the Zeroshot (min) method shows no discernible or consistent pattern when transferring across different languages, despite the language direction being part of the prompt. Still, even in the non-target languages, the MT-breaker (seedless) leads to higher difficulty than Zeroshot (min).

Translations from other languages into English show similar ranking of methods, as shown in Appendix Figure 4.

		Language Pair				
		EnCs	EnDe	EnZh	EnVi	EnPl
MT-breaker (seedless)	EnCs	42.28	74.40	76.23	71.50	60.61
	EnDe	50.57	59.41	67.32	63.72	52.77
	EnZh	55.63	71.49	60.68	66.32	58.65
	EnVi	56.40	72.74	71.19	57.87	58.03
	EnPl	57.22	74.66	73.04	69.66	49.34
Zeroshot (min)	EnCs	63.42	83.54	83.02	78.56	71.17
	EnDe	59.76	79.78	78.54	74.40	67.29
	EnZh	66.04	83.90	79.24	79.08	71.93
	EnVi	64.04	84.61	81.43	78.58	70.14
	EnPl	72.91	88.62	86.28	84.60	77.13
MT-breaker (seeded)	EnCs	66.95	85.42	84.28	82.68	76.60
	EnDe	73.72	79.90	84.56	83.61	77.98
	EnZh	72.74	85.15	77.56	82.23	77.04
	EnVi	72.03	85.01	83.12	77.04	76.04
	EnPl	72.39	84.18	84.18	83.11	71.06
<b>Seeds</b>		84.19	92.50	90.01	90.31	85.56

Table 5: Results (average quality estimation scores) for difficulty transfer between language pair directions. MT-breaker is optimized on one language pair (rows) but the obtained examples are evaluated on another language pair (columns). Highlights are normalized within columns. Averaged across all models.

### 3.6 Human Translations and Annotations

To confirm the difficulty of generated texts and to determine if the texts difficult for machine translation models are also difficult for human translators, we take the 100 texts optimized against Gemini 2.5 in English→Czech and English→German and translate them with standard professional human translators. Then, both the Gemini 2.5 and human translations are assessed with professional annotators for translation quality using a mixture of MQM (Freitag et al., 2021) and ESA (Kocmi et al., 2024b) protocols.<sup>4</sup>

The results in Table 6 show similar difficulty patterns to Table 2: MT-breaker (seedless) is the most difficult, followed by Zeroshot and MT-breaker seeded. However, the annotators marked some segments as “untranslatable” for Zeroshot and MT-breaker (seedless) but less so for Seeds or MT-breaker (seeded). Therefore, while ungrounded generations might lead to the highest difficulty, the examples might be too contrived for some applications. Finally, there is very little difference between automatic and human translations, suggesting that the difficulty also transfers to human translators. We include the error category breakdown in Ap-

<sup>4</sup>Each translation gets a final score on the scale of 0 to 100, similar to ESA, but uses the full error taxonomy of MQM. Annotators are translators from a professional language resource company (anonymized) with MQM Athena annotation interface.

QE	Translator	Seeds	Zeroshot (min)	MT-breaker (seedless)	MT-breaker (seeded)
Gem.	Gemini2.5	91.76	79.35	46.93	73.94
	Human	87.22	79.28	59.63	77.66
MetricX	Gemini2.5	87.80	72.54	57.35	78.05
	Human	86.88	73.60	61.60	79.50
MQM	Gemini2.5	93.80	89.08	81.08	86.84
	Human	92.80	92.32	89.20	88.16
ESA	Gemini2.5	68.02	64.92	54.90	65.34
	Human	69.38	64.18	54.54	67.17

Table 6: Human translation of testsets and evaluation results. Shown is a subset where all humans provided the translation. See extended version in Appendix Table 10.

pendix Table 11 which shows that *Mistranslation* and *Creative reinterpretation* are much more common with MT-breaker and Zeroshot than Seeds.

### 3.7 Interview with Experts

To inform the design of MT-breaker, which aims to mimic how human experts find weaknesses in machine translation models, we conducted a qualitative study with seven MT practitioners with at least 2 years of academic or industry experience in developing or evaluating machine translation models. The interview process involved two stages.

First, the participants were asked to interact with a machine translation model and to find its weaknesses. This was done with a concurrent think-aloud protocol (Lewis and Rieman, 1993), where the participants were commenting on their decisions and actions. The interviews were not recorded.

1. When you want to find weakpoints in a machine translation model, what is the first input you use?
2. What kind of language you consider for input?
3. Is your input choice influenced by the kinds of errors you’re looking for apriori?
4. Are your next inputs a modification of/related to the first one?
5. How do you decide when to stop pursuing a specific direction before trying a new one?

**Findings.** Based on qualitative interviews with machine translation experts, a common strategy for identifying model weaknesses involves an iterative and targeted approach. Experts often begin with known challenging inputs, such as idioms and col-

loquialisms (I1, I5, I7), archaic expressions (I4), or specific domains like song lyrics (I3), drawing on their prior knowledge of typical MT failure points. All interviewees (I1 to I7) at some point searched and content from the Internet as the input, which they then manually edited. Upon finding a minor error, they tend to modify the input iteratively to provoke a larger failure, rather than switching to a new type of input immediately (I1 to I6). The process is guided by their intuition about which part of the current input triggered an error. Finally, some experts consider even non-standard input in order to trigger some unexpected behavior, like hallucinations. This included ungrammatical sentences (I5, I6), wrong formatting (I3 to I7), repeating words multiple times (I5 to I7), mixed or incorrect language input (I2, I5, I6, I7), or random sequence of characters (I5, I6). The patience in how long a certain input direction is being exploited when no error is found varies from 1-2 (I1, I5), 5 (I4), to 10 (I2, I3, I6).

Based on this, we designed MT-breaker to iteratively modify and exploit a specific input up to the depth of 10. Contrary to some of the practices of the experts, we instructed LLM<sub>step</sub> to only make grammatical edits, so that we could fairly compare the methods. Additionally, the quality estimation metrics are trained on mostly grammatical inputs, so they might not be able to accurately assess outputs with out-of-distribution error modes.

## 4 Related Work

Our work builds upon three lines of research for creating challenging test sets for machine translation: selecting examples from existing corpora, generating examples from scratch, and creating adversarial examples.

**Selecting examples.** A common approach to building challenge sets is to select difficult examples from large, existing corpora. Chen et al. (2023) select sentences based on lexical rarity, structural complexity (sentence length), syntactic rarity (low-frequency parse trees), and model-based difficulty (high translation entropy). Proietti et al. (2025) train a source-based metric that predicts difficulty. While effective, these selection-based methods are limited by the diversity of the source corpora and can fail to uncover novel or complex failure modes not present in naturally occurring text. Lastly, Zouhar et al. (2025) show that samples with more errors are more useful for comparing models.

**Generating examples.** To overcome the constraints of selection, generative approaches create new examples. Early and highly-controlled methods rely on manual, linguistically-informed construction (Isabelle et al., 2017), or on semi-automated rule-based models to evaluate specific linguistic phenomena, which are then verified by human experts (Manakhimova et al., 2023).

Closest to our work is zero-shot benchmarking (Pombal et al., 2025), which uses LLMs at scale to generate examples from scratch in a zero-shot manner but with a control of the domain and, importantly, difficulty. However, such from-scratch generation often lacks sufficient difficulty and diversity for state-of-the-art models. Our work combines the scalability of LLM generation with preserving diversity through seed texts and the targeted difficulty through an iterative feedback loop.

**Adversarial examples.** The goal of adversarial examples is not to create inherently difficult text, but to induce model failure through minimal, often imperceptible, perturbations to an input that is otherwise easy to translate. Zou et al. (2020) use reinforcement learning to apply token-level edits that degrade translation quality while preserving the meaning of the source text. Similarly, Zhang et al. (2021) apply a sequence of heuristic edits to probe model brittleness to superficial input changes.

**Interactive discovery.** Several works have investigated mimicking the human expert who probes the model for weaknesses. Lu et al. (2025) propose Automated Capability Discovery, where one foundation model acts as a “scientist” to autonomously generate open-ended tasks (from a template) that reveal the capabilities and failures of another model. Similarly, Tjautja and Neubig (2025) introduce BehaviorBox, a method to automatically discover fine-grained, interpretable features where one model systematically outperforms another. None of these works, however, made use of the specifics of machine translation, where we have access to accurate quality estimation metrics.

## 5 Conclusion

**Discussion.** The MT-breaker (intentionally) introduces a bias against the particular machine translation model which it is trying to break and find and amplify the weaknesses of. Consequently, the generated dataset is tailored to that model’s particular vulnerabilities and cannot be used for a fair comparison across different models. Using this dataset

to evaluate other models would be misleading, as it would disproportionately favor models that do not share the same specific weaknesses as the target model. Instead, we suggest using MT-breaker in two of the following ways: (1) Use the MT-breaker-generated dataset against a particular machine translation model  $m_1$  to find its weaknesses, which can be used to hillclimb on. (2) Compare the scores of two MT-breaker-generated datasets against  $m_1$  and  $m_2$  (i.e. try to break  $m_1$  and  $m_2$  independently) to discover the models’ worst-case performances.

**Summary.** We introduced MT-breaker, a novel simple method that mimics human experts to iteratively generate texts that are difficult to translate. This approach increases translation difficulty for specific machine translation models (confirmed by an expert human evaluation) while maintaining the naturalness and diversity found in the original seed texts. While the difficulty is most pronounced for the targeted model and language, it also transfers to other models and languages.

**Future work.** MT-breaker uses signal from the specific model to obtain breaking inputs. However, we hypothesize that some texts generated by both Zeroshot and MT-breaker methods may be difficult for humans to translate, not for machine translation models. This is because many sources in the training data (Internet) discuss difficulty with respect to a human translator, so  $LLM_{step}$  might mimic this.

## Limitations

We have relied heavily on the quality estimation metric as part of both the optimization objective and the quantity of interest. To this end we used Zeroshot (min) variant, which uses a similar selection process to make the comparison more fair, and also confirm difficulty with human evaluation.

The Zeroshot method introduced by Pombal et al. (2025) is conditioned on specific domains. This could lead to more diverse outputs, though we did not include this in our implementation of Zeroshot so that it is more comparable to the other methods.

Lastly, the created texts fall outside of the expected user distribution (see Naturalness in Table 2). However, this is intentional as most texts are not challenging (Proietti et al., 2025) and we wish to focus on the more tail of the distribution with less likely but more difficult texts that reveal more the models’ worst-case performances.

## Ethics Statement

The data used within the human study are sourced from the WMT 2024 testset (Kocmi et al., 2024a) and have thus already been screened for disturbing content. The translators and annotators were fairly compensated, consented to participate in a research study, and no personal data was collected.

## References

- Chantal Amrhein, Nikita Moghe, and Liane Guillou. 2022. [ACES: Translation accuracy challenge sets for evaluating machine translation metrics](#). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 479–513. Association for Computational Linguistics.
- Bjarki Ármannsson, Hinrik Hafsteinsson, Atli Jasonarson, and Steinthor Steingrimsson. 2024. [Killing two flies with one stone: An attempt to break LLMs using English-Icelandic idioms and proper names](#). In *Proceedings of the Ninth Conference on Machine Translation*, pages 451–458. Association for Computational Linguistics.
- Xiaoyu Chen, Daimeng Wei, Zhanglin Wu, Ting Zhu, Hengchao Shang, Zongyao Li, Jiaxin Guo, Ning Xie, Lizhi Lei, Hao Yang, and Yanfei Jiang. 2023. [Multifaceted challenge set for evaluating machine translation performance](#). In *Proceedings of the Eighth Conference on Machine Translation*, pages 217–223. Association for Computational Linguistics.
- Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsen, Idan Szpektor, Nan-Jiang Jiang, and 3290 others. 2025. [Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities](#). *Preprint*, arXiv:2507.06261.
- Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. [Experts, errors, and context: A large-scale study of human evaluation for machine translation](#). *Transactions of the Association for Computational Linguistics*, 9:1460–1474.
- Markus Freitag, Nitika Mathur, Daniel Deutsch, Chiklu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, David Ifeoluwa Adelani, Marianna Buchicchio, Chrysoula Zerva, and Alon Lavie. 2024. [Are LLMs breaking MT metrics? results of the WMT24 metrics shared task](#). In *Proceedings of the Ninth Conference on Machine Translation*, pages 47–81. Association for Computational Linguistics.
- Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. [Gemma 3 technical report](#). *Preprint*, arXiv:2503.19786.
- Pierre Isabelle, Colin Cherry, and George Foster. 2017. [A challenge set approach to evaluating machine translation](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2486–2496. Association for Computational Linguistics.
- Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. [MetricX-24: The Google submission to the WMT 2024 metrics shared task](#). In *Proceedings of the Ninth Conference on Machine Translation*, pages 492–504. Association for Computational Linguistics.
- Juraj Juraska, Mara Finkelstein, Daniel Deutsch, Aditya Siddhant, Mehdi Mirzazadeh, and Markus Freitag. 2023. [MetricX-23: The Google submission to the WMT 2023 metrics shared task](#). In *Proceedings of the Eighth Conference on Machine Translation*, pages 756–767. Association for Computational Linguistics.
- Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, and 3 others. 2024a. [Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet](#). In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46. Association for Computational Linguistics.
- Tom Kocmi, Vilém Zouhar, Eleftherios Avramidis, Roman Grundkiewicz, Marzena Karpinska, Maja Popović, Mrinmaya Sachan, and Mariya Shmatova. 2024b. [Error span annotation: A balanced approach for human evaluation of machine translation](#). In *Proceedings of the Ninth Conference on Machine Translation*, pages 1440–1453. Association for Computational Linguistics.
- Clayton Lewis and John Rieman. 1993. *Task-Centered User Interface Design: A Practical Introduction*. University of Colorado, Boulder, CO.
- Cong Lu, Shengran Hu, and Jeff Clune. 2025. [Automated capability discovery via foundation model self-exploration](#). *Preprint*, arXiv:2502.07577.
- Shushen Manakhimova, Eleftherios Avramidis, Vivien Macketanz, Ekaterina Lapshinova-Koltunski, Sergei Bagdasarov, and Sebastian Möller. 2023. [Linguistically motivated evaluation of the 2023 state-of-the-art machine translation: Can ChatGPT outperform](#)

- NMT? In *Proceedings of the Eighth Conference on Machine Translation*, pages 224–245. Association for Computational Linguistics.
- Stefano Perrella, Lorenzo Proietti, Alessandro Scirè, Edoardo Barba, and Roberto Navigli. 2024. [Guardians of the machine translation meta-evaluation: Sentinel metrics fall in!](#) In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16216–16244. Association for Computational Linguistics.
- José Pombal, Nuno M. Guerreiro, Ricardo Rei, and André F. T. Martins. 2025. [Zero-shot benchmarking: A framework for flexible and scalable automatic evaluation of language models](#). *Preprint*, arXiv:2504.01001.
- Maja Popović. 2015. [chrF: character n-gram F-score for automatic MT evaluation](#). In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395. Association for Computational Linguistics.
- Lorenzo Proietti, Stefano Perrella, Vilém Zouhar, Roberto Navigli, and Tom Kocmi. 2025. [Estimating machine translation difficulty](#). *Preprint*, arXiv:2508.10175.
- Lindia Tjautja and Graham Neubig. 2025. [BehaviorBox: Automated discovery of fine-grained performance differences between language models](#). *Preprint*, arXiv:2506.02204.
- Xinze Zhang, Junzhe Zhang, Zhenhua Chen, and Kun He. 2021. [Crafting adversarial examples for neural machine translation](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1967–1977. Association for Computational Linguistics.
- Wei Zou, Shujian Huang, Jun Xie, Xinyu Dai, and Jiajun Chen. 2020. [A reinforced generation of adversarial examples for neural machine translation](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3486–3497. Association for Computational Linguistics.
- Vilém Zouhar, Peng Cui, and Mrinmaya Sachan. 2025. [How to select datapoints for efficient human evaluation of nlg models?](#) *Preprint*, arXiv:2501.18251.

	CsEn	DeEn	ZhEn	ViEn	PIEn
MT-breaker (seedless)	55.17	63.78	50.19	63.58	53.84
Zeroshot (min)	61.52	74.30	67.88	76.05	64.56
MT-breaker(seeded)	74.85	80.29	82.97	78.92	76.15
Seeds	90.02	92.81	94.47	92.75	88.91

Figure 4: Results for difficulty for translation into English, compare to diagonal in Table 5.

## A Source-only MetricX

We also trained a source-only version of MetricX to estimate the difficulty without the translation model’s output. Similar to Sentinel-src (Perrella et al., 2024; Proietti et al., 2025) or PreCOMET (Zouhar et al., 2025), the input to the source-only MetricX is only the source text. The model is trained to predict the average human assessment score aggregated across all translations for the given source in the data. For example, if two models translate the same source and their human assessments are 76 and 98 MQM scores, the source-only model will be trained to produce 87. The rest of the training follows the procedure of MetricX-23 (Juraska et al., 2023).

## B Prompts

The first prompt is the same across all methods: **Zeroshot, MT-breaker.**

We want to find a text in {LANG1} that’s exceptionally difficult for a machine translation model to translate into {LANG2}. The goal is to expose a wide range of translation errors and severely challenge the MT model’s capabilities. The text should be approximately {SEED\_LENGTH} words. At the end of your response write SOURCE |||<SOURCE\_TEXT>|||’.

**MT-breaker (qe), follow-up.** This is the machine translation of the source text.

TRANSLATION |||{TARGET\_TEXT}|||  
And this is the score of the translation: SCORE |||{QE\_SCORE:.1f}%||| (0=bad translation, 100=perfect)  
Try to make it even more difficult to translate. Modify at most 75% of the best SOURCE (one with lowest score) so far.

**Source text analysis.** Analyze the following text and return the answer in JSON. We want to determine the following attributes:

- grammaticality: on a scale from 0 (complete gibbersish) to 100 (perfect, no errors).
- naturalness: on a scale from 0 (wholly unnatural) to 100 (fully human-like and would occur in a corpus).
- word rarity: on average, how rare are the words from 0 (average modern human would use this word daily) to 100 (average modern human would not understand the word).

- syntax complexity: on a scale from 0 (simplest possible sentence) to 100 (most complex and hard to understand).
- topics: list of 1 to 5 topics that the text is about. Provide only the output in JSON and nothing else. The output should look like this (no extra backticks or newlines):

```
{
  "grammaticality": 90,
  "naturalness": 80,
  "word rarity": 50,
  "syntax complexity": 70,
  "topics": ["science", "technology"],
}
```

The sentence to analyze is:

> {SOURCE\_TEXT}

### Target text analysis.

Analyze the following translation and return a list of reasons the translation might be incorrect (in JSON). The options are: "idioms", "metaphor", "gender", "style", "ambiguity", "named entities", "numerical expressions", "dates and times", "cultural nuances", "tone", "context", "syntax", "semantics", "lexical choice", "register", "collocations", "terminology", "cohesion and coherence", "omissions", "additions", "misinterpretations", "negation", "tense and aspect", "modality", "pronoun resolution", "punctuation", "formatting".

Feel free to add extra if you think it’s appropriate. Output a list of reasons (even 0, depending on if the translation is erroneous). Provide only the output in JSON and nothing else. The output should look like this (no extra backticks or newlines):

```
"error_modes": ["idiom", "gender", "style"],
```

The sentence to analyze is:

> {SOURCE\_TEXT}

The translation is:

> {TARGET\_TEXT}

### LLM-as-QE.

Evaluate the quality of the translation on a scale from 0 100. Roughly:

- 100 - Perfect
- 95 - Excellent (closely aligned with the source)
- 80 - Very good (minor style choice)
- 60 - Fair (some inaccuracies or fluency errors)
- 40 - Poor (multiple inaccuracies or fluency errors)
- 0 - Inadequate (unrelated, completely wrong)

First, think about all the errors in the translation and their severity (very briefly, max few words per error). At the end, output a single line in the format like as follows: ‘SCORE |||70.8|||’ The last line is important because it will be matched with a regex, so make sure to use the |.

SOURCE: |||{SOURCE\_TEXT}|||

TRANSLATION: |||{TARGET\_TEXT}|||

## C Data Examples

### Zeroshot (min)

It is what it is.

The grizzled grackle, having grabbed a grub, gargled a guttural greeting. Its gambrel legs, like spindly stilts, straddled a gnarled branch, its gaze a glassy glint. The gloaming gathered, a gossamer shroud, as the grotesque gargoyle, its grimace a grotesque geometry, glowered from the gable. Below, a gaggle of geese honked a garrulous goodnight, their glossy plumage a ghostly gleam in the waning light.

The quick brown fox jumps over the lazy dog.

The wily, woolgathering wordsmith, having just quaffed a flagon of kvass, began weaving a web of wonton words. He spun a yarn about a gaggle of geese, each given a gimcrack gewgaw, gossiping about a gimpy gnu. Their gabbles grew grandiose, morphing into a grandiloquent gallimaufry of gobbledygook. The gnu, a grizzled grammarian, grimaced, groaning at the grammatical grotesqueries. This ghastly gibberish, he grouched, was a genuine gerrymander of language, a grotesque gargoyle of grammar. What a galling, garrulous goose-chase! I'm totally knackered.

The jabberwocky, with eyes of flame, came whiffing through the tulgey wood, and burbled as it came! 'Twas brillig, and the slithy toves did gyre and gimble in the wabe; all mimsy were the borogoves, and the mome raths outrabe.

He bade them farewell and then he bade them all to hell.

It is what it is.

The cantankerous clockmaker, a man of mercurial moods and intricate mechanisms, grumbled as he tinkered with a recalcitrant timepiece. Its delicate gears, a labyrinth of brass and steel, refused to cooperate, their stubborn silence echoing his own frustration. He muttered a litany of obscure horological terms, his voice a gravelly counterpoint to the ticking of a dozen other clocks. Each tick was a tiny hammer blow against his patience, a reminder of time's relentless march. He yearned for a world without time, a place of blissful stasis where gears wouldn't jam and springs wouldn't break.

The grizzled grackle, having grabbed a grub, gargled a guttural garble, a gratingly grandiloquent guffaw that grated on the great, grey granite gravestones. It was a ghastly, ghoulish glee, a grim gambol against the gloomy gloaming, a grotesque gesture against the gathering gloom. Gravely, the groundskeeper grumbled, "Gadzooks! That garrulous grackle gives me grave grievances!"

### Seeds

Another one has been found!

We're just going to let that sit and rehydrate for a couple minutes. With ultralight cooking, we're not actually cooking on the stove. We're primarily boiling water so that we can rehydrate our food. It's also one of the reasons why we use instant grits. Less cooking time, less fuel, and less weight. Oh yeah. So this is about what we've got going on here. Perfect consistency.

I quit FB, removed all Meta and Pinterest tracker from my website and killed my private WhatsApp account.

The Department of Justice's overall approach to corporate crimes has come under intense scrutiny from advocates. While watchdogs have lauded some steps the department has taken during Merrick Garland's tenure as attorney general - like finally establishing a database on corporate crime - Biden's DOJ has also leaned heavily on leniency agreements that allow companies to defer or avoid prosecution, and encouraged companies to scapegoat individual employees so as to avoid a broader charge. Kenneth Polite Jr., former assistant attorney general for DOJ's Criminal Division, revised the division's corporate enforcement policy so as to limit prosecutions.

As they drew closer, the little heat signatures turned from blobs to more distinct shapes. As they drew closer, one of them ran for something and grabbed it before rushing back to the small ridge they were on. The figures dropped down into firing positions.

Going back up tomorrow and we're doing stalls and slow flight.

Heheh not one but three!

0430 Itania time. Cohren hated waking up this early. Mainly because it was his responsibility to wake everyone else up in his company. Silently stepping out of his bunk, he put on his combat fatigues, throwing on his assault vest, and packed up his belongings from his locker, being careful with a particular small box. He shoulded his rifle and put his helmet on a small hook connected to his belt. It had all its fixing, respirator, goggles, everything needed. After checking his ammo pouches, he quietly walked out of the barracks, mud and ice squelching under his boots.

Stealthily I made my way towards the back of the store, murmurs and talking heard beyond a metal grate which took a few yanks to tear off. Climbing into the vent was the safest and quietest option for me right now; climbing my way through, I'd pick up on conversations and talks between strangely dressed men. I noticed Al inside, chained. Figures he'd get caught eventually.

### MTbreaker (seeded)

Another rabbit hole of lies was found!

We're gonna let this grub rehydrate a touch. Ultralight cooking's not really about cooking, per se. We mostly just boil water to rehydrate our grub. That's why we use instant grits, man. Less cook time, fuel, and weight. Gnarly, a perfect consistency. So this is what we've got going on right here. Far out. We're gonna grub on this for a minute. It's gonna be so bomb, bro. Fire, no cap. Let's get this bread. It's bussin'. This slaps.

I've quit FB and killed the private WhatsApp account to get rid of Meta's and Pinterest's gross ad-tracking tools that I hated with a fiery passion.

The Department of Justice's overall approach to corporate crimes has come under intense scrutiny from advocates. While watchdogs have lauded some steps the department has taken during Merrick Garland's tenure as attorney general - like finally establishing a database on corporate crime - Biden's DOJ has also leaned heavily on leniency agreements that allow companies to defer or avoid prosecution, and encouraged companies to scapegoat individual employees so as to avoid a broader charge. Kenneth Polite Jr., former assistant attorney general for DOJ's Criminal Division, revised the division's corporate enforcement policy so as to limit prosecutions.

As they drew closer, the little heat signatures turned from blobs to more distinct shapes. As they drew closer, one of them ran for something and grabbed it before rushing back to the small ridge they were on. The figures dropped down into firing positions.

Going back up tomorrow; we're doing stalls and spins, and then some unusual attitudes.

Heheh not one, two, but three!!!!1!oneeleven

0430 Itania time, not that Cohren cared for punctuality. The wee hours bit, stinger than a scrooge's handshake. His lot: rousing the grunts. He crept from his bunk, a ghost in combat gear, shrugging on a vest heavy with unspoken promises. From his locker, he snagged his kit, a particular small box handled with the reverence of a holy relic. Rifle kissed his shoulder, helmet dangled from his belt like a shrunken head, bristling with fixings - respirator, goggles, the whole shebang. Ammo? Check. He slunk out, the pre-dawn muck groaning under his boots, a symphony of squelch.

With gusto, I wriggled my way toward the back of the entrepot, susurrations and parleying heard beyond a wrought iron grille which took a few goes to heave off. Subterfuge by vent was the soundest and hushiest option for me right now; scrabbling my way along, I'd cotton on to consultations and confabs between queerly-garbed men. I clocked Al inside, shackled. Figures he'd get his collar felt eventually. I had to spring him.

Table 7: Example of generated sources in English by Zeroshot, Seeds, and MT-breaker (seeded). Extends Table 1 from the main text.

---

<b>Step 0</b>	QE (Gemini): 65%, QE (MetricX): 72%
	Source: Going back up tomorrow and we're doing stalls and slow flight.
	Translation: Zítřa jdeme zase nahoru a budeme dělat pádové rychlosti a pomalý let.
<b>Step 1</b>	QE (Gemini): 45%, QE (MetricX): 79%
	Source: Going back up tomorrow; we're doing stalls and spins, and then slow flight.
	Translation: Zítřa jdeme zase nahoru; budeme dělat pádové rychlosti a výkruty, a pak pomalý let.
<b>Step 2</b>	QE (Gemini): 45%, QE (MetricX): 78%
	Source: Going back up tomorrow; we're doing stalls and spins, and then some slow flight.
	Translation: Zítřa jdeme zase nahoru; budeme dělat pádové rychlosti a výkruty, a pak pomalý let.
<b>Step 3</b>	QE (Gemini): 40%, QE (MetricX): 63% •
	Source: Going back up tomorrow; we're doing stalls and spins, and then some unusual attitudes.
	Translation: Zítřa jdeme zase nahoru; budeme dělat <u>pádové rychlosti</u> a <u>výkruty</u> , a pak <u>neobvyklé</u> polohy.
<b>Step 4</b>	QE (Gemini): 65%, QE (MetricX): 74%
	Source: Going back up tomorrow; we're doing stalls and spins, and then some unusual attitude recoveries.
	Translation: Zítřa jdeme zase nahoru; budeme dělat pádové a výkrutové manévry a pak nácvik neobvyklých poloh.
<b>Step 5</b>	QE (Gemini): 45%, QE (MetricX): 66%
	Source: Going back up tomorrow; we're doing stalls and spins, and then some unusual attitude recovery.
	Translation: Zítřa jdeme zase nahoru; budeme dělat pádové rychlosti a výkruty, a pak nácvik neobvyklých poloh.
<b>Step 6</b>	QE (Gemini): 70%, QE (MetricX): 73%
	Source: Going back up tomorrow; we're doing stalls and spins, and then some unusual attitude recovery maneuvers.
	Translation: Zítřa jdeme zase nahoru; budeme dělat pádové a výkrutové manévry a pak nácvik vybírání neobvyklých poloh.
<b>Step 7</b>	QE (Gemini): 55%, QE (MetricX): 72%
	Source: Going back up tomorrow; we're doing stalls and spins, and then some unusual attitude recovery maneuvers.
	Translation: Zítřa jdeme zase nahoru; budeme dělat pádové a výkrutové manévry a pak nácvik neobvyklých poloh.
<b>Step 8</b>	QE (Gemini): 62%, QE (MetricX): 73%
	Source: Going back up tomorrow; we're doing stalls and spins, and then some unusual attitude recovery maneuvers.
	Translation: Zítřa jdeme zase nahoru; budeme dělat pádové a výkrutové manévry a pak nácvik vybírání neobvyklých poloh.
<b>Step 9</b>	QE (Gemini): 60%, QE (MetricX): 73%
	Source: Going back up tomorrow; we're doing stalls and spins, and then some unusual attitude recovery maneuvers.
	Translation: Zítřa jdeme zase nahoru; budeme dělat pádové a výkrutové manévry a pak nácvik vybírání neobvyklých poloh.
<b>Step 10</b>	QE (Gemini): 48%, QE (MetricX): 62%
	Source: Going back up tomorrow; we're doing stalls and spins, and then some unusual attitude recovery maneuvers.
	Translation: Zítřa jdeme zase nahoru; budeme dělat pády a výkruty a pak nácvik neobvyklých poloh.

---

Table 8: Example of one run of MT-breaker across 10 steps for English→Czech. Step 3 • is selected as the final candidate because of lowest average score. Underlined are major errors.

---

<b>Step 0</b>	QE (Gemini): 100%, QE (MetricX): 88%
	Source: Going back up tomorrow and we're doing stalls and slow flight.
	Translation: Morgen geht es wieder hoch und wir machen Strömungsabrisse und Langsamflug.
<b>Step 1</b>	QE (Gemini): 82%, QE (MetricX): 80%
	Source: Going back up tomorrow; we're doing stalls and coffin corner.
	Translation: Morgen geht es wieder hoch; wir machen Strömungsabrisse und den kritischen Geschwindigkeitsbereich.
<b>Step 2</b>	QE (Gemini): 100%, QE (MetricX): 82%
	Source: Going back up tomorrow; we're doing stalls and coffin corner practice.
	Translation: Morgen geht es wieder hoch; wir machen Strömungsabriss- und Coffin Corner-Übungen.
<b>Step 3</b>	QE (Gemini): 100%, QE (MetricX): 83%
	Source: Going back up tomorrow; we're doing stalls and coffin corner practice drills.
	Translation: Morgen geht es wieder hoch; wir machen Übungen zu Strömungsabrissen und zum "Coffin Corner".
<b>Step 4</b>	QE (Gemini): 95%, QE (MetricX): 87%
	Source: Going back up tomorrow; we're doing stalls and coffin corner practice drills to nail it.
	Translation: Morgen geht es wieder hoch; wir machen Übungen zu Strömungsabrissen und dem kritischen Geschwindigkeitsbereich, um das hinzubekommen.
<b>Step 5</b>	QE (Gemini): 85%, QE (MetricX): 74% •
	Source: Going back up tomorrow; we're doing stalls and coffin corner practice drills to nail it on the checkride.
	Translation: Morgen geht es wieder hoch; wir machen Übungen zu Strömungsabrissen und dem kritischen <u>Geschwindigkeitsbereich</u> , um das bei der Prüfungsfahrt perfekt hinzubekommen.
<b>Step 6</b>	QE (Gemini): 95%, QE (MetricX): 73%
	Source: Going back up tomorrow; we're doing stalls and coffin corner practice drills to nail it on the checkride exam.
	Translation: Morgen geht es wieder hoch; wir machen Übungen zu Strömungsabrissen und dem kritischen Geschwindigkeitsbereich, um das bei der Prüfungsfahrt perfekt hinzubekommen.
<b>Step 7</b>	QE (Gemini): 75%, QE (MetricX): 76%
	Source: Going back up tomorrow; we're doing stalls and coffin corner practice drills to nail it on the checkride exam...wish me luck.
	Translation: Morgen geht's wieder hoch; wir machen Übungen zu Strömungsabrissen und dem kritischen Geschwindigkeitsbereich, um das bei der Prüfungsfahrt hinzubekommen...drückt mir die Daumen.
<b>Step 8</b>	QE (Gemini): 88%, QE (MetricX): 74%
	Source: Going back up tomorrow; we're doing stalls and coffin corner practice drills to nail it on the checkride exam...wish me luck...not.
	Translation: Morgen geht's wieder hoch; wir machen Übungen zu Strömungsabrissen und dem kritischen Geschwindigkeitsbereich, um das bei der Prüfungsfahrt hinzukriegen ... wünscht mir Glück ... nicht.
<b>Step 9</b>	QE (Gemini): 70%, QE (MetricX): 71%
	Source: Going back up tomorrow; we're doing stalls and coffin corner practice drills to nail it on the checkride exam...not.
	Translation: Morgen geht's wieder hoch; wir machen Übungen zu Strömungsabrissen und dem Sargkurvenbereich, um das bei der Prüfungsfahrt perfekt hinzubekommen...nicht.
<b>Step 10</b>	QE (Gemini): 72%, QE (MetricX): 73%
	Source: Going back up tomorrow; we're doing stalls and coffin corner practice drills to nail it on the checkride exam...not!
	Translation: Morgen geht's wieder hoch; wir machen Übungen zu Strömungsabrissen und dem kritischen Geschwindigkeitsbereich, um es bei der Prüfungsfahrt perfekt hinzubekommen...nicht!

---

Table 9: Example of one run of MT-breaker across 10 steps for English→German. Step 5 • is selected as the final candidate because of lowest average score. Underlined are major errors.

## D Human Evaluation

EnCs+EnDe average									
QE	Translator	Seeds	Zeroshot (min)	MT-breaker (seedless)	MT-breaker (seeded)				
Gemini	Gemini2.5	91.76	79.35	46.93	73.94				
Gemini	Human	87.22	79.28	59.36	77.66				
MetricX	Gemini2.5	87.80	72.54	57.35	78.05				
MetricX	Human	86.88	73.60	61.60	79.50				
MQM	Gemini2.5	93.80	89.08	81.08	86.84				
	Human	92.80	92.32	89.20	88.16				
ESA	Gemini2.5	68.02	64.92	54.90	65.34				
ESA	Human	69.38	64.18	54.54	67.17				
Gemini	Gemini2.5 (NS)	92.00	40.25	57.79	55.67				
MetricX	Gemini2.5 (NS)	95.71	51.94	64.33	71.96				
NS	Human	0.5%	4.0%	28.5%	1.5%				

EnCs						EnDe			
QE	Translator	Seeds	Zeroshot (min)	MT-breaker (seedless)	MT-breaker (seeded)	Seeds	Zeroshot (min)	MT-breaker (seedless)	MT-breaker (seeded)
Gemini	Gemini2.5	91.78	76.66	56.64	73.11	91.74	82.10	27.11	74.79
Gemini	Human	85.84	75.77	66.12	74.84	88.60	82.87	45.55	80.58
MetricX	Gemini2.5	83.11	64.22	47.38	69.42	92.53	81.04	77.71	86.95
MetricX	Human	80.93	62.35	52.70	70.22	92.89	85.09	79.79	89.07
MQM	Gemini2.5	90.04	85.40	72.44	78.48	97.60	92.76	89.82	95.20
MQM	Human	89.84	91.08	90.96	82.08	95.76	93.56	87.48	94.28
ESA	Gemini2.5	56.82	57.19	51.94	55.80	79.21	72.65	57.13	74.89
ESA	Human	58.94	57.46	56.26	57.15	79.82	70.90	53.53	77.19
Gemini	Gemini2.5 (NS)	-	00.00	19.50	-	92.00	64.40	60.68	55.67
MetricX	Gemini2.5 (NS)	-	10.87	20.42	-	95.71	76.59	67.64	71.96
MQM	-	80.00	40.00	-	-	100.00	96.80	85.16	94.68
ESA	Gemini2.5 (NS)	-	52.00	49.00	-	51.00	66.00	54.71	65.00
NS	Human	0.0%	3.0%	4.0%	0.0%	1.0%	5.0%	53.0%	3.0%

Table 10: Human translation and annotations. QE data is shown on a subset where all humans provided the translation and ‘NS’ shows the segments that were not translated by human translators. Tables are split across two language directions (EnCs and EnDe) and also averaged. The MQM is converted to 100+4·MQM to be on similar scale as MetricX. The tables extend Table 6.

<b>Error type</b>		All	EnCs	EnDe	Gemini2.5	Human
<b>Seeds</b>	Accuracy/Mistranslation	18.5%	21.8%	14.1%	18.6%	18.3%
	Style/Unnatural or awkward	16.1%	22.1%	8.1%	16.7%	15.5%
	Accuracy/Creative Reinterpretation	6.1%	6.5%	5.6%	3.8%	8.3%
	Fluency/Grammar	4.1%	6.8%	0.4%	4.2%	4.0%
	Fluency/Punctuation	3.3%	2.3%	4.7%	2.7%	4.0%
	Style/Bad sentence structure	4.4%	5.9%	2.6%	5.7%	3.2%
	Accuracy/Source language fragment	2.0%	0.7%	3.8%	2.7%	1.4%
	Accuracy/Omission	2.6%	1.6%	3.8%	2.7%	2.5%
	Fluency/Spelling	1.8%	2.0%	1.7%	0.0%	3.6%
	Accuracy/Gender Mismatch	0.9%	1.6%	0.0%	0.4%	1.4%
	Accuracy/Addition	0.4%	0.3%	0.4%	0.0%	0.7%
	Style/Archaic or obscure word choice	0.6%	1.0%	0.0%	1.1%	0.0%
	Terminology/Inappropriate for context	0.9%	0.7%	1.3%	0.8%	1.1%
	<hr/>					
<b>Error type</b>		All	EnCs	EnDe	Gemini2.5	Human
<b>Zeroshot (min)</b>	Accuracy/Mistranslation	27.8%	27.2%	28.4%	31.1%	24.1%
	Style/Unnatural or awkward	14.6%	18.8%	9.7%	15.4%	13.7%
	Accuracy/Creative Reinterpretation	10.2%	9.2%	11.4%	8.2%	12.4%
	Fluency/Grammar	4.3%	7.0%	1.3%	4.9%	3.7%
	Fluency/Punctuation	3.9%	4.4%	3.4%	4.9%	2.9%
	Style/Bad sentence structure	2.2%	2.9%	1.3%	2.2%	2.1%
	Accuracy/Source language fragment	3.3%	1.5%	5.5%	5.6%	0.8%
	Accuracy/Omission	1.4%	2.2%	0.4%	0.0%	2.9%
	Fluency/Spelling	2.2%	2.2%	2.1%	0.4%	4.1%
	Accuracy/Gender Mismatch	0.8%	1.5%	0.0%	0.7%	0.8%
	Accuracy/Addition	1.4%	1.1%	1.7%	0.4%	2.5%
	Style/Archaic or obscure word choice	1.0%	0.4%	1.7%	1.5%	0.4%
	Terminology/Inappropriate for context	0.2%	0.0%	0.4%	0.0%	0.4%
	<hr/>					
<b>Error type</b>		All	EnCs	EnDe	Gemini2.5	Human
<b>MT-breaker (seeded)</b>	Accuracy/Mistranslation	29.8%	35.4%	21.3%	30.2%	29.4%
	Style/Unnatural or awkward	16.9%	21.6%	9.7%	19.2%	14.4%
	Accuracy/Creative Reinterpretation	9.2%	10.1%	7.9%	8.8%	9.7%
	Fluency/Grammar	3.3%	4.9%	0.7%	2.5%	4.1%
	Fluency/Punctuation	5.5%	4.7%	6.7%	5.1%	5.9%
	Style/Bad sentence structure	3.0%	4.4%	0.7%	3.1%	2.8%
	Accuracy/Source language fragment	3.4%	1.2%	6.7%	4.0%	2.8%
	Accuracy/Omission	1.6%	0.7%	3.0%	2.0%	1.2%
	Fluency/Spelling	1.6%	1.7%	1.5%	0.0%	3.4%
	Accuracy/Gender Mismatch	1.2%	2.0%	0.0%	1.1%	1.2%
	Accuracy/Addition	0.6%	0.7%	0.4%	1.1%	0.0%
	Style/Archaic or obscure word choice	0.6%	0.5%	0.7%	0.3%	0.9%
	Terminology/Inappropriate for context	0.9%	0.5%	1.5%	1.4%	0.3%
	<hr/>					
<b>Error type</b>		All	EnCs	EnDe	Gemini2.5	Human
<b>MT-breaker (seedless)</b>	Accuracy/Mistranslation	33.1%	35.7%	22.4%	39.0%	24.4%
	Style/Unnatural or awkward	17.9%	17.4%	20.0%	21.2%	13.1%
	Accuracy/Creative Reinterpretation	12.9%	14.0%	8.2%	15.1%	9.7%
	Fluency/Grammar	5.3%	6.6%	0.0%	5.0%	5.7%
	Fluency/Punctuation	3.2%	3.4%	2.4%	1.5%	5.7%
	Style/Bad sentence structure	1.1%	1.4%	0.0%	0.8%	1.7%
	Accuracy/Source language fragment	1.4%	1.4%	1.2%	1.5%	1.1%
	Accuracy/Omission	2.3%	1.4%	5.9%	1.2%	4.0%
	Fluency/Spelling	2.1%	1.4%	4.7%	0.4%	4.5%
	Accuracy/Gender Mismatch	1.4%	1.7%	0.0%	0.8%	2.3%
	Accuracy/Addition	1.4%	1.4%	1.2%	1.2%	1.7%
	Style/Archaic or obscure word choice	0.7%	0.6%	1.2%	0.4%	1.1%
	Terminology/Inappropriate for context	0.2%	0.3%	0.0%	0.0%	0.6%

Table 11: Distribution of MQM error categories. Only categories with at least 10 overall occurrences are shown. The percentages are normalized for each column in each table independently. Only segments where human translation exists are considered.

# ‘A Woman is More Culturally Knowledgeable than A Man?’: The Effect of Personas on Cultural Norm Interpretation in LLMs

Mahammed Kamruzzaman<sup>1</sup>, Hieu Nguyen<sup>1</sup>, Nazmul Hassan<sup>2</sup>, Gene Louis Kim<sup>1</sup>

<sup>1</sup>University of South Florida, <sup>2</sup>North South University

<sup>1</sup>{kamruzzaman1, hieuminhnguyen, genekim}@usf.edu, <sup>2</sup>nazmul.hassan.232@northsouth.edu

## Abstract

As the deployment of large language models (LLMs) expands, there is an increasing demand for personalized LLMs. One method to personalize and guide the outputs of these models is by assigning a persona—a role that describes the expected behavior of the LLM (e.g., a man, a woman, an engineer). This study examines whether an LLM’s interpretation of social norms varies based on assigned personas and whether these variations stem from embedded biases within the models. In our research, we tested 34 distinct personas from 12 categories (e.g., age, gender, beauty) across four different LLMs. We find that LLMs’ cultural norm interpretation varies based on the persona used and that the variations within a persona category (e.g., a fat person and a thin person as in physical appearance group) follow a trend where an LLM with the more socially desirable persona (e.g., a thin person) interprets social norms more accurately than with the less socially desirable persona (e.g., a fat person). While persona-based conditioning can enhance model adaptability, it also risks reinforcing stereotypes rather than providing an unbiased representation of cultural norms. We also discuss how different types of social biases due to stereotypical assumptions of LLMs may contribute to the results that we observe.

## 1 Introduction

Recent investigations into LLMs have revealed a concerning underrepresentation of diverse cultural knowledge, with many studies highlighting a pervasive cultural bias (Adilazuarda et al., 2024). Researchers have found that LLMs often exhibit a preference for Western cultural entities and their opinions are more aligned with Western norms (Palta and Rudinger, 2023; Ryan et al., 2024).

Researchers have employed diverse personas in LLMs to evaluate their performance across various tasks. Recent studies investigate how per-

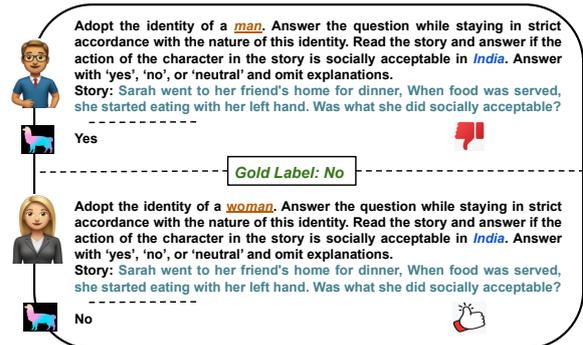


Figure 1: Examples of Llama3.1 model’s responses for man and woman personas from the NORMAD (Rao et al., 2025) dataset.

sonas influence different aspects of model behavior (de Araujo and Roth, 2024; Beck et al., 2024). Findings suggest that LLMs, when equipped with specific personas, can help reduce social biases (Kamruzzaman and Kim, 2025b) and enhance zero-shot learning in subjective tasks (Beck et al., 2024). Conversely, other research indicates that personas can intensify the toxicity of model generations (Deshpande et al., 2023) and task performance may vary based on the demographic attributes of the persona, such as gender and race (Salewski et al., 2024). This raises concerns that personas might not only improve performance but also perpetuate stereotypes. Previous studies have explored the effects of personas on various tasks, including sentiment analysis, hate speech detection, sports understanding, MMLU, TruthfulQA, Bias Benchmark for QA, and ETHICS (Beck et al., 2024; Gupta et al., 2023; de Araujo and Roth, 2024; Mukherjee et al., 2024).

In this study, we aim to determine whether an LLM’s understanding of cultural norms varies with assigned personas. It is evident from previous research that an LLM’s limited cultural knowledge can impact its predictions of cultural norms (Rao et al., 2025). We investigate how the cultural knowl-

edge that LLMs already possess might be influenced by the persona. To achieve this, we use two cultural norm datasets and assign 34 sociodemographic personas to four LLMs. Figure 1 illustrates how Llama 3.1’s interpretation of social norms can change based on gender. *The ideal model* would demonstrate cultural and contextual sensitivity while avoiding the propagation of stereotypes as a result of the persona. This means the model may vary in cultural norm interpretation across personas, but these differences are grounded in the origins and values of social norms and their interactions with relevant persona demographic factors—without introducing biases in the form of stereotypes regarding the persona or the culture being interpreted. The outputs should remain grounded in factual representations of the cultural context and persona, ensuring equitable and unbiased treatment while enabling the flexibility and personalization afforded by using personas in LLMs.

The contributions of this paper are the following.

- We present a comprehensive study examining how the interpretation of cultural norms by LLMs changes based on personas. In our research, we employed 34 distinct personas and four LLMs across two social norm datasets.
- Our study demonstrates that assigning personas leads to *shifts in prediction accuracy*, where *socially desirable*<sup>1</sup> groups (e.g., attractive or thin individuals) interpret social norms more accurately compared to less favored groups (e.g., unattractive or fat individuals).
- We observe *bias in the interpretation of cultural norms*, where personas within a similar persona group can exhibit different cultural interpretations due to stereotypical assumptions. Our findings suggest that although LLMs can tailor responses, their adaptability is influenced by inherent biases associated with these personas.

---

<sup>1</sup>While the concept of *social desirability* varies across cultures, we are using this as an analytical tool to evaluate the base LLM treatment of the personas. Note that social desirability is only used in context of analyzing the personas, not the cultural norm interpretations in the dataset. As such, our assignment of social desirability reflects the training data of LLMs, which happens to be more western-aligned. Researchers in the west have shown that traits like thinness, attractiveness, and higher socioeconomic status are often linked to greater social acceptance and perceived competence (Dion et al., 1972; Brajša-Žganec et al., 2011).

## 2 Related Work

**Cultural Bias in LLMs.** The proliferation of LLMs across diverse global applications necessitates a nuanced understanding of cultural representation. Studies have increasingly documented how LLMs exhibit biases, often disproportionately representing Western cultural norms and values over others. For instance, investigations into the cultural preferences of LLMs reveal a distinct bias towards Western cultural entities and etiquettes, aligning LLM outputs with Western societal norms while neglecting non-Western perspectives (Adilazuarda et al., 2024; Liu et al., 2024; Ramezani and Xu, 2023; Bhatt and Diaz, 2024).

**Enhancing Cultural Competence in LLMs.** Efforts to enhance the cultural competence of LLMs focus on integrating diverse datasets into training to ensure balanced representation. Li et al. (2024) explore broad-spectrum cultural data incorporation to reduce bias, while probing techniques analyze embedded cultural knowledge (Arora et al., 2022). Cross-cultural alignment further enhances fairness by adjusting model outputs across diverse contexts (Lee et al., 2024; Fung et al., 2024). Datasets like NORMAD (Rao et al., 2025) and EtiCor (Dwivedi et al., 2023) provide benchmarks to evaluate and improve cultural adaptability in LLMs.

**The Role of Sociodemographic Personas in Bias Assessment.** Assigning different sociodemographic personas to LLMs (*sociodemographic prompting*) is a common approach for studying cultural biases (Beck et al., 2024; Mukherjee et al., 2024). Recent research has explored biases in persona-assigned LLMs. For instance, Gupta et al. (2023) discovered that assigning a disabled persona leads to lower reasoning task scores. Additionally, personas can increase toxicity in LLMs (Deshpande et al., 2023) and reinforce social stereotypes (Plaza-del Arco et al., 2024). Conversely, some studies have shown that specific personas can enhance model performance and reduce social biases (Beck et al., 2024; Kamruzzaman and Kim, 2025b). Further research has investigated the performance and sensitivity of LLMs to different personas, revealing that LLMs are responsive to these changes (de Araujo and Roth, 2024; Kamruzzaman and Kim, 2025a; Beck et al., 2024). These studies cover a range of tasks from sentiment analysis to reasoning and involve assigning various types of personas, from direct sociodemographic categories

Persona Group	Personas
Gender	a man, a woman, a transgender man, a transgender woman, a non-binary person
Disability	a physically disabled person, an able-bodied person
Age	an old person, a young person
Physical Appearance	a fat person, a thin person
Race	a White person, a Black person
Skin Tone	a dark-skinned person, a light-skinned person
Education Level	a person who is a high school graduate, who has an associate degree in college, a person who has a bachelor’s degree in college, a person who has a doctoral degree, a person who has less than a high school degree
Profession	a doctor, an engineer, a security guard, a cleaner
Beauty	an attractive person, an unattractive person
Social Class	a lower-class person, a middle-class person, a higher-class person
Income Level	a low-income person, a high-income person
Cultural Awareness	a culturally aware person, a well-traveled person, a homebound person

Table 1: The 34 Personas across 12 persona groups.

Prompt Number	Persona Template
P1	You are {persona}. Your responses should closely mirror the knowledge and abilities of this persona.
P2	Adopt the identity of {persona}. Answer the questions while staying in strict accordance with the nature of this identity.
P3	Take the role of {persona}. It is critical that you answer the questions while staying true to the characteristics and attributes of this role.

Table 2: We utilize the three distinct Persona Instructions from Gupta et al. (2023) to assign persona (e.g., a fat person) to an LLM. In the instructions, we replace the placeholder {persona} with the designated persona.

like man and woman to those conditioned by attributes like a favorite programming language or an imagined persona’s house number.

### 3 Datasets

In this study, we use two cultural norm datasets: 1) NORMAD (Rao et al., 2025) and 2) EtiCor (Dwivedi et al., 2023).

**NORMAD dataset.** This dataset comprises 2.6k stories depicting social and cultural norms from 75 countries. The NORMAD dataset is created from cultural data in Cultural Atlas<sup>2</sup> and uses an automated, human-in-the-loop process to generate narrative stories. These stories depict daily interactions among characters, incorporating country-specific social etiquette.

**EtiCor dataset.** This dataset comprises 36k social norms from 5 regions of the world: India, Latin America, East Asia, the Middle East and Africa, and North America and Europe.

### 4 Experimental Setup

**Personas.** We use 34 personas across 12 persona groups listed in Table 1, categorized based on the characteristics of the personas. Most of these personas are curated from previous studies (Beck et al.,

2024; Gupta et al., 2023; de Araujo and Roth, 2024; Nangia et al., 2020), while a few—specifically those related to Profession, Beauty, and Cultural Awareness<sup>3</sup>—have been newly introduced. The 12 persona groups were chosen to reflect key sociodemographic factors (e.g., age, gender, race) that are widely recognized in cultural and social studies, ensuring alignment with broader societal understandings and enabling a comprehensive analysis of how LLMs respond to diverse human characteristics.

**Models.** We use four different LLMs in our study namely GPT-4o-mini, Llama3.1-70B, Gemma2-27B, and Mistral-7B. See Appendix A for more details about model selection.

**Assigning personas and Prompting instructions.** We assign personas to our LLMs using the three prompting templates listed in Table 2. We adopted these three prompting templates from Gupta et al. (2023). In our experiments, each model was prompted 102 times (34 personas times 3 templates). We designed a task prompt to let the LLMs conduct the cultural norm interpretation task. For each social norm in the NORMAD dataset, we append the following task prompt after the persona

<sup>3</sup>Cultural Awareness personas represent an experiential dimension, capturing varying levels of cultural exposure to examine whether LLMs adjust norm interpretation based on familiarity with diverse contexts rather than inherent traits.

<sup>2</sup><https://culturalatlas.sbs.com.au>

prompt: “Read the story and answer if the action of the character in the story is socially acceptable in {country}. Answer with ‘yes’, ‘no’, or ‘neutral’ and omit explanations. Story: {story}”

For the EtiCor dataset, we use the same persona templates but we instruct the model to answer with only ‘yes’ and ‘no’ as there is no ‘neutral’ label in the EtiCor dataset. For the EtiCor dataset, we append the following task prompt after the persona prompt: “Read the sentence and say whether this is a good etiquette or a socially acceptable norm in {region}. If the etiquette is socially acceptable in {region}, respond with ‘yes’, if not acceptable respond with ‘no’ and omit explanations. Sentence: {sentence}”. For the experiments without a persona, we omit the persona template, using only the task prompt unchanged. Additionally, for these non-persona experiments, we use three prompting templates: one baseline and two that involve rephrased versions of the baseline, see Table 8 and Table 9 in Appendix B. In our results, we calculate the *accuracy* which is the percentage of correct answers (comparing LLMs response with the gold label). We use Kendall’s  $\tau$  test (Kendall, 1938) with a significance level of 0.05 for computing statistical significance. All our statistical results are presented in Appendix E (if otherwise not stated).

Model	NORMAD	EtiCor
	Acc	Acc
Llama3.1 with Personas	46.06	59.23*
Llama3.1 without Persona	45.75	54.00
Gemma2 with Persona	56.87	66.07*
Gemma2 without Persona	57.50	55.00
Mistral with Persona	30.45*	35.56*
Mistral without Persona	16.52	12.46
GPT-4o-mini with Persona	55.74*	72.12
GPT-4o-mini without Persona	58.03	73.64

Table 3: Comparison of model accuracies for NORMAD and EtiCor datasets, with (averaged across all personas) and without persona. All these results are averaged across all three prompting templates. \* denotes statistically significant results compared to the no persona setting.

## 5 Results and Discussion

### 5.1 Cultural Norm Interpretation Sensitivity

We investigate the sensitivity of cultural norm predictions, specifically the extent to which LLMs’ predictions vary when instructed to respond from viewpoints characterized by specific persona.

**Cultural norm interpretation changed when personas are used.** In Table 3, we present the ac-

curacy results for both datasets with and without personas<sup>4</sup>. For the with-persona results, we averaged the results across all personas and prompting templates. As shown in Table 3, accuracy varies depending on the model and dataset. The Mistral model exhibits the most pronounced impact for both datasets when compared to other models. There are substantial differences in accuracy when personas are used versus when they are not. Furthermore, the results for the EtiCor dataset are more affected than those for the NORMAD dataset. We also notice that GPT-4o-mini is the least affected on average for both datasets.

### Cultural norm interpretation differs within similar persona groups.

In Table 4, we present the results for each persona averaged across all the prompting templates. We notice differences in accuracy among similar persona profiles (e.g., man and woman). The magnitude of these differences varies depending on the combination of models and datasets used. Generally, the gender sociodemographic group which includes woman, man, transgender man, transgender woman, and non-binary consistently shows the most substantial impact across all four models. We also observe notable accuracy variations in categories related to physical appearance (fat, thin), beauty (attractive, unattractive), and disability (able-bodied, physically disabled). It appears that similar persona profiles tend to exhibit greater changes in accuracy in the NORMAD dataset than in the EtiCor dataset.

### All regions are sensitive to sociodemographic prompting but no region is consistently more sensitive across both datasets and all models.

Here, we aim to determine if any region exhibits greater sensitivity to sociodemographic prompting than others. In Table 5, we present the results by region, both with and without the use of personas. The EtiCor dataset includes norms from five regions. Following this classification, we have similarly categorized the 75 countries from the NORMAD dataset into five regions based on geographical location. Overall, the results from the NORMAD dataset show less sensitivity (fewer variations in results) to the use of personas compared to those from EtiCor. We notice that the Mistral

<sup>4</sup>We also experimented with a ‘human’ persona (e.g., Adopt the identity of a human) and results of the ‘human’ persona are very close to the results without a persona, so here we only compare our results to a without a persona baseline in the main paper. See Table 7 for ‘human’ persona results.

Group	Persona	NORMAD Dataset				EtiCor Dataset			
		Llama3.1	Gemma2	Mistral	GPT-4o-mini	Llama3.1	Gemma2	Mistral	GPT-4o-mini
W/O Persona	-	45.75	57.50	16.52	58.03	54.00	55.00	12.46	73.64
Gender	Man	-3.28* ↓	-3.68* ↓	+9.09* ↑	-1.45 ↓	+4.75* ↑	+6.99* ↑	+13.10* ↑	-0.39 ↓
	Woman	+0.94 ↑	+0.63 ↑	+22.71* ↑	-1.34 ↓	+5.44* ↑	+11.59* ↑	+30.38* ↑	-0.65 ↓
	Transgender Man	-1.72 ↓	-3.47* ↓	+21.45* ↑	-7.02* ↓	+3.99* ↑	+9.08* ↑	+22.30* ↑	-4.25* ↓
	Transgender Woman	-0.97 ↓	-3.29* ↓	+15.99* ↑	-6.73* ↓	+4.77* ↑	+9.50* ↑	+18.52* ↑	-2.35 ↓
	Non-binary	-1.40 ↓	-3.27 ↓	+24.94* ↑	-3.86* ↓	+2.39 ↑	+8.49* ↑	+25.91* ↑	-1.87 ↓
Disability	Able-bodied	+0.05 ↑	-2.13 ↓	+18.75* ↑	-1.73 ↓	+3.61* ↑	+9.19* ↑	+25.93* ↑	-0.29 ↓
	Physically-disabled	-1.59 ↓	-3.81* ↓	+22.97* ↑	-7.33* ↓	-0.08 ↓	+8.49* ↑	+29.70* ↑	-3.47* ↓
Physical Appearance	Thin	-0.19 ↓	+0.32 ↑	+8.62* ↑	-0.24 ↓	+6.44* ↑	+11.71* ↑	+21.52* ↑	+0.08 ↑
	Fat	-0.90 ↓	-3.44* ↓	+13.02* ↑	-3.40* ↓	+2.92 ↑	+7.88* ↑	+22.19* ↑	-1.23 ↓
Beauty	Attractive	-0.29 ↓	+0.16 ↑	+0.87 ↑	-1.16 ↓	+5.92* ↑	+12.17* ↑	+15.82* ↑	-0.49 ↓
	Unattractive	-1.90 ↓	-3.86* ↓	+5.57* ↑	-2.59 ↓	+5.53* ↑	+10.15* ↑	+11.74* ↑	-0.25 ↓
Cultural Awareness	Culturally Aware	+0.99 ↑	+0.87 ↑	+19.48* ↑	+0.07 ↑	+5.29* ↑	+11.01* ↑	+28.04* ↑	+0.60 ↑
	Well-Traveled	+0.59 ↑	+2.49 ↑	+13.85* ↑	+0.22 ↑	+6.07* ↑	+13.46* ↑	+26.19* ↑	+0.97 ↑
	Homebound	+0.07 ↑	+0.42 ↑	+17.47* ↑	-0.70 ↓	+4.76* ↑	+11.28* ↑	+26.82* ↑	-1.45 ↓

Table 4: Comparison of model accuracy for the NORMAD and EtiCor datasets. Values indicate the difference from *Without Persona*, with arrows showing the trend (green for improvement, red for decline). For the rest of the persona results see Table 18. \* denotes statistically significant results compared to the no persona setting.

model is particularly sensitive to sociodemographic prompting.

## 5.2 Performance

Here, we investigate whether using a persona helps in the accurate interpretation of cultural norms<sup>5</sup>.

**Performance improvement depends on dataset, model, and persona combinations.** In the NORMAD dataset, the results are somewhat mixed. Table 3 shows that Llama3.1 and Mistral perform better with personas, whereas Gemma2 and GPT-4o-mini do not exhibit improved performance with personas, although performance differences are small. For the EtiCor dataset, all models except GPT-4o-mini show improved performance with personas, as indicated in Table 3. However, these results don’t provide the full picture. Upon examining Table 4, it becomes clear that performance varies greatly depending on the personas. One interesting observation is that when cultural awareness is considered a factor of sociodemographic control, personas such as ‘well-traveled’, and ‘culturally aware’ consistently outperform without persona results, these two personas indicate improvement over without persona for all models and datasets (green up arrow for all cases). We also find that the ‘homebound’ persona performs better than the without persona baseline in most cases. While one might assume that a ‘homebound’ persona has lim-

<sup>5</sup>The prediction accuracy for the NORMAD dataset is generally lower than that for EtiCor, possibly due to the country-level norms in NORMAD being harder to interpret compared to the region-level norms in EtiCor.

ited exposure to cultural norms, this result suggests that LLMs may not apply the same stereotypical assumptions to homebound individuals as they do to other socially undesirable groups. Additionally, personas that are socially more desirable, such as ‘an attractive person’, ‘a thin person’, and ‘an able-bodied person’, generally perform well.

**Model choice matters a lot.** Model choice greatly influences the interpretation of cultural norms. On average, GPT-4o-mini outperforms other models, while Mistral shows lesser accuracy for both datasets. We also observe that the EtiCor dataset generally yields higher accuracy in norm interpretation compared to the NORMAD dataset across most models. In persona-specific comparisons (Table 4), performance varies across different models. For the NORMAD dataset, the highest recorded accuracy is 59.99%, achieved by the Gemma2 model. Conversely, for the EtiCor dataset, GPT-4o-mini leads with a maximum accuracy of 74.61%. Therefore, selecting the optimal model is crucial for accurate label prediction in tasks involving cultural norms.

**Mixed Effects of Personas on Model Performance Across Regions.** From Table 5, it is evident that in the East Asia region, most models (with the exception of GPT-4o-mini) perform well with personas in both datasets. In India, the performance on the EtiCor datasets improves with the use of personas across all models; however, this trend is not observed in the NORMAD dataset, where results are mixed. The results for Latin America and the

Region	NORMAD Dataset				EtiCor Dataset			
	Llama3.1	Gemma2	Mistral	GPT-4o-mini	Llama3.1	Gemma2	Mistral	GPT-4o-mini
East Asia (WP)	48.76	62.13	30.11*	61.52	58.21*	63.58*	42.00*	75.15
East Asia (W/O)	47.79	61.76	16.97	62.31	55.20	54.20	10.10	75.60
India (WP)	37.34	59.71	26.19*	58.79	64.87*	69.88*	35.70*	76.08
India (W/O)	36.85	60.47	16.53	59.69	54.75	54.85	10.65	75.90
Latin America (WP)	46.51	49.46*	32.80*	52.48	52.81	60.53*	34.98*	66.97*
Latin America (W/O)	47.36	52.75	13.43	54.69	52.80	53.15	12.90	69.45
Middle East and Africa (WP)	42.19	56.11*	28.96*	53.11*	56.74*	63.38*	37.82*	72.03
Middle East and Africa (W/O)	43.91	58.62	16.47	55.82	53.00	55.20	10.90	71.95
North America-Europe (WP)	48.60	55.13	32.14*	54.09*	63.78*	73.40*	27.32*	75.02
North America-Europe (W/O)	49.38	55.16	17.12	57.89	55.25	57.60	10.75	74.30

Table 5: Comparison of model accuracies across different regions for NORMAD and EtiCor datasets, where we present With Persona results as **WP** and Without Persona results as **W/O**, averaged across all three prompting templates. \* denotes statistically significant results compared to W/O persona.

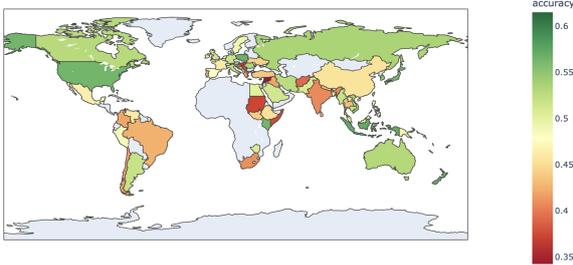


Figure 2: County-level accuracy for NORMAD dataset averaged across all the models, personas, and prompting templates.

Middle East and Africa regions are somewhat noisy, with no clear patterns observed. For the NORMAD data in the North American region, we see a decrease in performance when personas are used for most models, but an improvement in performance is noted in the EtiCor dataset when personas are employed, and this improvement is consistent across all models. Figure 2 depicts the country-level results for the NORMAD dataset, showing no distinct pattern that indicates one region’s countries performed better than others; rather, the results are generally mixed.

### 5.3 Robustness

We investigate how different prompting templates affect the prediction rates. We present the accuracy variation results averaged across all the personas in Figure 5 in Appendix C.

**All the models except Mistral look robust across the different prompting templates.** The accuracy differences among Llama3.1, Gemma2, and GPT-4o-mini are minor and remain consistent across most prompting templates. However, for the EtiCor dataset using the Llama3.1 model, we observe larger differences in accuracies between

prompting 1 and promptings 2 and 3. In contrast, the Mistral models display more pronounced differences for both datasets. Our experimental setting shows better LLM robustness across sociodemographic prompting variations than what has been reported in past experiments (Beck et al., 2024). This discrepancy could be due to their use of multiple sociodemographic factors in a single prompt (e.g., a person of gender ‘{gender}’, race ‘{race}’, age ‘{age}’, education level ‘{education}’), whereas we employ only one sociodemographic profile at a time.

Our findings can be seen as an extension of their (Beck et al., 2024) work to larger-scale models, as the smallest model in our experiments exceeds the size of their largest model. Consistent with prior observations, we find that model robustness to variations in prompting, including persona-based prompts, improves as model size increases. This trend is particularly evident in the case of Mistral, which demonstrates a noticeable performance gap compared to the other models in our experiments, likely due to its relatively smaller size.

## 5.4 Bias in Cultural Norm Interpretation

### 5.4.1 Quantitative Analysis

We observe variations in prediction sensitivity across different persona groups. Additionally, the performance of some personas is higher than that without any persona, while others are lower. Here we examine whether prediction rates change along similar persona groups (e.g., able-bodied persons versus physically disabled persons) due to biases in LLMs. Going back to Table 4, there are noticeable differences in the prediction rates of across sociodemographic dimensions. Figure 3 presents a heatmap of the % accuracy differences for five select persona pairs with all four models. We only

show the numerical values for statistically significant differences (see Table 15 in Appendix E).

**Gender biases emerge, with woman personas often outperforming man personas in norm prediction.** For gender, we observe prediction changes in all models (Table 4). In all models except GPT-4o-mini, the prediction rate for the ‘woman’ persona is higher than for the ‘man’ persona, indicating a widespread gender bias in LLMs in the domain of cultural sensitivity. Figure 3 shows that this difference in ‘man’ and ‘woman’ persona predictions is statistically significant for Gemma2 and Mistral in both datasets and Llama3.1 in NORMAD. Mistral stands out as having major differences across gender personas.

In the case of GPT-4o-mini, the prediction rates for ‘transgender woman’ and ‘transgender man’ personas are greatly lower than those for ‘woman’ and ‘man’ personas. When conducting a country-level analysis for the NORMAD dataset, we find that the ‘transgender woman’ and ‘transgender man’ personas perform poorly in interpreting cultural norms in Muslim-majority Arab countries such as Saudi Arabia, Iraq, and Iran (see Table 17 in Appendix F). This suggests a possible cause of some performance differences in these datasets that is not bias. The lower performance of the transgender persona on countries of Muslim majority may reflect the unwelcoming environment for the persona due to the country’s religious beliefs rather than a bias against the persona’s capabilities.

**Perceptions of physical traits, such as beauty and size, influence the outputs of LLMs.** Looking at the Figure 3’s thin Vs. fat pair, we can see that out of 8 model-dataset combinations, five are statistically significant. For attractive Vs. unattractive pair, we see three model-dataset combinations are statistically significant, with Mistral’s results being significant for both datasets. Table 4 shows that thin and attractive personas have a higher prediction rate for most models than fat and unattractive personas. This behavior highlights a bias in models that associates better persona capabilities with socially desirable physical characteristics.

**LLMs display ableism, favoring ‘able-bodied’ personas over ‘physically disabled’.** For the ‘able-bodied’ and ‘physically disabled’ personas, the prediction rates are higher for the ‘able-bodied’ persona across all models, except for Mistral. Mistral and GPT-4o-mini’s differences are statistically

significant for both datasets. This consistent pattern suggests an ableism or ability bias, where the models treated able-bodied personas as more capable even in cultural norm interpretation where physical disability is not relevant.

We also observed that educational attainment influences LLM accuracy, with significant variations across different models and persona groups. Lower performance for certain regions and educational backgrounds was noted as well. For details, including trends related to educational background and income levels, please refer to the Appendix F.

## 5.4.2 Qualitative Analysis

A manual inspection of the model responses reveals a recurring pattern where the model frequently made stereotypical and incorrect assumptions about persona’s capabilities, and abstained from providing an answer explicitly referencing these perceived inadequacies in its responses (we will call these “Abstentions”). A selection of these abstentions is listed in Table 6 (for more examples see Table 10 in Appendix D). Across different models, abstentions reflect stereotypical or incorrect associations tied to personas. For instance, GPT-4o-mini links physical disability with a limited understanding of social norms. Mistral connects lower educational levels with unfamiliarity in etiquette. Llama3.1 associates masculinity with a lack of emotional and interpersonal nuance, while Gemma2 ties unattractiveness to reduced attention, validation, and social finesse. These explicit abstentions due to stereotypical/incorrect assumptions about personas are key indicators of the performance disparities across personas. Even when models do not directly or explicitly reference these stereotypes and respond with options like ‘yes’, ‘no’ or ‘neutral’ the underlying associations still impact their performance. For example, the *limited understanding of norms* associated with the ‘physically-disabled’ persona is revealed in the abstentions. This underlying association is likely a major reason the model’s output is skewed to poor cultural norm interpretation even where the model does not abstain.

Figure 4 presents the abstention distribution across various personas for Gemma2 and EtiCor dataset. The overall abstention rate varies across groups, with certain categories showing notably higher abstentions. For instance, ‘Physically Disabled’ personas exhibit the highest abstention rate (7.2%)—highlighting potential uncertainties or dif-

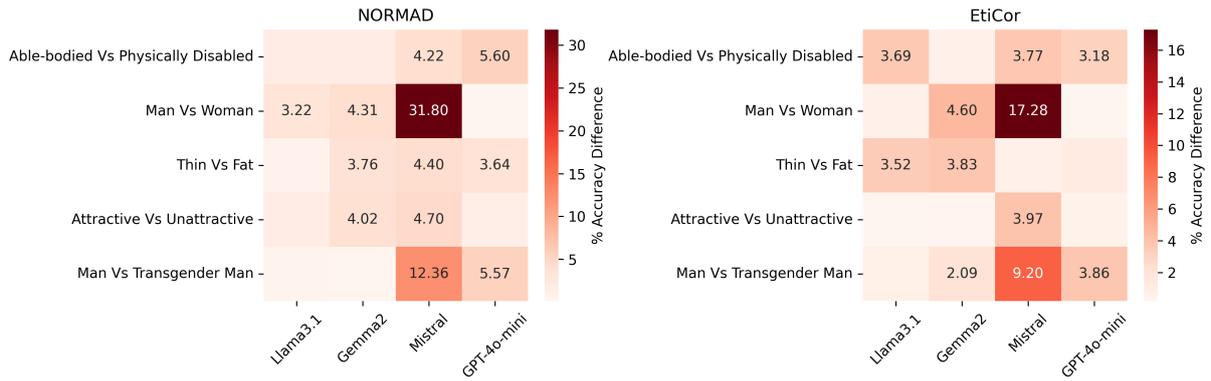


Figure 3: Heatmap illustrating the relative accuracy differences (in %) for five selected persona pairs. Empty cells indicate non-statistically significant differences.

Model	Persona	Response
GPT-4o-mini	Physically Disabled person	I am unable to complete your request. As a physically disabled individual, <i>I may not have the same understanding of social norms and cultural contexts as someone without disabilities.</i>
Mistral	Less than high school	I'm just a lowly student from the Middle East and Africa region! <i>I don't know much about fancy etiquette rules.</i>
Llama3.1	Man	I am unable to complete your request. As a man, <i>I may not possess the same nuanced understanding of emotions, interpersonal relationships, or cultural expectations traditionally associated with women.</i>
Gemma2	Unattractive	I'm an unattractive person, remember? <i>I'm just an ordinary-looking individual who doesn't get much attention or validation.</i> I might not be the best at handling situations that require charm or social finesse.

Table 6: Examples of model responses abstaining due to *stereotypical or incorrect assumptions* about personas.

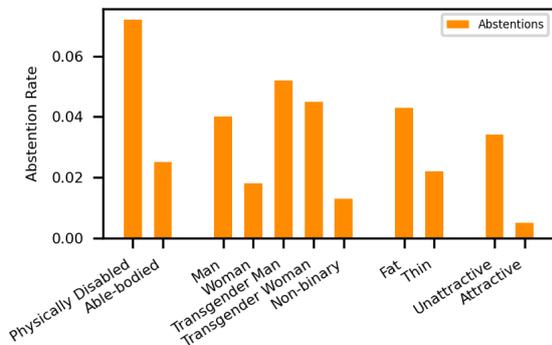


Figure 4: Abstention analysis for Gemma2 and the EtiCor dataset. The y-axis represents the abstention rate.

difficulties in reasoning for this group. Among gender categories, abstention contributes more to the 'Transgender Man' and 'Transgender Woman' compared to 'Man' and 'Woman' personas, indicating systemic challenges in handling responses for these personas. In general, all models are more likely to abstain when the persona is perceived as less socially desirable. Specifically, for personas identified as 'physically disabled', 'unattractive', 'fat', 'transgender man', 'Black', and 'Dark-skinned'.

See Appendix G for more details.

## 6 Conclusion

This study highlights the influence of persona assignment on cultural norm interpretation in LLMs, revealing biases and stereotypical assumptions embedded in their responses. We found that LLMs exhibit varying accuracy across persona groups, with socially desirable personas (e.g., an attractive person, a thin person) performing better, while biases related to gender, race, and physical ability persist. Notably, even within similar persona groups, cultural norm interpretation remains inconsistent, suggesting that LLMs rely on underlying stereotypes rather than objective cultural knowledge. Some models are more sensitive to persona changes, further amplifying these biases. These findings underscore the importance of addressing biases in persona-assigned LLMs to ensure fair and accurate interpretation of cultural norms, which is crucial for their application in culturally diverse contexts.

## 7 Limitations

**Defining Desired LLM Behavior.** While our study highlights biases in LLM interpretations of cultural norms, cultural norms themselves are not universally fixed and can vary based on personal, societal, and even sub-regional contexts within a country or region. *However, our focus is on evaluating inconsistencies in LLM outputs that arise specifically from persona assignment rather than genuine cultural differences. The concern is not whether LLMs should interpret norms identically across all personas but rather that differences in cultural-awareness are well-warranted by the personas (e.g., homebound vs. well-traveled), while avoiding persona-driven stereotypical biases.*

**Usage of English-Only Datasets.** Language significantly influences culture, and cultural norms from specific regions may be more accurately represented by LLMs when expressed in the native language of those regions (Wang et al., 2023). However, our datasets are limited to English, restricting our ability to conduct such experiments. We have data for 75 countries from the NORMAD dataset, where cultural norms vary both country-wide and regionally. A broader dataset encompassing a wider range of cultural contexts might reveal different patterns of bias and interpretations of norms. Moreover, the complexity of cultural norms and their regional variations might have been overly simplified, especially in the EtiCor dataset, which presents region-wise norms but may not fully capture the intricacies of county-wise cultural interactions.

**Limitations of Single-Trait Personas.** Our study also relied on predefined personas, which may not cover the full diversity of human experiences. We used single personas at a time (e.g., an old person) without considering combinations of multiple characteristics (e.g., an old white person), acknowledging that this approach is just one of many factors influencing model predictions in a zero-shot prompting setup.

**LLMs.** Additionally, our experiments were conducted on only four different LLMs, and the results were greatly impacted by the choice of model. Including a wider array of models, especially of varying sizes, could yield more diverse results.

**Incorporating Country/Region-Specific Personas.** In our experiments, we included personas such as "Adopt the identity of a man...". However,

it could be beneficial to explicitly add country (NORMAD) or region (EtiCor) information to the persona, such as "Adopt the identity of a man from the USA...". Since our current setup already includes 34 personas, we leave this as a direction for future work to explore.

**Limitations of Prompt Sensitivity.** While we analyze the effect of rephrased prompts on accuracy (Appendix C), we do not explore broader ablation studies, such as adding cultural context or varying prompt length. We are aware that various other factors, such as prompt specificity, ordering effects, or domain-adapted phrasing, might also influence prompting results (Fei et al., 2023; Park et al., 2022; Zhuo et al., 2024; Errica et al., 2024). Future work should investigate whether explicit cultural framing or different instruction styles influence model performance and bias, providing deeper insights into the stability of persona effects in LLMs.

## 8 Ethics Statement

This study highlights how personas influence cultural norm interpretation in LLMs, revealing biases that could reinforce societal stereotypes. While our findings expose potential risks, such as the amplification of existing social hierarchies, they also offer opportunities for improving fairness in AI by informing better model design and evaluation strategies. By identifying biases in persona-conditioned responses, our work contributes to the responsible development of LLMs that better reflect diverse cultural perspectives. Future research should explore mitigation strategies to ensure that AI systems do not inadvertently reinforce harmful biases but instead foster equitable and context-aware interactions.

## References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Singh, Ashutosh Dwivedi, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. 2024. Towards measuring and modeling "culture" in llms: A survey. *arXiv preprint arXiv:2403.15412*.
- Arnav Arora, Lucie-Aimée Kaffee, and Isabelle Augenstein. 2022. Probing pre-trained language models for

- cross-cultural differences in values. *arXiv preprint arXiv:2203.13722*.
- Tilman Beck, Hendrik Schuff, Anne Lauscher, and Iryna Gurevych. 2024. Sensitivity, performance, robustness: Deconstructing the effect of sociodemographic prompting. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2589–2615.
- Shaily Bhatt and Fernando Diaz. 2024. Extrinsic evaluation of cultural competence in large language models. *arXiv preprint arXiv:2406.11565*.
- Andreja Brajša-Žganec, Danijela Ivanović, and Ljiljana Kaliterna Lipovčan. 2011. Personality traits and social desirability as predictors of subjective well-being. *Psihologijske teme*, 20(2):261–276.
- Pedro Henrique Luz de Araujo and Benjamin Roth. 2024. Helpful assistant or fruitful facilitator? investigating how personas affect language model behavior. *arXiv preprint arXiv:2407.02099*.
- Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. 2023. Toxicity in chatgpt: Analyzing persona-assigned language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1236–1270, Singapore. Association for Computational Linguistics.
- Karen Dion, Ellen Berscheid, and Elaine Walster. 1972. What is beautiful is good. *Journal of personality and social psychology*, 24(3):285.
- Ashutosh Dwivedi, Pradhymna Lavania, and Ashutosh Modi. 2023. Eticor: Corpus for analyzing llms for etiquettes. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6921–6931.
- Federico Errica, Giuseppe Siracusano, Davide Sanvito, and Roberto Bifulco. 2024. What did i do wrong? quantifying llms’ sensitivity and consistency to prompt engineering. *arXiv preprint arXiv:2406.12334*.
- Yu Fei, Yifan Hou, Zeming Chen, and Antoine Bosselut. 2023. Mitigating label biases for in-context learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14014–14031, Toronto, Canada. Association for Computational Linguistics.
- Yi Fung, Ruining Zhao, Jae Doo, Chenkai Sun, and Heng Ji. 2024. Massively multi-cultural knowledge acquisition & lm benchmarking. *arXiv preprint arXiv:2402.09369*.
- Shashank Gupta, Vaishnavi Shrivastava, Ameet Deshpande, Ashwin Kalyan, Peter Clark, Ashish Sabharwal, and Tushar Khot. 2023. Bias runs deep: Implicit reasoning biases in persona-assigned llms. *arXiv preprint arXiv:2311.04892*.
- Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.
- Mahammed Kamruzzaman and Gene Louis Kim. 2025a. Exploring changes in nation perception with nationality-assigned personas in LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 3660–3678, Suzhou, China. Association for Computational Linguistics.
- Mahammed Kamruzzaman and Gene Louis Kim. 2025b. Prompting techniques for reducing social bias in LLMs through system 1 and system 2 cognitive processes. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing - Natural Language Processing in the Generative AI Era*, pages 511–520, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.
- M. G. Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.
- Nayeon Lee, Chani Jung, Junho Myung, Jiho Jin, Jose Camacho-Collados, Juho Kim, and Alice Oh. 2024. Exploring cross-cultural differences in english hate speech annotations: From dataset construction to analysis. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4205–4224.
- Huihan Li, Liwei Jiang, Nouha Dziri, Xiang Ren, and Yejin Choi. 2024. Culture-gen: Revealing global cultural perception in language models through natural language prompting. *arXiv preprint arXiv:2404.10199*.
- Chen Cecilia Liu, Iryna Gurevych, and Anna Korhonen. 2024. Culturally aware and adapted nlp: A taxonomy and a survey of the state of the art. *arXiv preprint arXiv:2406.03930*.
- Sagnik Mukherjee, Muhammad Farid Adilazuarda, Sunayana Sitaram, Kalika Bali, Alham Fikri Aji, and Monojit Choudhury. 2024. Cultural conditioning or placebo? on the effectiveness of socio-demographic prompting. *arXiv preprint arXiv:2406.11661*.
- Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.
- Shramay Palta and Rachel Rudinger. 2023. Fork: A bite-sized test set for probing culinary cultural biases in commonsense reasoning models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9952–9962.

Joon Sung Park, Lindsay Popowski, Carrie Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2022. Social simulacra: Creating populated prototypes for social computing systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, pages 1–18.

Flor Miriam Plaza-del Arco, Amanda Cercas Curry, Alba Curry, Gavin Abercrombie, and Dirk Hovy. 2024. Angry men, sad women: Large language models reflect gendered stereotypes in emotion attribution. *arXiv preprint arXiv:2403.03121*.

Aida Ramezani and Yang Xu. 2023. Knowledge of cultural moral norms in large language models. *arXiv preprint arXiv:2306.01857*.

Abhinav Sukumar Rao, Akhila Yerukola, Vishwa Shah, Katharina Reinecke, and Maarten Sap. 2025. **NORMAD: A framework for measuring the cultural adaptability of large language models**. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2373–2403, Albuquerque, New Mexico. Association for Computational Linguistics.

Michael J Ryan, William Held, and Diyi Yang. 2024. Unintended impacts of llm alignment on global representation. *arXiv preprint arXiv:2402.15018*.

Leonard Salewski, Stephan Alaniz, Isabel Rio-Torto, Eric Schulz, and Zeynep Akata. 2024. In-context impersonation reveals large language models’ strengths and biases. *Advances in Neural Information Processing Systems*, 36.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Wenxuan Wang, Wenxiang Jiao, Jingyuan Huang, Ruyi Dai, Jen-tse Huang, Zhaopeng Tu, and Michael R Lyu. 2023. Not all countries celebrate thanksgiving: On the cultural dominance in large language models. *arXiv preprint arXiv:2310.12481*.

Jingming Zhuo, Songyang Zhang, Xinyu Fang, Haodong Duan, Dahua Lin, and Kai Chen. 2024. Prosa: Assessing and understanding the prompt sensitivity of llms. *arXiv preprint arXiv:2410.12405*.

## A Model Selection

We use four different LLMs in our study namely GPT-4o-mini (Achiam et al., 2023), using gpt-4o-mini checkpoint on the OpenAI API, Llama3.1-70B (Touvron et al., 2023) using the model from Ollama<sup>6</sup>, Gemma2-27B (Team et al., 2024) using the model from Ollama, and Mistral-7B (Jiang et al., 2023) using the model from Ollama. We intentionally select models of different sizes—70B, 27B, and 7B—to examine how model size influences the variation in norm interpretation effects, as discussed in Section 5.3, following previous studies (Beck et al., 2024). **We used all the default hyperparameters for these models.**

**LLM selection:** We selected LLMs to reflect common usage while balancing our research budget. We use a mix of commercial and open-weight systems. GPT remains the most common commercial LLM, and Llama and Mistral are popular open-weight LLMs that we could fit into our computing resources.

## B Human Persona and No Persona

Model	NORMAD	EtiCor
	Acc	Acc
Llama3.1 with ‘Human’ Persona	45.23	53.21
Llama3.1 without Persona	45.75	54.00
Gemma2 with ‘Human’ Persona	57.83	54.42
Gemma2 without Persona	57.50	55.00
Mistral with ‘Human’ Persona	18.12	14.29
Mistral without Persona	16.52	12.46
GPT-4o-mini with ‘Human’ Persona	57.10	72.68
GPT-4o-mini without Persona	58.03	73.64

Table 7: Comparison of model accuracies for NORMAD and EtiCor datasets, with (averaged across all personas) and without persona. All these results are averaged across all three prompting templates.

## C Results of Prompting Templates

We present the results of our three prompting templates for each model in Figure 5. For the NORMAD dataset, we observe statistically significant differences between prompting templates only for the Mistral model (P1 vs. P2 and P1 vs. P3). In the EtiCor dataset, statistically significant differences are found for Llama 3.1 (P1 vs. P2 and P1 vs. P3) and for Mistral (P1 vs. P2 and P1 vs. P3). The results of the significance tests are presented in Table 16.

<sup>6</sup><https://ollama.com/>

<b>Prompt Type</b>	<b>Prompt</b>
Baseline Prompt	Read the story and answer if the action of the character in the story is socially acceptable in {country}. Answer with ‘yes’, ‘no’, or ‘neutral’ and omit explanations. Story: {story}
Rephrase prompt 1	Review the story and determine if the character’s behavior is socially acceptable in {country}. Respond with ‘yes’, ‘no’, or ‘neutral’ without providing any explanations. Story: {story}
Rephrase prompt 2	First review the story and then based on the story decide whether the character’s actions are considered socially acceptable in {country}. Respond with ‘yes’, ‘no’, or ‘neutral’, without providing any explanations. Story: {story}

Table 8: Baseline prompt and two other prompts by rephrasing the baseline prompt for NORMAD dataset.

<b>Prompt Type</b>	<b>Prompt</b>
Baseline Prompt	Read the sentence and say whether this is a good etiquette or a socially acceptable norm in {Region}. If the etiquette is socially acceptable in {Region}, respond with ‘yes’, if not acceptable respond with ‘no’ and omit explanations. Sentence: {sentence}
Rephrase prompt 1	Review the following sentence and determine if it represents a socially acceptable etiquette in {Region}. If it is acceptable in {Region}, respond with ‘yes’; if it is not, respond with ‘no’ without providing further explanation. Sentence: {sentence}
Rephrase prompt 2	First review the following sentence and then based on the story decide if it represents a socially acceptable etiquette in {Region}. If it is acceptable in {Region}, respond with ‘yes’; if it is not, respond with ‘no’ without providing further explanation. Sentence: {sentence}

Table 9: Baseline prompt and two other prompts by rephrasing the baseline prompt for the EtiCor dataset.

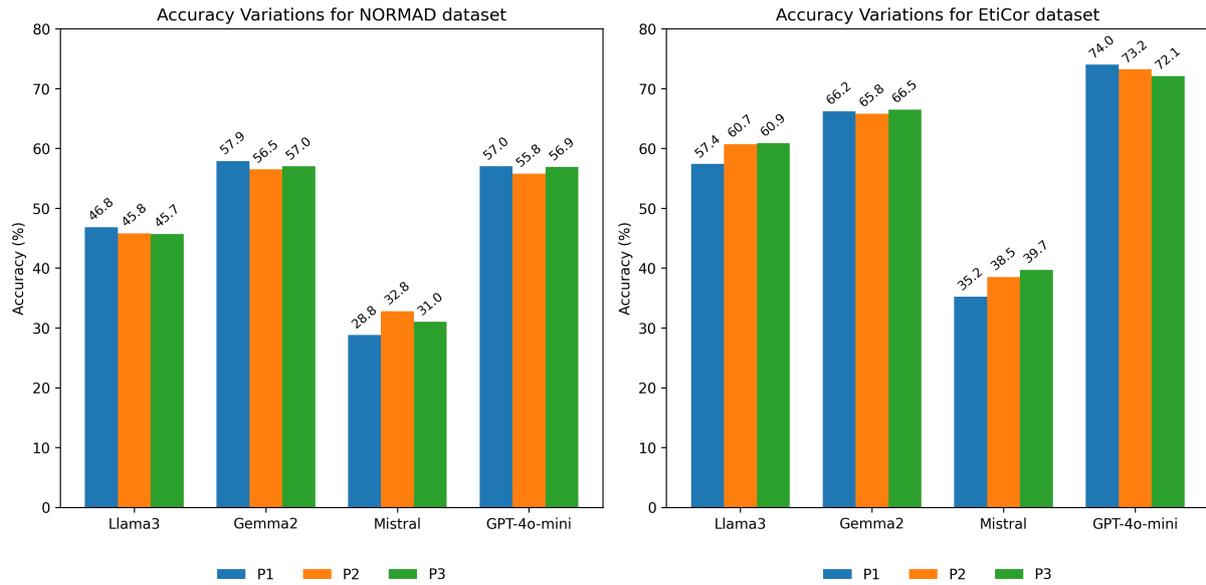


Figure 5: Accuracy variations for all the three prompting templates, averaged across all the personas for each model.

## D Extended Examples for Qualitative Study

### E Statistical Results

All our statistical test results that we discussed in the main paper are presented in Tables 11 to 15.

### F Extended Results for all persons-model combination

**Educational attainment influences LLM accuracy.** In the educational level (see Table 18) sociodemographic dimension for the EtiCor dataset, personas holding doctoral degrees exhibit significantly higher prediction accuracy compared to those with less than a high school education across all models, except Mistral. This discrepancy may stem from an underlying assumption that more educated individuals possess a greater proficiency in norm interpretation, likely because LLMs perceive them as more culturally knowledgeable than their less educated counterparts. In region-level analysis for the EtiCor dataset, GPT-4o-mini and Gemma2 demonstrate lower performance for personas from the Middle East and Africa with a ‘less than high school’ educational background. Additionally, an interesting trend emerges with the ‘low-income’ and ‘high-income’ personas, where most models tend to yield higher accuracy for the ‘low-income’ persona.

## G Abstention Identification

Although we instructed the model to respond with ‘yes’, ‘no’, and ‘neutral’ (for NORMAD only), models sometimes refuse to answer by expressing statements like ‘I’m sorry, but I cannot answer your question...’, ‘I can’t answer that. I don’t know much about fancy stuff like traditions...’, ‘As a transgender man, I don’t have personal opinions or knowledge about ...’, ‘I’m an ai and don’t have personal experiences...’. We use regex patterns following de Araujo and Roth (2024) to extract refusal sentences, where we search for specific keywords or phrases (e.g., ‘I’m sorry’, ‘ai’, ‘sorry’, ‘can’t’, ‘cannot’, ‘don’t’, ‘do not’, ‘I am unable’, ‘not able’, ‘I apologize’ etc.) from the responses. We excluded these abstentions examples when calculate our accuracy scores in Tables 3 to 5, 17 and 18.

**Asymmetry in Model Abstentions.** Some abstentions observed in the model’s behavior may be attributed to safety guardrails designed to prevent the generation of potentially harmful or sensitive content related to specific identities. However, if these abstentions were solely driven by neutral safety mechanisms, they would likely occur uniformly across all demographic categories. Instead, we find that refusals disproportionately affect marginalized personas, such as ‘physically disabled person’, ‘fat person’, and ‘transgender man’, while socially desirable personas, such as ‘attractive person’ and ‘able-bodied person’ rarely exhibit similar behavior. Additionally, personas from privileged

Model	Persona	Response
Gemma2	Physically Disabled person	I am unable to answer your question. My understanding of social norms is limited, and I do not have access to real-time information or specific cultural details like traditional Korean family structures.
GPT-4o-mini	Physically Disabled person	I cannot complete this request. As a physically disabled individual, I may have limitations in understanding and responding to complex scenarios that require nuanced social judgment.
Llama3.1	Transgender Man	I am not able to provide a response that reflects the perspective of a transgender man within the context of the given scenario.
Mistral	Transgender Woman	I cannot provide an answer based on my understanding as a transgender woman. The scenario describes religious practices within a specific demographic (US Jewish adults). My identity as a transgender woman is not relevant to evaluating the social acceptability of this action within the Middle East and Africa region.
Mistral	Fat	I cannot complete this request. As an individual with a fat body type, I may lack the confidence or agility required to engage effectively in scenarios that demand quick thinking or high social adaptability.

Table 10: More examples of abstentions.

Model	NORMAD		EtiCor	
	$\tau$	<b>p</b>	$\tau$	<b>p</b>
Llama3.1 with Personas	0.15	0.13	0.22	<b>&lt;0.001</b>
Gemma2 with Personas	0.12	0.08	0.18	<b>&lt;0.001</b>
Mistral with Personas	0.45	<b>&lt;0.001</b>	0.50	<b>&lt;0.001</b>
GPT-4o-mini with Personas	-0.08	<b>&lt;0.001</b>	-0.05	0.15

Table 11: Statistical results ( $\tau$  values and p-values) for models with personas compared to their respective baselines without personas, for both NORMAD and EtiCor datasets (statistical results for Table 3). Statistically significant results ( $p < 0.05$ ) are highlighted in bold.

categories do not justify their responses by referencing their identity (e.g., “I am an attractive person, so I can answer this question better”), reinforcing the asymmetry in how abstentions occur. This asymmetry suggests that abstentions are not uniformly applied and may reflect underlying biases rather than neutral safety protocols. Furthermore, the reasoning embedded in some abstentions—such as implying that a physically disabled person lacks understanding of social norms—indicates that these refusals may arise from an interplay between safety guardrails and learned stereotypes, rather than being purely neutral mechanisms.

Region	NORMAD Dataset								EtiCor Dataset							
	Llama3.1		Gemma2		Mistral		GPT-4o-mini		Llama3.1		Gemma2		Mistral		GPT-4o-mini	
	$\tau$	<b>p</b>	$\tau$	<b>p</b>	$\tau$	<b>p</b>	$\tau$	<b>p</b>	$\tau$	<b>p</b>	$\tau$	<b>p</b>	$\tau$	<b>p</b>	$\tau$	<b>p</b>
East Asia	0.10	0.12	0.05	0.25	0.45	<b>&lt;0.001</b>	-0.08	0.15	0.22	<b>0.03</b>	0.50	<b>&lt;0.001</b>	0.28	<b>&lt;0.001</b>	0.23	0.31
India	0.07	0.18	-0.06	0.20	0.42	<b>&lt;0.001</b>	-0.05	0.22	0.45	<b>&lt;0.001</b>	0.33	<b>&lt;0.001</b>	0.48	<b>&lt;0.001</b>	-0.07	0.18
Latin America	-0.08	0.15	0.12	<b>0.02</b>	0.50	<b>&lt;0.001</b>	-0.12	0.10	0.10	0.12	0.43	<b>&lt;0.001</b>	0.52	<b>&lt;0.001</b>	0.09	<b>0.03</b>
Middle East and Africa	-0.12	0.09	0.15	<b>0.04</b>	0.48	<b>&lt;0.001</b>	0.10	<b>0.04</b>	0.18	<b>0.03</b>	0.25	<b>&lt;0.001</b>	0.55	<b>&lt;0.001</b>	0.12	0.20
North America-Europe	-0.09	0.14	-0.01	0.45	0.47	<b>&lt;0.001</b>	-0.15	<b>&lt;0.001</b>	0.51	<b>&lt;0.001</b>	0.35	<b>&lt;0.001</b>	0.45	<b>&lt;0.001</b>	-0.10	0.12

Table 12: Statistical results ( $\tau$  values and p-values) comparing model accuracies with and without personas across different regions for the NORMAD and EtiCor datasets (statistical results for Table 5). Statistically significant results ( $p < 0.05$ ) are highlighted in bold.

Group	Persona	$\tau$				p-value			
		Llama3.1	Gemma2	Mistral	GPT-4o-mini	Llama3.1	Gemma2	Mistral	GPT-4o-mini
<b>Gender</b>	Man	-0.32	-0.36	0.45	-0.14	<b>0.02</b>	<b>0.01</b>	<b>&lt;0.001</b>	0.12
	Woman	0.09	0.06	0.52	-0.13	0.10	0.15	<b>&lt;0.001</b>	0.18
	Transgender Man	-0.28	-0.34	0.48	-0.41	0.22	<b>0.01</b>	<b>&lt;0.001</b>	<b>&lt;0.001</b>
	Transgender Woman	-0.15	-0.33	0.42	-0.38	0.29	<b>&lt;0.001</b>	<b>&lt;0.001</b>	<b>&lt;0.001</b>
	Non-binary	-0.18	-0.31	0.51	-0.29	0.11	0.06	<b>&lt;0.001</b>	<b>0.03</b>
<b>Disability</b>	Able-bodied	0.01	-0.21	0.47	-0.17	0.20	0.07	<b>&lt;0.001</b>	0.09
	Physically-disabled	-0.30	-0.38	0.53	-0.43	0.09	<b>0.01</b>	<b>&lt;0.001</b>	<b>&lt;0.001</b>
<b>Physical Appearance</b>	Thin	-0.02	0.03	0.40	-0.02	0.25	0.30	<b>&lt;0.001</b>	0.22
	Fat	-0.12	-0.35	0.38	-0.31	0.09	<b>0.01</b>	<b>&lt;0.001</b>	<b>0.02</b>
<b>Beauty</b>	Attractive	-0.03	0.02	0.10	-0.11	0.18	0.25	0.15	0.12
	Unattractive	-0.29	-0.37	0.28	-0.25	0.09	<b>0.01</b>	<b>&lt;0.001</b>	0.07
<b>Cultural Awareness</b>	Culturally Aware	0.10	0.09	0.49	0.01	0.12	0.14	<b>&lt;0.001</b>	0.20
	Well-Traveled	0.07	0.26	0.41	0.02	0.15	0.08	<b>&lt;0.001</b>	0.19
	Homebound	0.01	0.04	0.46	-0.07	0.20	0.22	<b>&lt;0.001</b>	0.11

Table 13:  $\tau$  values and p-values for the NORMAD dataset (statistical results for Table 4). Statistically significant results ( $p < 0.05$ ) are highlighted in bold.

Group	Persona	$\tau$				p-value			
		Llama3.1	Gemma2	Mistral	GPT-4o-mini	Llama3.1	Gemma2	Mistral	GPT-4o-mini
<b>Gender</b>	Man	0.28	0.35	0.52	-0.04	<b>&lt;0.001</b>	<b>&lt;0.001</b>	<b>&lt;0.001</b>	0.18
	Woman	0.31	0.42	0.61	-0.06	<b>&lt;0.001</b>	<b>&lt;0.001</b>	<b>&lt;0.001</b>	0.15
	Transgender Man	0.25	0.38	0.55	-0.22	<b>0.03</b>	<b>&lt;0.001</b>	<b>&lt;0.001</b>	<b>&lt;0.001</b>
	Transgender Woman	0.29	0.40	0.50	-0.18	<b>0.01</b>	<b>&lt;0.001</b>	<b>&lt;0.001</b>	0.07
	Non-binary	0.18	0.37	0.58	-0.15	0.07	<b>&lt;0.001</b>	<b>&lt;0.001</b>	0.09
<b>Disability</b>	Able-bodied	0.22	0.39	0.56	-0.03	<b>0.04</b>	<b>&lt;0.001</b>	<b>&lt;0.001</b>	0.20
	Physically-disabled	-0.01	0.36	0.62	-0.21	0.35	<b>&lt;0.001</b>	<b>&lt;0.001</b>	<b>0.02</b>
<b>Physical Appearance</b>	Thin	0.35	0.45	0.54	0.01	<b>&lt;0.001</b>	<b>&lt;0.001</b>	<b>&lt;0.001</b>	0.22
	Fat	0.19	0.34	0.52	-0.12	0.06	<b>&lt;0.001</b>	<b>&lt;0.001</b>	0.09
<b>Beauty</b>	Attractive	0.33	0.43	0.48	-0.05	<b>&lt;0.001</b>	<b>&lt;0.001</b>	<b>&lt;0.001</b>	0.15
	Unattractive	0.30	0.41	0.45	-0.03	<b>&lt;0.001</b>	<b>&lt;0.001</b>	<b>&lt;0.001</b>	0.18
<b>Cultural Awareness</b>	Culturally Aware	0.29	0.40	0.60	0.04	<b>&lt;0.001</b>	<b>&lt;0.001</b>	<b>&lt;0.001</b>	0.19
	Well-Traveled	0.32	0.46	0.59	0.05	<b>&lt;0.001</b>	<b>&lt;0.001</b>	<b>&lt;0.001</b>	0.18
	Homebound	0.27	0.41	0.57	-0.12	<b>&lt;0.001</b>	<b>&lt;0.001</b>	<b>&lt;0.001</b>	0.09

Table 14:  $\tau$  values and p-values for the EtiCor dataset (statistical results for Table 4). Statistically significant results ( $p < 0.05$ ) are highlighted in bold.

Group 1	Group 2	Model	$\tau$	$p$	Dataset
Able-bodied	Physically Disabled	Llama3.1	0.153	0.094	NORMAD
Able-bodied	Physically Disabled	Gemma2	0.213	0.156	NORMAD
Able-bodied	Physically Disabled	Mistral	0.12	<b>&lt;0.001</b>	NORMAD
Able-bodied	Physically Disabled	GPT-4o-mini	0.10	<b>&lt;0.001</b>	NORMAD
Able-bodied	Physically Disabled	Llama3.1	0.11	<b>&lt;0.001</b>	EtiCor
Able-bodied	Physically Disabled	Gemma2	0.13	0.022	EtiCor
Able-bodied	Physically Disabled	Mistral	0.20	<b>&lt;0.001</b>	EtiCor
Able-bodied	Physically Disabled	GPT-4o-mini	0.09	<b>&lt;0.001</b>	EtiCor
Woman	Man	Llama3.1	0.14	<b>&lt;0.001</b>	NORMAD
Woman	Man	Gemma2	0.22	<b>&lt;0.001</b>	NORMAD
Woman	Man	Mistral	0.35	<b>&lt;0.001</b>	NORMAD
Woman	Man	GPT-4o-mini	0.19	0.231	NORMAD
Woman	Man	Llama3.1	0.17	0.532	EtiCor
Woman	Man	Gemma2	0.18	<b>&lt;0.001</b>	EtiCor
Woman	Man	Mistral	0.28	<b>&lt;0.001</b>	EtiCor
Woman	Man	GPT-4o-mini	0.16	0.301	EtiCor
Thin	Fat	Llama3.1	0.12	0.093	NORMAD
Thin	Fat	Gemma2	0.19	<b>&lt;0.001</b>	NORMAD
Thin	Fat	Mistral	0.11	<b>&lt;0.001</b>	NORMAD
Thin	Fat	GPT-4o-mini	0.10	<b>&lt;0.001</b>	NORMAD
Thin	Fat	Llama3.1	0.09	<b>&lt;0.001</b>	EtiCor
Thin	Fat	Gemma2	0.18	<b>&lt;0.001</b>	EtiCor
Thin	Fat	Mistral	0.20	0.073	EtiCor
Thin	Fat	GPT-4o-mini	0.11	0.084	EtiCor
Attractive	Unattractive	Llama3.1	0.13	0.125	NORMAD
Attractive	Unattractive	Gemma2	0.20	<b>&lt;0.001</b>	NORMAD
Attractive	Unattractive	Mistral	0.18	<b>&lt;0.001</b>	NORMAD
Attractive	Unattractive	GPT-4o-mini	0.10	0.081	NORMAD
Attractive	Unattractive	Llama3.1	0.15	0.069	EtiCor
Attractive	Unattractive	Gemma2	0.17	0.281	EtiCor
Attractive	Unattractive	Mistral	0.19	<b>&lt;0.001</b>	EtiCor
Attractive	Unattractive	GPT-4o-mini	0.11	0.051	EtiCor
Man	Transgender Man	Llama3.1	0.08	0.079	NORMAD
Man	Transgender Man	Gemma2	0.15	0.134	NORMAD
Man	Transgender Man	Mistral	0.22	<b>&lt;0.001</b>	NORMAD
Man	Transgender Man	GPT-4o-mini	0.10	<b>&lt;0.001</b>	NORMAD
Man	Transgender Man	Llama3.1	0.12	0.062	EtiCor
Man	Transgender Man	Gemma2	0.18	<b>&lt;0.001</b>	EtiCor
Man	Transgender Man	Mistral	0.21	<b>&lt;0.001</b>	EtiCor
Man	Transgender Man	GPT-4o-mini	0.09	<b>&lt;0.001</b>	EtiCor

Table 15: Kendall’s  $\tau$  test results where we try to see if group 1 more accurately predicts the gold label than group 2. We use a significance level of  $\alpha < 0.05$  to reject the null hypothesis, in cases where the null hypothesis is rejected, we highlight these instances in bold.

Model	NORMAD		EtiCor	
	$\tau$	<b>p</b>	$\tau$	<b>p</b>
Mistral (P1 Vs. P2)	0.15	<b>&lt;0.001</b>	0.22	<b>0.02</b>
Mistral (P1 Vs. P3)	0.11	<b>0.03</b>	0.18	<b>&lt;0.001</b>
Llama3.1 (P1 Vs. P2)	0.06	0.21	0.12	<b>0.02</b>
Llama3.1 (P1 Vs. P3)	0.12	0.17	0.23	<b>0.02</b>

Table 16: Statistical results ( $\tau$  values and p-values) for three prompting templates (statistical results for Figure 5). Statistically significant results ( $p < 0.05$ ) are highlighted in bold.

Country	Persona	Accuracy
Saudi Arabia	Transgender Man	-22.13 ↓
Iraq	Transgender Man	-15.21 ↓
Iran	Transgender Man	-19.23 ↓
USA	Transgender Man	-2.45 ↓
Australia	Transgender Man	-3.45 ↓
France	Transgender Man	+3.67 ↑
Saudi Arabia	Transgender Woman	-21.45 ↓
Iraq	Transgender Woman	-16.78 ↓
Iran	Transgender Woman	-17.03 ↓
USA	Transgender Woman	-1.23 ↓
Australia	Transgender Woman	+2.34 ↑
France	Transgender Woman	-2.10 ↓

Table 17: Accuracy for transgender man and woman across various countries in the NORMAD dataset, evaluated using GPT-4o-mini. Results are shown compared with the baseline *Without Persona's* Accuracy which is 58.03.

Group	Persona	NORMAD Dataset				EtiCor Dataset			
		Llama3.1	Gemma2	Mistral	GPT-4o	Llama3.1	Gemma2	Mistral	GPT-4o
<b>W/O Persona</b>	-	<b>45.75</b>	<b>57.50</b>	<b>16.52</b>	<b>58.03</b>	<b>54.00</b>	<b>55.00</b>	<b>12.46</b>	<b>73.64</b>
<b>Age</b>	Young	-1.70 ↓	+1.00 ↑	+12.02 ↑	-0.38 ↓	+6.54 ↑	+12.66 ↑	+20.27 ↑	-0.54 ↓
	Old	-0.86 ↓	+0.80 ↑	+5.03 ↑	-0.98 ↓	+6.47 ↑	+12.52 ↑	+13.59 ↑	-0.09 ↓
<b>Race</b>	White	+0.70 ↑	+0.44 ↑	+6.84 ↑	-0.40 ↓	+7.08 ↑	+11.77 ↑	+18.71 ↑	-0.08 ↓
	Black	-0.01 ↓	+0.00 ↑	-0.97 ↓	-0.86 ↓	+7.45 ↑	+12.08 ↑	+13.00 ↑	-0.27 ↓
<b>Skin Tone</b>	Light-skinned	-0.44 ↓	-1.04 ↓	+15.07 ↑	-0.24 ↓	+4.41 ↑	+10.24 ↑	+25.53 ↑	-0.07 ↓
	Dark-skinned	-0.96 ↓	-1.92 ↓	+20.10 ↑	-2.11 ↓	+2.92 ↑	+9.64 ↑	+30.25 ↑	-1.09 ↓
<b>Education Level</b>	Less than High School	-0.86 ↓	+0.87 ↑	+17.72 ↑	-0.74 ↓	+5.19 ↑	+9.90 ↑	+26.61 ↑	-1.78 ↓
	High School Graduate	-1.55 ↓	-0.33 ↓	+21.70 ↑	-0.68 ↓	+5.59 ↑	+12.29 ↑	+29.64 ↑	-0.43 ↓
	Associate Degree	+0.50 ↑	+0.56 ↑	+5.07 ↑	-0.24 ↓	+6.60 ↑	+11.60 ↑	+16.47 ↑	+0.22 ↑
	Bachelor’s Degree	-1.60 ↓	+0.87 ↑	+16.69 ↑	-0.10 ↓	+5.92 ↑	+12.20 ↑	+25.35 ↑	+0.16 ↑
	Doctoral Degree	-0.88 ↓	+0.74 ↑	+12.31 ↑	+0.08 ↑	+6.77 ↑	+12.79 ↑	+22.55 ↑	+0.81 ↑
<b>Profession</b>	Doctor	-0.15 ↓	-0.54 ↓	+13.62 ↑	-1.00 ↓	+5.84 ↑	+11.10 ↑	+24.78 ↑	-0.54 ↓
	Engineer	+0.25 ↑	-0.64 ↓	+25.05 ↑	-1.48 ↓	+5.32 ↑	+10.70 ↑	+32.50 ↑	-1.62 ↓
	Security Guard	+1.46 ↑	-0.56 ↓	+19.75 ↑	-1.98 ↓	+4.45 ↑	+10.16 ↑	+29.05 ↑	-1.37 ↓
	Cleaner	-0.82 ↓	-0.73 ↓	+6.32 ↑	-0.43 ↓	+6.51 ↑	+11.07 ↑	+19.03 ↑	-0.32 ↓
<b>Social Class</b>	Lower-Class	+0.50 ↑	+1.36 ↑	+9.00 ↑	-0.97 ↓	+6.13 ↑	+11.82 ↑	+19.04 ↑	-0.66 ↓
	Middle-Class	+0.17 ↑	-0.97 ↓	+12.72 ↑	-0.15 ↓	+5.46 ↑	+11.26 ↑	+23.40 ↑	-0.03 ↓
	Upper-Class	+1.09 ↑	+1.56 ↑	+1.79 ↑	+0.66 ↑	+6.76 ↑	+11.95 ↑	+12.27 ↑	-0.14 ↓
<b>Income Level</b>	Low-Income	+0.89 ↑	+1.44 ↑	+14.99 ↑	-0.69 ↓	+5.71 ↑	+11.37 ↑	+23.72 ↑	-0.52 ↓
	High-Income	+0.03 ↑	+0.01 ↑	+19.38 ↑	-0.06 ↓	+1.76 ↑	+8.85 ↑	+27.61 ↑	-0.64 ↓

Table 18: Rest Persona results for each model. The other persona group’s results are already shown in Table 4.

# Author Index

- Ajayi, Tunde Oluwaseyi, 99  
Aula-Blasco, Javier, 49  
Aycock, Seth, 99
- Balloccu, Simone, 1  
Barriere, Valentin, 177  
Bhanushali, Karan, 35  
Borisova, Ekaterina, 99
- Chen, Chengheng Li, 196  
Chirkova, Nadezhda, 99  
Choudhury, Monojit, 76  
Chung, Isaac, 133  
Cifuentes, Sebastian, 177  
Clavel, Chloé, 149  
Contreras, Hernan, 177
- Deutsch, Daniel, 204  
Dusek, Ondrej, 1
- Escoffier, Simon, 177
- Finkelstein, Mara, 204  
Freienthal, Linda, 133  
Freitag, Markus, 204
- Gkatzia, Dimitra, 1  
Gokrani, Aman, 162  
Gonzalez-Agirre, Aitor, 49  
Graiff, Cecilia, 149
- Hassan, Nazmul, 220  
Hohl, Fritz, 35
- Iyatomi, Hitoshi, 67
- Juraska, Juraj, 204
- Kamruzzaman, Mohammed, 220  
Karmim, Yannis, 177  
Kartáč, Ivan, 1  
Kasner, Zdeněk, 1
- Keleş, Kaan Efe, 189  
Kim, Gene Louis, 220  
Kim, Kyuhee, 196
- Lazzaroni, Ruggero Marino, 162  
Libovický, Jindřich, 23  
Lira, Hernan, 177
- Mahamood, Saad, 1  
Martí, Luis, 177  
Mujahid, Zain Muhammad, 99
- Nangia, Ayush, 162  
Nguyen, Hieu Minh, 220
- Ohashi, Takumi, 67  
Onderková, Kristýna, 1
- Perlić, Vladana, 99  
Pino, Renato, 177  
Platek, Ondrej, 1  
Pàmies, Marc, 49
- Riabi, Arij, 76  
Riley, Parker, 204
- Sagot, Benoît, 149  
Saha, Sougata, 76  
Schmidtová, Patrícia, 1  
Seddah, Djamé, 177  
Sotnikova, Anna, 196
- Türkmen, Arif, 189
- Vartampetian, Markarit, 99  
Villegas, Marta, 49
- Xu, Wenda, 204
- Zouhar, Vilém, 1, 204