

ExCAM: Explainable Cultural Awareness Metrics

Christoph Leiter^{1,2}, Haiyue Song³, Hour Kaing³, Jin Tei³,
Hideki Tanaka³, Masao Utiyama³, Steffen Eger^{1,2}

¹University of Mannheim, Germany,

²University of Technology Nuremberg, Germany,

³National Institute of Information and Communications Technology (NICT), Japan

Correspondence: christoph.leiter@uni-mannheim.de

Abstract

Evaluating the cultural awareness of large language models is crucial to ensure the fairness of generated text and the generalizability of applications across the world. Recent benchmarks explore cultural goods like food or values like behavior in stressful situations through the lens of question answering or text generation tasks. However, creating these benchmarks requires time-intensive and costly human annotations. Also, benchmarks that evaluate cultural awareness in free text are scarce and often rely on dated evaluation mechanisms. To address this gap, we introduce ExCAM, an **Explainable Cultural Awareness Metric**, which is, to our knowledge, the first dedicated evaluation metric that identifies, rates and explains cultural errors in instruction-output pairs. To train and evaluate ExCAM, we introduce *ExCAM40k*, a dataset comprising nine existing benchmarks that we reformat and enhance with synthetic errors. Compared to several baselines, including GPT-5, ExCAM achieves the highest error detection rate with up to 80% accuracy on a balanced test set. Therefore, ExCAM opens the pathway towards fine-grained and explainable cultural evaluation of free text.¹

1 Introduction

Recent advances in large language models (LLMs) have enabled impressive real-world applications, like code generation, tool usage and multilingual reasoning. However, recent benchmarks highlight shortcomings in LLMs’ alignment with cultural values and their knowledge about cultures. For example, Chiu et al. (2025) show that human performance still surpasses LLMs in answering culture-related questions. Also, Durmus et al. (2024); Sukiennik et al. (2025) show that LLMs’ default responses are more skewed towards US norms. While these benchmarks are valuable in tracking

¹We make our code available at <https://github.com/NL2G/ExCAM/>.

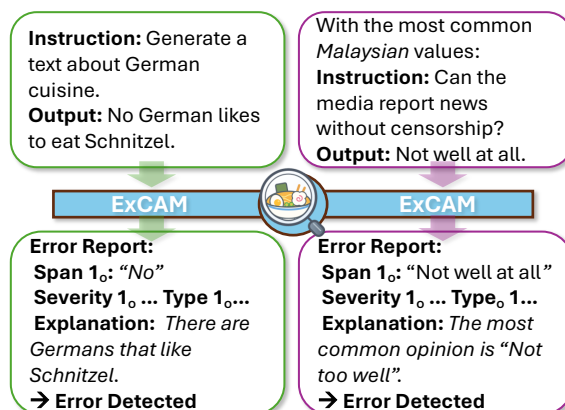


Figure 1: Evaluating the cultural awareness of an instruction and its generated output text with ExCAM. The error report contains details on every detected error. In the example, “O” indicates an error in the output (not in the instruction). **Left:** a *free-text* example with the cultural overgeneralization that *no* Germans like Schnitzel. **Right:** an *impersonation* example modified from GlobalOpinionQA (Durmus et al., 2024). In Malaysia, the most common answer to the instruction is “Not too well” instead of the given “Not well at all”.

the progress of cultural alignment, they have shortcomings: (1) they require costly human supervision, (2) publishing the benchmarks may lead to LLM contamination (Chen et al., 2025), (3) single benchmarks often have a limited focus (e.g., only material goods like food) and (4) their evaluation format follows a fixed scheme, e.g., question answering (QA).

Other strands of NLP address similar challenges with *automatic evaluation metrics* that judge the quality of the generated outputs; e.g., COMET (Rei et al., 2020) and GEMBA (Kocmi and Federmann, 2023b) for machine translation (MT) or MENLI (Chen and Eger, 2023) and BERTScore (Zhang et al., 2020) for summarization. To this end, we introduce ExCAM, an **Explainable Cultural Awareness Metric**, to automatically produce fine-grained measurements of cultural awareness in

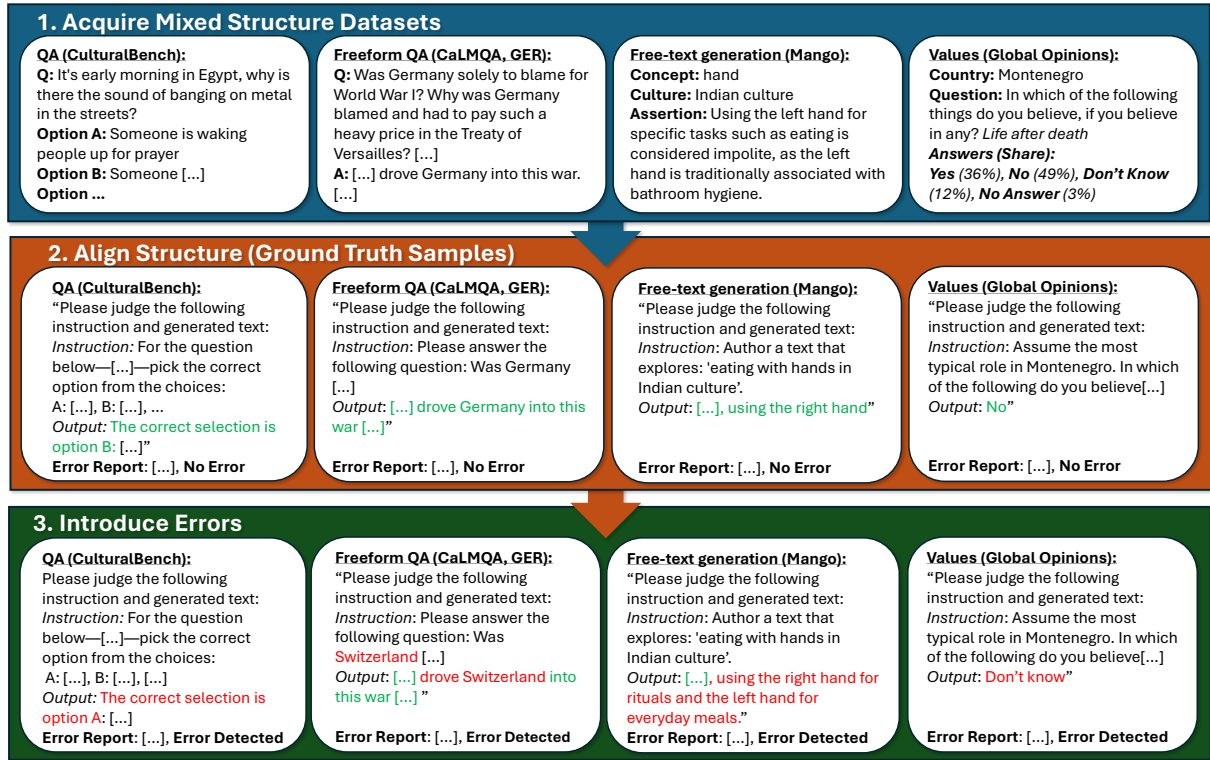


Figure 2: Example of the ExCAM data creation pipeline (also see §3). (1) We collect benchmarks for cultural awareness. We show examples of 4 datasets with unaligned data formats. (2) We load each dataset with an aligned structure, mapping each data sample into an instruction-output pair and wrapping it as an input prompt for ExCAM. This is our ground truth data, because it is sourced from existing verified datasets. Therefore, we assign an empty error report, where no error is detected.² (3) We generate negative examples that contain cultural errors, either by using given incorrect answer options from source datasets or by generating errors with different conditions using LLMs. The CaLMQA sample is originally in German and was translated for this figure.

LLM-generated content. We operationalize cultural awareness through the presence and severity of cultural errors in instruction-output pairs, as compared to other aspects such as cultural diversity, granularity, and specificity that are, e.g., used in MAKIEval (Zhao et al., 2025). Because ExCAM is an automated metric, at inference time, it can judge generations on arbitrary data *without* human quality labels (addressing cost issues and short-lived benchmarks). Additionally, it is trained on diverse datasets and input formats, enabling a focus on varying cultural aspects (addressing the specific focus of previous benchmarks).

Inspired by MQM (multidimensional quality metrics)-style evaluation metrics, (e.g. Lommel et al., 2014; Kocmi and Federmann, 2023a; Xu et al., 2023), ExCAM is a suite of LLMs that is tuned to create a fine-grained cultural error report (see Figure 1). This report further contains an explanation and fine-grained span, severity and

type predictions. To train ExCAM, we adapt data from 9 existing LLM benchmarks covering *QA* (answering a question with given answers), *free-form QA* (answering a question without given answers), *text generation* (generating a text about a given topic) and *impersonation* (assuming the role of a member of a given culture to answer questions related to norms and **values**). These benchmarks employ human verification, so we consider them as **ground truth** without cultural errors. Because they have various input and output formats, in a pre-processing step, we wrap their data with templates that transform them into a set of *instruction-output* pairs. Finally, for each data sample, we construct one or more error samples by either selecting an incorrect answer provided in the original dataset or by generating synthetic errors. During practical usage, any text can be judged with ExCAM by wrapping it in an instruction-output format, e.g., with simple instructions like “Generate a culturally correct text”. In summary, we make the following contributions:

²The empty report has the form {“Explanation”: “The instruction and generated text are without cultural errors.”}.

- ✓ **Cultural Evaluation Metric:** We introduce *ExCAM*, an **Explainable Cultural Awareness Metric**. To our knowledge, *ExCAM* is the first dedicated metric tuned to produce fine-grained measurements of cultural awareness in LLM-generated content. Besides error-detection, the generated reports can be used with a scoring heuristic to obtain quality scores.
- ✓ **Creating a Semi-Supervised Benchmark:** We create a new dataset *ExCAM40k* that consolidates existing datasets for cultural awareness evaluation, wraps them into instruction-output templates and introduces additional generated error samples. This dataset can be used to train and meta-evaluate cultural awareness metrics. We verify its validity with a human evaluation.
- ✓ **Meta-Evaluation of Cultural Awareness metrics:** Compared to several strong prompting-based baselines, our metric achieves the highest accuracy of error detection with 0.798, outperforming, e.g., GPT-5 (Singh et al., 2026) by more than 10%.
- ✓ **Out-of-Domain Evaluation:** We analyze performance differences of *ExCAM* across the 9 source datasets in in-domain and out-of-domain scenarios by training additional LoRAs in a leave-one-out setting. Additionally we evaluate it on the external dataset CARE (Guo et al., 2025b). In most settings it outperforms the baselines, highlighting the generalizability of our approach.

2 Related Work

Our work is related to evaluation metrics in other fields of generative AI and existing benchmarks of cultural awareness.

Evaluation Metrics Open-ended generation is commonly evaluated using lexical overlap, embeddings, or LLM judges (Sai et al., 2022; Celikyilmaz et al., 2021; Leiter et al., 2023; Li et al., 2024). *ExCAM* is inspired by the explainable MQM framework, which identifies error spans, types, and severities and aggregates them into a quality score (Lommel et al., 2014). For example, several modern metrics adapt it for MT evaluation (Kocmi and Federmann, 2023a; Xu et al., 2023; Junczys-Dowmunt, 2025).

Cultural Awareness Benchmarks Several surveys summarize LLM evaluation on culture-related tasks (Pawar et al., 2025; Adilazuarda et al., 2024; Liu et al., 2025). They show that existing works evaluate culture through specific *proxies* (Adilazuarda et al., 2024), i.e., they pick specific culture-related aspects as their focus, for example, *food*, e.g., WorldCuisines for visual QA (Winata et al., 2025). Benchmarks for cultural awareness also vary based on the task type like QA, freeform QA or impersonation. Our work shifts focus from creating new benchmarks to defining a metric that can automatically test cultural awareness in LLMs. Because *ExCAM* is promptable at inference time, it can be used for many cultural aspects and task types at once. Few other cultural awareness benchmarks evaluate generated free text. CaLMQA (Arora et al., 2025), a benchmark for culture-related long-form QA, uses VERIScore (Song et al., 2024), a framework for fact verification, and GPT-4o (OpenAI, 2024b) to judge the relevance of generated answers. NativQA (Hasan et al., 2025) is a benchmark containing free-form questions originating from different regions. They evaluate the quality with reference answers and BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), BERTScore and GPT-4o. CURE (Vo and Koyejo, 2025) evaluates LLMs on their ability to grade the social acceptability of a situation on four different datasets. Among others, it is evaluated by an LLM as a judge that grades properties (one for each of 4 base datasets) of the reasoning process, like coverage (does the reasoning involve every aspect) and coherence (whether the aspects of the reasoning are well connected). CURE’s main focus is the judgment of social situations. Also, the LLM-as-a-judge metrics prompt definitions judge reasoning quality but will not necessarily reveal and highlight cultural mistakes. Lastly, MAKIEval (Zhao et al., 2025) is an automated framework to evaluate the cultural awareness of generated free text. LLMs are systematically prompted with culturally diverse instructions. Then, all generated culturally relevant entities are extracted with GPT-4o-mini (OpenAI, 2024a) and their Wikidata entries are parsed. Cultural alignment is evaluated by aspects such as *entity diversity* and *how many entities match the culture specified* in the input prompt. In contrast to the evaluation techniques used in these benchmarks, *ExCAM* is a tuned, reference-free metric dedicated to cultural evaluation that combines evaluation knowledge from several existing works. Another key

difference from MAKIEval is the evaluation focus: we focus on a fine-grained per-sample evaluation that returns a report of cultural errors. This also includes customs and norms that are difficult to evaluate with entity matching, e.g., greetings. Another related work is CARE (Guo et al., 2025b). CARE and ExCAM are complementary but fundamentally distinct in purpose. CARE is a human-annotated preference dataset designed to study and improve multilingual cultural alignment, whereas ExCAM (which is inspired by MQM-style machine translation metrics) is a dedicated, fine-grained metric that evaluates cultural awareness through cultural errors in arbitrary instruction-output pairs.

3 Methodology

In this section, we describe the output format and construction pipeline of ExCAM. A (reference-free) metric M assigns a quality score s to an instruction i and the respective output o , i.e., $M(i, o) = s$. s describes how well i and o fulfill evaluation aspects that are inherent to the metric or passed as external arguments.

For ExCAM, this aspect is cultural awareness, operationalized through cultural errors. Following the proxy-based framing of culture by Adilazuarda et al. (2024), we view samples as culture-related if they contain at least one explicit or implicit demographic proxy that relates it to a group of people, such as region, ethnicity and religion. We define a cultural error as a factually incorrect or contextually inappropriate statement or behavior concerning the culture or group identified by such a proxy. To do so, we design ExCAM to return fine-grained explanations in *MQM-style*, i.e., it returns error types, error count, error severities, error spans and an explanation (see Figure 1 for an example). Depending on the use-case, it can be beneficial to handle cultural errors in the instruction and in the output either jointly or separately. A user may want to use ExCAM to test if their generation is without cultural issues. In that case, it is beneficial to take into account instruction errors. When rating a whole generation model, it is more suitable to only take into account the errors in the output, as the system should not be punished for errors in the instruction, although the effect will even out because every model receives the same incorrect instruction. ExCAM’s error reports contain separate fields for instruction and output errors (see the Paragraph “Error Construction”). In our main

evaluation, we compute joint scores for instruction and output errors. Appendix N additionally discusses the effect this has on the evaluation result and reports output-only error results.

The score s is computed as error count weighted by severity. We include the severities *minor* (-1) and *major* (-5).³ In other words, the best possible s is 0 and for every cultural error in the instruction or output points are deducted. A cultural error could, for example, be an incorrect fact or an incorrect representation of the most common opinion when assuming a role. The function of the score is to compare system outputs (by averaging across outputs, entire systems). Like MQM-heuristics for MT, the actual score does not take into account the error type, span, and explanation. These provide additional explainability (and increase the length of the reasoning chain in LLMs). Because severity can be subjective, our main evaluation mode is the prediction of error absence/presence, with severity scores evaluated separately through correlation analyses (see §4, *Evaluation*).

ExCAM is a suite of LLMs that we fine-tune to return the described schema. We produce the training data in three steps that are explained in the following paragraphs: (1) selecting verified ground truth data that is error-free, (2) wrapping these ground truth samples into a uniform instruction-output structure and (3) introducing fine-grained errors to the ground truth samples to produce diverse samples that should receive negative scores. Figure 2 gives an overview of this pipeline.

Ground Truth from Existing Benchmarks

First, we select data from existing cultural awareness benchmarks that have been verified with some form of human supervision (see §4). Because of this supervision, we make the assumption of cultural correctness and set $\text{ExCAM}(\text{verified data}) = 0$, which is the highest score.

Data Formatting The generation of culturally aware content has many sub-tasks like *culturally aware MT*, *summarization*, *text generation* and *QA*. ExCAM should judge instruction-output pairs, hence we wrap the selected data into a suitable common prompt structure (see Appendix A). That means, we transform our data such that we can write our equation as $\text{ExCAM}(\text{input}, \text{output}) = 0$ (where input and output are wrapped in a joint

³These values are used with MT MQM heuristics, for example, by GEMBA-MQM (Kocmi and Federmann, 2023a).

prompt). Figure 2 shows one example for each of the four dataset types that we include: QA, freeform QA, free-text generation and impersonation.⁴ To increase generalizability, we use 50 different paraphrases (a high number chosen to increase diversity) for the instruction and output templates that are selected randomly (but stratified) for every data sample. Appendix D shows samples of dataset-specific transformations.

Error Construction In the previous steps, we have constructed data samples that receive a perfect quality score. However, ExCAM can only be effective if it can detect and judge cultural errors. We construct the training data in two ways: (1) constructing **hard negative samples** from incorrect options provided in the source datasets (e.g., incorrect answer options for QA) and (2) constructing soft negative samples with LLMs when no incorrect options are provided. For each sample that is constructed from QA with answer options, we include an additional positive and negative QA pair with a single answer (freeform QA). As these samples are not generated, we assign a default severity of major (-5) to every hard error and template-based explanations that describe the correct answer that should have been picked instead.⁵

Soft negative examples are created by prompting LLMs to introduce cultural errors to the instruction and/or output. There is one prompt variant to introduce minor and major errors, i.e., the distinction is assigned by the LLM. Based on it, we assign -1 respectively -5 to the soft errors. We also ask the model to freely suggest an error type like “stereotyping” or “misrepresentation” that should help high-level explainability. Additionally, we ask the model to create an explanation of the errors it introduces, as well as judgments on whether the sample is indeed related to culture. We filter out samples that are found as *not culture-related* in this step. For the ground truth error reports, we include the error type, error severities, error spans and explanation. Here, errors are reported separately for the instruction and the output. Appendix C describes the templates and Appendix E shows a full example of an ExCAM error report.

⁴Freeform QA is closely related to free-text generation, as both involve open-ended text generation.

⁵For CulturalBench (Chiu et al., 2025), the hard examples without answer options coincide with the CulturalBench-hard subset. However, for us “hard” vs. “soft” refers to manually introduced vs. LLM-generated error cases and not to difficulty.

Type	Instruction	Output
Soft	Please state your answer clearly for: ‘What is the history of the language in Bangladesh?’	Answer: [...] Until the 14th century, there was little linguistic difference between Bangla and Urdu.
Hard	What is the most important family holiday in China? Pick from these options: Option A: christmas; [...]; Option C: lunar new year; [...]	Option A: christmas

Table 1: Example comparing a soft (LLM-generated) error and a hard error (built from QA-dataset answers). Errors are marked in red.

Score Heuristic Let R be an error report generated by ExCAM. Also, let sev_i denote the error severity of error i , for $i = 1, \dots, n$.

In practice, for cases where the metric does not find any error, we take the certainty of the answer as a secondary rating:

$$s^* = \sum_{i=1}^n sev_i \quad s = \begin{cases} s^* & s^* < 0 \\ p(R) & s^* = 0 \end{cases} \quad (1)$$

$p(R)$ is the cumulative probability of the entire token sequence of the error report normalized by length. Therefore, the highest possible score is 1 (samples without errors and the highest certainty of the token sequence). This fallback helps resolve score ambiguities in cases with hard negative samples (see Appendix K).

4 Experiment Setup

In this section, we describe the source datasets from which we construct *ExCAM40k*, the hyperparameters of the error construction, the fine-tuning process of ExCAM, the evaluation setup and a human validation setup.

Base Datasets For QA, we include (1) BLEND (Myung et al., 2024), a large-scale QA benchmark for everyday knowledge across multiple cultures and languages, (2) CulturalBench (Chiu et al., 2025), a benchmark for cultural knowledge, and (3) INCLUDE, a benchmark that measures if LLMs can answer multilingual exam questions sourced

from various cultures (Romanou et al., 2025). For **freeform text generation**, we include Mango (Nguyen et al., 2024), a dataset of LLM-generated assertions on diverse cultural topics that was evaluated through human judgments on sampled assertions collected via Amazon Mechanical Turk. We reframe this dataset by creating instructions that prompt for text generation on the topic of the assertion while we consider the output as generated text. For **freeform QA**, we include the subset of (1) CaLMQA (Arora et al., 2025) that has outputs provided on Hugging Face. It is a dataset with multilingual questions sourced from internet forums and freelancers and generated answers. The authors of CaLMQA verify the quality of these generations with human annotations in 5 languages. Another freeform QA dataset is (2) NativQA (Hasan et al., 2025) with question-answer pairs sourced from search engines and verified by human annotators. For **impersonation** (where the LLM is prompted to assume a typical role in order to test for represented values), we include (1) NormAd (Rao et al., 2025) which uses LLMs to transform etiquette descriptions from Mosaica (2024) into scene descriptions including proper or improper behavior for a culture (probed in human evaluation). Then LLMs are evaluated on judging if a scene description fits the etiquette. We also include (2) GlobalOpinionQA (Durmus et al., 2024), a dataset that uses global survey data to examine the extent to which LLM responses can match human responses and (3) EPIC (Freunda et al., 2023), a dataset that benchmarks irony detection capability for five English speaking countries. For NormAd and GlobalOpinionQA, we ask the LLM to assume the most typical role from the respective countries. For EPIC, we provide further demographic data from their dataset as context.

As described in §3, we format every data sample of these datasets in an *instruction-output* form. Appendix D shows one example of this transformation for every dataset. Our templates are built hierarchically, thus new types of datasets and transformations can easily be added to the evaluation. The licenses of models and datasets are in Appendix O.

Baselines As baselines, we evaluate the pre-trained LLMs *Qwen3-14B* (Yang et al., 2025), *DeepSeek-R1-Distill-Qwen-14B* (Guo et al., 2025a), *Gemma3-27B* (Gemma Team, 2025), *Phi3-medium-4k* (Abdin et al., 2024b), *Phi-4* (Abdin et al., 2024a) and *Mistral-Small-24b*

Instruct (Mistral AI, 2025) (see Appendix B for a comparison of model sizes, libraries and sampling parameters), and separately GPT-5 (2025-08-07 version). For this evaluation we define two prompts: a *counting* prompt that instructs the LLM to count the errors in the input and output and a *binary* prompt that simply differentiates error presence and absence. The full prompts are shown in Appendix G. Additionally, we test the baselines with the same disambiguation mechanism as ExCAM and use the sequence probability as score where 0 would be returned.

Error Generation, Cluster Setup and Tuning

ExCAM40k consists of 12.4k hard and 6.3k soft (LLM-generated) error samples with 18.8k error-free samples, where the soft errors are generated with Qwen3.5-122B-A10B (Qwen Team, 2026) in thinking mode (Appendix F for preprocessing steps and detailed statistics).

For ExCAM, we tune LoRAs (Hu et al., 2022) for Phi-3-medium-4k-instruct and Gemma3-27B with supervised fine-tuning (SFT) on the error reports in the training data (see Appendix B for training details).

Evaluation As our main evaluation, we report the scaled accuracy, defined as $2 * \text{Acc} - 1$ (thus, on our balanced dataset, random accuracy of 0.5 maps to 0), of predicting whether a cultural error is present or not. Additionally, we test the correlation of ExCAM’s scores with the synthetic ground truth scores (based on error count and severity). To do so, we report Kendall correlation and tie-calibrated accuracy (Deutsch et al., 2023), limited to 1,000,000 pairs. Further, we ablate the effect on out-of-domain predictions by training an additional 9 LoRAs per model in a leave-one-out setting and testing their performance separately on data constructed from each source dataset. As a significance test, we use a paired permutation test with $n=1000$ and reduce the number of tie acc. pairs to 1000.

Human Validation of the ExCAM Dataset To verify the validity of the synthetic errors in *ExCAM40k*, we perform a human evaluation, where 4 annotators (2 native German, 1 Japanese and 1 Chinese annotator) rate 100 samples (25 from each dataset for which we generate soft errors).⁶ 80 samples are in English and annotated by 3 annotators and 8 German, 6 Japanese and 6 Chinese samples

⁶This sample provides sufficient evidence for the targeted validation of error-introduction success.

Criterion	3/3	2/3	1/3	0/3
Culture rel.	87.5	10.0	2.5	0.0
Error valid (strict)	61.2	22.5	11.2	5.0
Error valid (partial)	93.8	6.2	0.0	0.0
Expl. valid	78.8	18.8	1.2	1.2

Table 2: Vote distributions over samples annotated by all 3 annotators (N=80). Values are percentages. *Strict* indicates that all fine-grained error spans are truly part of an error.

each are annotated by two annotators. Among other fields, annotators assess (1) the cultural relation of the source data (boolean), (2) whether the modification introduces a culture-related error (valid, partially valid, not valid), (3) whether the generated explanation is correct (valid, partially valid, not valid) and (4) agreement with the error severity label. *Partially valid error* is selected if parts of the error span selection highlight unnecessary words. Partially valid explanations are selected if parts of the error are not addressed or in some parts described incorrectly. Table 2 shows the distribution of annotator selections for the positive classes. For example, all annotators agree in 61.2% of cases that the error spans highlight only a cultural error and in 93.8% of cases that a true cultural error is contained in the sample. In none of the cases do two or more annotators agree that no cultural error is present. See Appendix H for further annotation details and results regarding severity, error types and explanation persuasiveness.

5 Results

Main Results: ExCAM40k Table 3 reports the scaled accuracy, Kendall correlation and tie-calibrated accuracy for the different baselines and ExCAM variants on the ExCAM40k test set. Our metric ExCAM_r (Gemma3)⁷ achieves the best performance across all measures (see Appendix J for the confusion matrix and Appendix N for an evaluation that does not take source errors into account), followed by ExCAM_r (Phi3) (0.099 difference in scaled acc.). Size-wise, Gemma3-27B is the largest model, followed by Mistral-24B. For all baselines, the *counting* prompt (cnt) yields the best results. The weakest performance of the tested models is achieved by Phi-3m-4k and Mistral-24B. For the best baseline, we report the correlations with the

⁷ExCAM_r is denoted with r, because it returns report-style outputs.

Model	S.Acc.	Kd	T.Acc.
DS-R1-14B _{cnt}	0.331	0.305	0.474
Gemma3-27B _{cnt}	0.311	0.284	0.455
Mistral-24B _{cnt}	0.282	0.246	0.452
Phi-3m-4k _{cnt}	0.270	0.303	0.458
Phi-4 _{cnt}	0.351	0.349	0.495
Phi-4 _{cnt+p}	0.346	0.318	0.491
Qwen3-14B _{cnt}	0.280	0.285	0.464
ExCAM _r (Gemma3)	0.591	0.569	0.680
ExCAM _r (Phi3)	0.492	0.493	0.617

Table 3: Scaled **accuracy** (0 is random), **Kendall**, and **tie calibrated accuracy** between metric scores and the ground truth scores of our test set. For each baseline model, we only show the **counting** prompt baseline (as it achieved the highest scores). The “+p” indicates that we take the sequence probability as score, when the score is 0. Underlined values are significantly higher than all other values in the respective column ($p < 0.05$).

probability as fallback score (+p), which has an adverse effect. ExCAM improves scaled accuracy by 0.280 for Gemma3 and 0.222 for Phi3, so for both models tuning has a benefit. Appendix P explores the results for 5 different languages.

Out-of-Domain Performance Figure 3 shows an overview of ExCAM’s out-of-domain (ood) performance. For each column (besides ALL and AVG), a separate LoRA is trained in a leave-one-out manner. The ood performance is significantly better than all baselines for BLENd, CulturalBench, INCLUDE and Mango. For CaLMQA and GlobalOpinionQA, the baselines are outperformed, but not significantly. For EPIC, the ood performance is random for all models. EPIC’s semantic proxy, irony detection, is very distinct from the other datasets, offering a rather unaligned cultural relation. Appendix Q shows the LoRA performance on all subsets. To further test ExCAM’s generalization, we evaluate ExCAM_r (Gemma3) and the strongest baselines (DeepSeek, Gemma3 and Phi-4) on unmodified CARE responses with human cultural-appropriateness ratings (Guo et al., 2025b). ExCAM_r achieves significantly higher tie-calibrated accuracy than the baselines ($p < 0.05$; see Appendix M for more details), further highlighting its generalization.

Comparison to GPT-5 We compute the performance of in-domain ExCAM and GPT-5 with the error counting prompt and two additional prompts

Scaled Acc.	ExCAM _r (Gemma3)	0.591*	0.374	0.579*	0.254	0.649*	-0.002	0.363	0.263*	0.768*	0.274	0.222
	ExCAM _r (Phi3)	0.492	0.258	0.525	0.140	0.547	0.002	0.196	0.110	0.385	0.096	0.319
	B: Gemma3-27B _{ent}	0.311	0.304	0.453	0.239	0.527	0.000	0.271	0.208	0.455	0.298	0.285
	B: Phi-3m-4k _{ent}	0.270	0.254	0.425	0.045	0.533	0.006	0.355	0.108	0.443	0.076	0.299
		ALL (ID)	AVG (OOD)	BLEND	CaLMQA*	CulturalBench	EPIC	GlobalOpinionQA	INCLUDE	Mango*	NativQA*	Normad**
		Dataset										

Figure 3: Out-of-domain LoRA performance of ExCAM compared to the Baselines. Each column shows the test set we evaluate on and, for ExCAM, the rows show the performance of LoRAs that were trained on all train sets except the column dataset. For example, BLEND with ExCAM_r (Gemma3) means that a Gemma3 LoRA was trained on all ExCAM train sets besides BLEND and evaluated on only the BLEND test set. ALL (ID) shows the in-domain performance from Table 3. AVG (OOD) shows the average of all ood columns. Cell values with * are significant in a column ($p \leq 0.05$). Datasets with soft errors are marked with * behind their column name.

on a subset of 100 samples per dataset, i.e., on 900 samples in total. The two additional prompts aim to reduce the output-format mismatch between ExCAM and the baselines: (1) a zero-shot prompt that describes ExCAM’s structured error-report schema and (2) a few-shot prompt that additionally provides one error-containing example from each source dataset. Both prompts instruct GPT-5 to produce an explanation followed by the same structured error report as ExCAM and a score is determined using the same scoring heuristic.

Table 4 shows the performance of in-domain ExCAM on the different data subsets. ExCAM_r (Gemma3) outperforms all tested GPT-5 prompts on 7 out of 9 test subsets, 3 of which are significant, and on the macro average (due to the small subset size, we use a significance threshold of $p \leq 0.1$ for this comparison). Only on CaLMQA, two GPT-5 prompts and the few-shot prompt on INCLUDE achieve higher scores than in-domain ExCAM. For EPIC, only ExCAM and few shot GPT-5 recognize some errors in culture-conditioned irony detection as cultural errors, highlighting the importance of fine-tuning for this setting. The zero-shot error-report prompt performs substantially worse than the other approaches. Few-shot examples improve its overall score from 0.213 to 0.427, approximately matching the error-counting prompt at 0.429, although their relative performance varies across datasets.

Explanation Quality We evaluate ExCAM’s metric explanations separately for hard and soft errors. For *hard errors*, the ground truth is template-based and contains a text stating which answer should be given instead. In 87% respectively 78%

of samples without errors, ExCAM_r (Gemma3) and ExCAM_r (Phi3) correctly predict no error. Likewise, they correctly detect an error and return an explanation for 72% respectively 73% of error samples. We evaluate these explanations semantically by using Qwen2.5-7B-Instruct (Qwen Team, 2024) to compare them with the ground-truth explanations. For ExCAM_r (Gemma3), 70% of the explanations for correctly detected hard errors are judged consistent with the ground truth. Appendix I describes the evaluation in detail.

Soft errors are generated, so during the validation process described above, the same annotators rate the similarity of 100 explanations for both in-domain ExCAM variants with the ground truth explanations on a 5-point Likert scale, where 5 is the best. We then filter for samples where the ground truth is valid as per majority vote ($n=90$). Among these, ExCAM_r (Gemma) predicts an error in 67% of cases and (Phi3) in 60% of cases. The average of ExCAM_r (Gemma) is 4.01 and for ExCAM_r (Phi3) 3.88, highlighting good explanation quality. In Appendix L, we compute the tie-calibrated accuracy for samples where the annotators labeled the error spans and severities as correct.

Overall, while there are failure cases, the explanations can inform human processes of judging the cultural awareness of generated text.

6 Conclusion

We present ExCAM, an **Explainable Cultural Awareness Metric**. Trained on a novel dataset that is unified from recent cultural evaluation benchmarks and enhanced with synthetic error samples, ExCAM judges instruction-output pairs with regard

Model	ALL	BLenD	CaLMQA	EPIC	CultB	INCLUDE	Mango	NativQA	NormAd	GOQA
GPT-5 _{rep} (zero-shot)	0.213	0.160	0.200	-0.080	0.240	0.240	0.180	0.360	0.500	0.120
GPT-5 _{rep} (few-shot)	0.427	0.300	0.380	0.200	0.420	0.560	0.380	0.420	0.680	0.500
GPT-5 _{cnt}	0.429	0.580	0.260	0.000	0.660	0.240	0.700	0.540	0.580	0.300
ExCAM _r (Gemma3)	0.596*	0.740*	0.240	0.300	0.800*	0.420	0.780	0.600	0.700	0.780*
ExCAM _r (Phi3)	0.453*	0.600*	0.071	0.180	0.600*	0.220	0.800	0.140	0.620	0.840*

Table 4: Scaled accuracy of GPT-5 prompting variants and ExCAM on balanced 100-sample subsets. Zero represents random performance. ALL is the macro-average across datasets; *rep* and *cnt* denote error-report and error-counting prompts. CultB and GOQA denote CulturalBench and GlobalOpinionQA. Bold indicates the highest result in each column. The * indicates ExCAM values that are significantly higher than all GPT-5 results ($p \leq 0.1$)

to their cultural correctness and returns detailed reports explaining the exact errors. By automating cultural evaluation, ExCAM addresses cost issues with collecting human annotations for newer benchmarks. Because ExCAM can evaluate newly collected inputs without requiring a fixed public test set, it helps guard against benchmark-specific test-set contamination. Another benefit of ExCAM is that it is trained on different cultural evaluation tasks and can take arbitrary input formats (wrapped in instruction-output structure), which addresses a shortcoming of, e.g., QA benchmarks. Our experiments show that ExCAM outperforms several LLM-based baselines in in- and out-of-domain settings. These results highlight the capability of ExCAM to aid the verification of LLM-generated content, the construction of novel, flexible benchmarks and the general detection of culturally problematic content.

Limitations

One limitation of ExCAM is that, when no errors are detected, we return the sequence probability as a score, which can be interpreted as certainty but does not provide textual explanations. Further, our explanations are evaluated for plausibility (which is useful for accessibility and understandability) rather than faithfulness to model internals. Our soft training errors are synthetically generated and may therefore not fully reflect naturally occurring human errors. Hard errors derived from incorrect options in curated source datasets partly mitigate this issue. Because *ExCAM40k*’s training and test splits share the same construction pipeline, performance may partly reflect construction-specific patterns, particularly for LLM-generated soft errors (although our human evaluation did not observe such artifacts). Our leave-one-dataset-out experiments and evaluation on CARE responses provide evidence of generalization beyond the main

test set. Future work could further probe dataset-specific cues by manually rewriting synthetic errors. We also cannot fully rule out overlap between the source datasets and the evaluated models’ pretraining data, although our instruction-format transformations reduce direct textual overlap. Another limitation is potential self-preference in LLM-based metrics (Panickssery et al., 2024): an evaluator may favor outputs generated by itself or related models. This could bias ExCAM when evaluating outputs from its own base-model family. Using evaluators from multiple model families may help reduce this effect. Further, ExCAM is limited by the cultural knowledge acquired during pre-training and fine-tuning. It may therefore fail on cultural knowledge that is absent from, or cannot be inferred from, these sources. We also note that there is no clear-cut separation between minor and major cultural errors and that it also depends on the recipients’ views. Here, we left the “decision” with the error generation LLM that might itself be biased. Lastly, for impersonation samples, we treat the respective dataset responses as ground truth. This may overemphasize majority opinions within a country and underrepresent within-culture diversity. Future work may try to reflect more diverse opinions from value surveys.

Ethical Considerations

ExCAM introduces the field of fine-grained automated metrics for cultural evaluation. We believe that this can help the development of fairer and more considerate LLMs. As stated in the limitations, due to its training setup, ExCAM may be biased towards the majority opinions in each country. If it is used to evaluate a stance, this should be taken into consideration. Additionally, synthetic cultural errors can contain offensive content. We do not mitigate this, as offensive content is a valid part of the cultural errors that we are identifying.

Acknowledgments

The first author conducted part of this research during an internship at the National Institute of Information and Communications Technology (NICT), Japan. This work was partially supported by JSPS KAKENHI Grant-in-Aid for Early-Career Scientists 25K21290. The NLLG group (first and last author) gratefully acknowledges support from the Federal Ministry of Education and Research (BMBF) via the research grant “Metrics4NLG” and the German Research Foundation (DFG) via the Heisenberg Grant EG 375/5-1. The authors gratefully acknowledge the scientific support and HPC resources provided by the Erlangen National High Performance Computing Center (NHR@FAU) of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU). The hardware is funded by the German Research Foundation (DFG). We partly used AI assistants for coding and writing assistance.

References

- Marah I Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio CT Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, and 8 others. 2024a. [Phi-4 technical report](#). Technical Report MSR-TR-2024-57, Microsoft.
- Marah I Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Hassan Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Caio César Teodoro Mendes, Weizhu Chen, Vishrav Chaudhary, Parul Chopra, and 66 others. 2024b. [Phi-3 technical report: A highly capable language model locally on your phone](#). Technical Report MSR-TR-2024-12, Microsoft.
- 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.
- 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, Vienna, Austria. Association for Computational Linguistics.
- Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2021. [Evaluation of text generation: A survey](#). Preprint, arXiv:2006.14799.
- Simin Chen, Yiming Chen, Zexin Li, Yifan Jiang, Zhongwei Wan, Yixin He, Dezhi Ran, Tianle Gu, Haizhou Li, Tao Xie, and Baishakhi Ray. 2025. [Benchmarking large language models under data contamination: A survey from static to dynamic evaluation](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 10080–10098, Suzhou, China. Association for Computational Linguistics.
- Yanran Chen and Steffen Eger. 2023. [MENLI: Robust evaluation metrics from natural language inference](#). *Transactions of the Association for Computational Linguistics*, 11:804–825.
- 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.
- Daniel Deutsch, George Foster, and Markus Freitag. 2023. [Ties matter: Meta-evaluating modern metrics with pairwise accuracy and tie calibration](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12914–12929, Singapore. Association for Computational Linguistics.
- 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*.
- Simona Frenda, Alessandro Pedrani, Valerio Basile, Soda Marem Lo, Alessandra Teresa Cignarella, Raffaella Panizzon, Cristina Marco, Bianca Scarlini, Viviana Patti, Cristina Bosco, and Davide Bernardi. 2023. [EPIC: Multi-perspective annotation of a corpus of irony](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13844–13857, Toronto, Canada. Association for Computational Linguistics.
- Gemma Team. 2025. [Gemma 3 technical report](#). Preprint, arXiv:2503.19786.

- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhushu Li, Ziyi Gao, Aixin Liu, and 175 others. 2025a. [Deepseek-r1 incentivizes reasoning in llms through reinforcement learning](#). *Nature*, 645(8081):633–638.
- Geyang Guo, Tarek Naous, Hiromi Wakaki, Yukiko Nishimura, Yuki Mitsufuji, Alan Ritter, and Wei Xu. 2025b. [CARE: Multilingual human preference learning for cultural awareness](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32866–32895, Suzhou, China. Association for Computational Linguistics.
- Md. Arif Hasan, Maram Hasanain, Fatema Ahmad, Sahinur Rahman Laskar, Sunaya Upadhyay, Vrunda N Sukhadia, Mucahid Kutlu, Shammur Absar Chowdhury, and Firoj Alam. 2025. [NativQA: Multilingual culturally-aligned natural query for LLMs](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 14886–14909, Vienna, Austria. Association for Computational Linguistics.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [LoRA: Low-rank adaptation of large language models](#). In *International Conference on Learning Representations*.
- Marcin Junczys-Dowmunt. 2025. [GEMBA v2: Ten judgments are better than one](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 926–933, Suzhou, China. Association for Computational Linguistics.
- Tom Kocmi and Christian Federmann. 2023a. [GEMBA-MQM: Detecting translation quality error spans with GPT-4](#). In *Proceedings of the Eighth Conference on Machine Translation*, pages 768–775, Singapore. Association for Computational Linguistics.
- Tom Kocmi and Christian Federmann. 2023b. [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.
- 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*, pages 117–138, Bali, Indonesia. Association for Computational Linguistics.
- Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024. [Llms-as-judges: A comprehensive survey on llm-based evaluation methods](#). *Preprint*, arXiv:2412.05579.
- 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.
- 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.
- Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. [Multidimensional quality metrics \(mqm\): A framework for declaring and describing translation quality metrics](#). *Tradumàtica tecnologies de la traducció*, pages 455–463.
- Mistral AI. 2025. [Mistral-small-24b-instruct-2501](#). <https://huggingface.co/mistralai/Mistral-Small-24B-Instruct-2501>.
- Mosaica. 2024. [The cultural atlas](#). <https://culturalatlas.sbs.com.au/>. Online; accessed 20.12.2025.
- 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.
- 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.
- OpenAI. 2024a. [GPT-4o mini: Advancing cost-efficient intelligence](#). Accessed 2026-06-01.
- OpenAI. 2024b. [Gpt-4o system card](#). *Preprint*, arXiv:2410.21276.
- Arjun Panickssery, Samuel R. Bowman, and Shi Feng. 2024. [LLM evaluators recognize and favor their own generations](#). In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- 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.

- Siddhesh Pawar, Junyeong Park, Jiho Jin, Arnav Arora, Junho Myung, Srishti Yadav, Faiz Ghifari Haznitrana, Inhwa Song, Alice Oh, and Isabelle Augenstein. 2025. [Survey of cultural awareness in language models: Text and beyond](#). *Computational Linguistics*, 51(3):907–1004.
- Qwen Team. 2024. [Qwen2.5: A party of foundation models](#).
- Qwen Team. 2026. [Qwen3.5: Towards native multi-modal agents](#).
- 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.
- Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. [COMET: A neural framework for MT evaluation](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.
- Angelika Romanou, Negar Foroutan, Anna Sotnikova, Sree Harsha Nelaturu, Shivalika Singh, Rishabh Maheshwary, Micol Altomare, Zeming Chen, Mohamed A. Haggag, Snegha A, Alfonso Amayuelas, Azril Hafizi Amirudin, Danylo Boiko, Michael Chang, Jenny Chim, Gal Cohen, Aditya Kumar Dalmia, Abraham Dires, Sharad Duwal, and 38 others. 2025. [INCLUDE: Evaluating multilingual language understanding with regional knowledge](#). In *The Thirteenth International Conference on Learning Representations*.
- Ananya B. Sai, Akash Kumar Mohankumar, and Mitesh M. Khapra. 2022. [A survey of evaluation metrics used for nlg systems](#). *ACM Comput. Surv.*, 55(2).
- Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, Akshay Nathan, Alan Luo, Alec Helyar, Aleksander Madry, Aleksandr Efremov, Aleksandra Spyra, Alex Baker-Whitcomb, Alex Beutel, Alex Karpenko, and 467 others. 2026. [Openai gpt-5 system card](#). *Preprint*, arXiv:2601.03267.
- Yixiao Song, Yekyung Kim, and Mohit Iyyer. 2024. [VeriScore: Evaluating the factuality of verifiable claims in long-form text generation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9447–9474, Miami, Florida, USA. Association for Computational Linguistics.
- Nicholas Sukiennik, Chen Gao, Fengli Xu, and Yong Li. 2025. [An evaluation of cultural value alignment in llm](#). *Preprint*, arXiv:2504.08863.
- Truong Vo and Sanmi Koyejo. 2025. [Cure: Cultural understanding and reasoning evaluation - a framework for "thick" culture alignment evaluation in llms](#). *Preprint*, arXiv:2511.12014.
- 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 Willie, Candy Olivia Mawalim, Cheng Ching Lam, Daud Abolade, Emmanuele Chersoni, and 32 others. 2025. [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.
- 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.
- 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. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. [Bertscore: Evaluating text generation with bert](#). In *International Conference on Learning Representations*.
- Raoyuan Zhao, Beiduo Chen, Barbara Plank, and Michael A. Hedderich. 2025. [MAKIEval: A multilingual automatic WiKidata-based framework for cultural awareness evaluation for LLMs](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 23104–23136, Suzhou, China. Association for Computational Linguistics.

A Evaluation Prompts

Table 5 shows the evaluation prompt used for Ex-CAM.

B Model Parameters & Tuning Configuration

Table 6 shows further implementation details of our project, including the baseline models reported in our experiments and the main software libraries

Name	Prompt
System Prompt	You compute a metric that judges the cultural correctness of a generation task.
Report Style	Please judge the following instruction and generated text: Instruction: {Instruction} Text: {Text} Return an error report in JSON format.

Table 5: Evaluation prompts used for ExCAM_r. What constitutes a score or an error report is learned during training.

used for inference and training. For the SFT training, we supervise the models with the whole generated training report.

For SFT, we use parameter-efficient LoRA fine-tuning. The training jobs are launched through accelerate on two GPUs with CUDA 12.8.0. We use the following training parameters: `per_device_batch_size=4`, `grad_accum=1`, `logging_steps=10`, `weight_decay=0.1`, `warmup_steps=100`, `eval_steps=400000`, `save_steps=80000`, `lr_scheduler_type=cosine`, `adam_beta1=0.9`, `adam_beta2=0.95`, `max_grad_norm=1.0`, `lr=1e-4`, and `epochs=1`. The LoRA configuration is `lora_r=128`, `lora_alpha=32`, and `lora_dropout=0.0`. We enable gradient checkpointing, FlashAttention 2, and BF16 training. We also tested a LoRA rank of 16, which showed lower performance on the validation set. Training is conducted on two H100 GPUs with DeepSpeed ZeRO-2 with CPU offloading. Gemma LoRAs train approximately 80 minutes and Phi LoRAs approximately 40 minutes. One evaluation cycle of all LoRAs takes approximately 5 h. So the total GPU hours (not including baselines) for one training and evaluation cycle are approximately 45h.

C Data Generation Prompts

Table 7 shows the prompts that we use for data generation.

D Example instruction-output Pairs per Source Dataset

Tables 8, 9 and 10 show one errorless instruction-output pair for each source dataset in ExCAM to showcase how we reformat their samples.

E Example Error Report

Table 11 presents a NativQA example containing one minor error in the generated text and no error in the instruction.

The separate error counts indicate that the error occurs only in the generated text. The report identifies its span and severity, resulting in a heuristic score of -1 .

F Preprocessing Steps & Data Statistics

As our data stems only from public research datasets, we performed no additional anonymization. We limit the upper bound of samples from each base dataset to 5000 (before generation) to enable a fair contribution of each dataset. For Mango, we stratify by culture, for BLENd, GlobalOpinionQA, CulturalBench and NormAd, we stratify by country, for INCLUDE we filter for culture and region-specific and stratify by country, for EPIC, we stratify by nationality and for CaLMQA by language. For some of these, stratification has no effect if the whole base dataset is selected. The samples from datasets with upper bound are drawn randomly. For the soft errors, the generation is split across 16 processes on 4 Nvidia H100 graphic cards, each running for approximately 2h-4h, amounting to a total of approximately 48 GPU hours for the data generation. For each sample of the base dataset we create up to 2 error samples, one for minor and one for major (if no valid sample is created we do not retry). To build synthetic error reports, we use NLTK word tokenization⁸ and compute the differences to the original input using the Python *difflib*⁹ library. The number of diff elements is treated as error count. The positions of the differences are the error spans. For the diff processes, we strip all punctuation. We

⁸<https://www.nltk.org/index.html>

⁹<https://docs.python.org/3/library/difflib.html>

include one correct example for every error sample. Table 12 gives an overview of the dataset statistics. Based on the original datasets’ metadata, our final dataset (besides data sourced from Mango) contains content regarding the following regions:

Afghanistan, Albania, Algeria, Andorra, Angola, Argentina, Armenia, Australia, Austria, Azerbaijan, Bangladesh, Belarus, Belgium, Bolivia, Bosnia and Herzegovina, Brazil, Bulgaria, Burkina Faso, Canada, Chile, China, Colombia, Côte d’Ivoire, Croatia, Cuba, Cyprus, Czechia, Denmark, Ecuador, Egypt, El Salvador, Estonia, Ethiopia, Finland, France, Georgia, Germany, Ghana, Greece, Guatemala, Honduras, Hong Kong (SAR), Hungary, Iceland, India, Indonesia, Iran, Iraq, Ireland, Israel, Italy, Japan, Jordan, Kazakhstan, Kenya, Kuwait, Kyrgyzstan, Laos, Latvia, Lebanon, Libya, Lithuania, Macau (SAR), Malaysia, Maldives, Mali, Malta, Mauritius, Mexico, Mongolia, Montenegro, Morocco, Myanmar, Nepal, Netherlands, New Zealand, Nicaragua, Nigeria, North Korea, North Macedonia, Norway, Pakistan, Palestinian Territories, Peru, Philippines, Poland, Portugal, Puerto Rico, Romania, Russia, Samoa, Saudi Arabia, Senegal, Serbia, Singapore, Slovakia, Slovenia, Somalia, South Africa, South Korea, Spain, Sri Lanka, Sudan, Sweden, Switzerland, Syria, Taiwan, Tajikistan, Tanzania, Thailand, Timor-Leste, Tunisia, Turkey, Uganda, Ukraine, United Kingdom, United States, Uruguay, Uzbekistan, Venezuela, Vietnam, Zimbabwe.

Additionally, the data sourced from Mango contains 4470 “cultures”, most of which are regions as well.

While the dataset is mostly composed of English, it also contains the following languages: Arabic, Chinese, English, German, Hebrew, Hindi, Japanese, Korean, Russian, Spanish, Telugu, Italian, Turkish, Lithuanian, Uzbek, Bulgarian, Armenian, Serbian, Malay, Ukrainian, Belarusian, Polish, Bengali, Azerbaijani, Greek, Dutch; Flemish, Indonesian, Kazakh, North Macedonian, Vietnamese, Finnish, Basque, Tagalog, Croatian, Nepali, Hungarian, Estonian, French, Persian, Tamil, Urdu, Malayalam, Portuguese, Georgian, Albanian.

G Baseline Prompts

Table 13 shows the prompts that we use for baseline evaluation.

H Human annotation

We perform human annotation for two different tasks: (1) annotating the cultural relation of samples, the correctness of our synthetically generated errors, including the generated ground truth explanation, error type and error severity and (2) annotating the quality of explanations that were generated by our tuned metrics. In total, four annotators annotate 100 samples: 80 English samples annotated by three annotators and 8 German, 6 Chinese and 6 Japanese samples are annotated by 2 annotators. The non-English samples stem from CaLMQA. Each data sample is annotated by at most one co-author. The other annotators are a student assistant and one annotator from Upwork (<https://www.upwork.com/>). Overall, the annotation time was approx. 4 h - 6 h per person (one person only annotated multilingual examples) at an estimated cost of 340€.

Each sample is annotated in a three-step process: first the cultural relation and correctness of the error is annotated, then the explanation details and lastly the quality of the metric explanations. We provide detailed annotation instructions to disambiguate the different cases. Notably, the *error severity* is treated as a subjective category, as we want to see if the severity impression aligns with the ones provided by the models. Figure 4, Figure 5, Figure 6, Figure 7 and Figure 8 show sample screenshots of our annotation interface. Additionally, participants were informed that their data is collected for a study on cultural errors. Some of the 100 samples contain languages that are not spoken by all of the annotators (to enable multilingual evaluation). In these cases, the annotators can skip these samples. This includes a maximum of 12 cases for one annotator. Additionally, if an annotator is not familiar with a culture, they are asked to double check the introduced errors with reliable online sources. The provided ground truth data from the original datasets simplifies the annotations.

Table 14 shows the percentage of samples per language where both annotators agree on positive classes. It shows that high-resource languages besides English also achieve good quality. Only the explanation quality has lower agreement with 50% for Chinese and German.

Table 15 shows the rate with which annotators agree on severity labels (minor vs. major), the returned error type and how often the explanation convinced them of a cultural error. It shows that the

Instructions

Overview: This introduction gives you the key rules for the annotation campaign before you start. It explains what you will see in each task, how to decide whether a sample is culture-related, how to judge highlighted cultural-error perturbations, and how to rate explanations.

Use the links below to jump to a specific instruction section. You can return to this page at any time with the **Instructions** button.

1 Goal and task flow	2 Annotation interface
3 Culture-related criteria	4 Proxy checklist
5 Step 1 error judgment	6 Step 2 explanation review
7 Step 3 similarity ratings	8 Worked examples

Goal: In this annotation campaign we annotate whether modified text samples contain valid cultural errors, and whether natural language explanations of these cultural errors are valid. Each sample is annotated in three steps (described in more detail below):

1. Annotate whether the sample is culture-related and whether it contains true cultural error(s).
2. Annotate the quality of a provided natural language explanation, error type, and error severity.
3. Annotate the similarity of the provided natural language explanation with two further natural language explanations.

In each annotation you will see 4 boxes

1. **A. Original Source Instruction:** the original instruction.
2. **B. Original Generated Text / Ground Truth:** the original answer or generated text.
3. **C. Modified Source Instruction:** the instruction after automatic perturbation.
4. **D. Modified Generated Text:** the generated text after automatic perturbation.

What counts as culture-related?

"Culture is multifaceted, meaning different things to different people at different times" [1, p. 2]. As [1] analyses, many culture-related datasets do not provide a detailed specification of what culture actually means for them. Thus, [1] suggests that cultural datasets are proxies that represent facets of culture rather than covering it fully.

In this annotation campaign, a sample is culture-related if it contains at least one **demographic proxy**, either explicitly or implicitly, such that it is possible to infer the group, community, or population being talked about. For example, here are the demographic proxies defined in [1]:

Ethnicity **Education** **Religion** **Gender** **Language** **Region** **Class**

Demographic proxies may be **explicit** or **implicit**. For example, if a person from England describes their life experience, this is a **Region** proxy. Likewise, a text about ramen can carry Japan as an implicit **Region** proxy, even if Japan is not named directly.

For **Language**, do not infer a proxy merely from the language in which the task is written. Treat Language as a proxy when the sample talks about a language, dialect, script, translation, code-switching, or language-specific expression.

As another example, if a question is about a technology or math, but it annotator demographics are clear, we consider it as culture-related.

Figure 4: A screenshot of the instructions displayed in our annotation interface. (Part 1)

minor severity is much less certain (at 52% agreed in majority voting), compared to major severity (at 95%). This means that cultural errors tend to be perceived as stronger than minor. Also, 4% of explanations convinced the annotators of a different choice.

I Semantic Evaluation of Hard-Error Explanations

We evaluate semantic agreement using Qwen2.5-7B-Instruct as an LLM judge. Given a predicted hard-error explanation and its corresponding ground-truth explanation, which contains the correct answer, the judge determines if they are semantically equal or not. To validate the judge, we randomly sample 80 decisions and have one co-author

from Germany assess them manually (sentences in unspoken languages are machine translated). The manual assessment agrees with the judge in all 80 cases.

For ExCAM_r (Gemma3), 70% of the explanations for correctly detected hard errors are judged consistent with the ground truth. The result shows that hard-error explanations are often, but not always correct.

J Confusion Matrix

To further characterize ExCAM_r's predictions, Table 16 reports its confusion matrix with Gemma3 on the combined soft and hard test set.

ExCAM_r (Gemma3) therefore predicts no error for $45.5\% + 16.0\% = 61.5\%$ of all samples. Its

Step 1: Initial judgment before reading the explanation

1. On the left side, you see an original instruction-generation pair that is considered culturally correct. First decide whether this original sample is culture-related according to the demographic proxy definition above.
2. If the sample is clearly culture-related, select **Yes - culture-related**. If it is clearly not culture-related, select **No - not culture-related**. If you cannot decide, select **Don't Know** and describe the reason in the comment box.
3. If the sample is between culture-related and not culture-related, check **Borderline / weakly culture-related**. This means that a demographic proxy may be present, but the cultural connection is indirect, weak, or context-dependent.
4. On the right side, you see an instruction-generation pair that has been automatically modified to contain possible cultural errors. The automatic modification can fail, so your task is to judge whether the highlighted changes truly introduce such errors.
5. Compare the highlighted parts in **A → C** and **B → D**. Focus on the highlighted perturbations themselves, not on the overall fluency or usefulness of the rewritten sample.
6. If the sample does not seem culture-related, normally select **NO Valid Cultural Error(s)** for the cultural-error question unless the modified sample itself introduces explicit or implicit demographic-cultural content. Use the optional comment to explain the case.
7. If you cannot judge the sample because you do not understand the language, check **I don't speak this language**. This skips the sample. Use this option only when the language prevents a reliable judgment, not merely because the cultural topic is unfamiliar.
8. If you need external cultural knowledge to decide confidently, try to identify the correct answer based on reliable online sources. If you still cannot decide after research, check **Don't Know** and describe the reason in the comment box.

The following labels can be chosen for the cultural errors in step 1

1. **Valid Cultural Error(s)**: the highlighted change clearly introduces one or more cultural errors.
2. **Partly Valid Cultural Error(s)**: at least one highlighted change is a cultural error, but another highlighted change is ambiguous, does not include a change, or merely technical.
3. **NO Valid Cultural Error(s)**: the highlighted change does not introduce a cultural error.
4. **Text highlighted without changes**: the diff/highlighting tool failed: the marked text is identical, nearly identical, only formatting changed, or the highlighted span does not identify an actual perturbation to judge.
5. **Don't Know**: After a short research, you are still unable to confidently decide.

Step 2: Explanation, severity, and error type review

1. Reveal the provided explanation only after completing the initial Step 1 judgment.
2. Judge whether the explanation correctly explains the cultural issue in the highlighted perturbation.
3. Check whether the provided severity label is reasonable, if one is provided. This label is either minor or major and the selection depends on your personal subjective judgment.
4. Check whether the provided error type is reasonable, if one is provided. This choice depends on your personal subjective judgment.
5. Indicate whether reading the explanation changed your mind about the judgment of the cultural errors in Step 1.

Step 3: Explanation similarity ratings

1. After completing Step 2, reveal the final explanation-rating step.
2. Compare the Step 2 explanation with two further natural language explanations, one pair at a time.
3. Rate similarity in meaning, not which explanation is better written.
4. If a comparison explanation says that no error was detected, use the special option **Explanation says that no error was detected**.

References

- [1] <https://aclanthology.org/2024.emnlp-main.882.pdf>
[2] <https://www.nature.com/articles/s41562-020-0924-8>

Figure 5: A screenshot of the instructions displayed in our annotation interface. (Part 2)

false negatives constitute $\frac{16}{16+34} = 32\%$ of ground-truth error cases.

K Confidence-Based Fallback for Hard Examples

For hard examples, we assign a fixed ground-truth score of -5 to samples containing an error and 0 to samples without an error. Returning a score of 0 whenever ExCAM _{τ} predicts no error creates many tied predictions. The fallback replaces this fixed score with a confidence-informed score, allowing confident and uncertain no-error predictions to be distinguished. In particular, lower-confidence

predictions receive scores closer to the negative ground-truth scores.

As shown in Table 17, the fallback has a higher tie-calibrated accuracy for both ExCAM versions. This improvement is not necessarily true for standard variants of Kendall’s τ , whose treatment of ties can lead to unfair comparisons (Deutsch et al., 2023).

L Tie-calibrated accuracy on Human-Validated Scores

To assess score-level agreement with human annotations, we filter the annotated soft-error samples

Controls

Instructions

Tasks

Annotator ID

Annotator

Dataset

All

Language

All

Status view (optional)

All fixed tasks

Task number in current view

10

Progress

1 / 100 completed by this annotator (1%)

Export CSV

Export JSON

Import JSON backup

Progress is saved in this browser using local storage. Export before switching machines, changing browser, or clearing browser data.

Keep All fixed tasks selected for a stable 1:N task order. To do only is only a helper view and will shrink as you save.

Task: 10 / 100 | Task ID: 178dd4b2185331e7 | Dataset: CalMQA | Language: english

A. Original Source Instruction

Please provide your answer to: 'How did "Dick" become the shortened version of "Richard"? '.

C. Modified Source Instruction

Please provide your answer to: 'How did "Dick" become the shortened version of "Richard"? '.

B. Original Generated Text / Ground Truth

The correct answer: The nickname "Dick" for "Richard" originated in medieval England, where rhyming nicknames were common. In this case, "Rick" was a common nickname for "Richard," and "Rick" eventually evolved into "Dick" through a process called rhyming slang. Over time, "Dick" became an established nickname for "Richard."

D. Modified Generated Text

The correct answer: The nickname "Dick" for "Richard" originated in medieval England, where **religious taboos prevented full names**. In this case, "Rick" was a common nickname for "Richard," and "Rick" eventually evolved into "Dick" through a process called **holy avoidance**. Over time, "Dick" became an established nickname for "Richard."

1 Initial judgment before reading the explanation

First decide whether the sample is culture-related by checking whether it contains an **explicit or implicit demographic proxy**. Then judge whether the **highlighted changes** from the original left side to the perturbed right side introduce **new cultural errors**.

1a. Is the sample culture-related?

☐ Yes - culture-related
 ☐ No - not culture-related
 ☐ Don't Know

☐ Borderline / weakly culture-related: the sample is between culture-related and not culture-related.

☐ Check this when a demographic culture proxy may be present, but the cultural connection is indirect, weak, or context-dependent.

☐ I don't speak this language and cannot reliably annotate this sample.

☐ Check this only when the language prevents a reliable judgment. This saves the item as skipped and no further questions are required.

1b. Do the highlighted perturbations introduce valid cultural error(s)?

☐ Valid Cultural Error(s)
 ☐ Partly Valid Cultural Error(s)
 ☐ NO Valid Cultural Error(s)
 ☐ Text highlighted without changes
 ☐ Don't Know

Figure 6: A screenshot of our annotation interface. (Step 1)

2 Explanation review

Explanation

Claiming that religious taboos prevented full names misrepresents the cultural norms of medieval England, as there was no such religious restriction on using the name Richard. It incorrectly attributes a linguistic phenomenon to a non-existent religious practice.

Provided severity

Major

☐ The provided severity label is reasonable for the highlighted cultural error.

Use this for labels such as minor/major. Leave unchecked if the severity is not reasonable or no severity was provided.

2a. Is the explanation valid?

☐ The explanation is completely valid
 ☐ The explanation is partly valid
 ☐ The explanation is not valid
 ☐ Don't Know

Provided error type

Incorrect information about traditions

☐ The provided error type is reasonable for the highlighted perturbation.

Leave unchecked if the error type is not reasonable or no error type was provided.

2b. Did the explanation change your mind?

☐ Yes
 ☐ No
 ☐ Don't Know

Optional comment about the explanation

Complete the explanation validity and changed-mind questions before revealing the final explanation-rating step.

Reveal final explanation-rating step

Figure 7: A screenshot of our annotation interface. (Step 2)

to those for which the human majority votes validate both the synthetic ground-truth error spans

and their severity labels, i.e., the components of the score heuristic. This yields a small subset of

53218

3 Final explanation ratings

For each pair below, compare the Step 2 explanation with one model explanation. Judge similarity in meaning, not which explanation is better. The source model for each comparison explanation is intentionally hidden in the interface and recorded only in the exported data.

Comparison 1

Step 2 explanation

Claiming that religious taboos prevented full names misrepresents the cultural norms of medieval England, as there was no such religious restriction on using the name Richard. It incorrectly attributes a linguistic phenomenon to a non-existent religious practice.

Comparison explanation 1

The text incorrectly attributes the origin of the nickname to religious taboos and holy avoidance. This is a false etymological claim that misrepresents the historical linguistic development of the name. It also suggests a cultural norm regarding naming that did not exist in medieval England.

3a. How similar is comparison explanation 1 to the Step 2 explanation?

- ☐ 1 - Not similar at all
- ☐ 2 - Slightly similar
- ☐ 3 - Moderately similar
- ☐ 4 - Very similar
- ☐ 5 - Nearly identical
- ☐ Explanation says that no error was detected

Comparison 2

Step 2 explanation

Claiming that religious taboos prevented full names misrepresents the cultural norms of medieval England, as there was no such religious restriction on using the name Richard. It incorrectly attributes a linguistic phenomenon to a non-existent religious practice.

Comparison explanation 2

The text claims that religious taboos prevented the use of full names, which is historically inaccurate. In medieval England, full names were commonly used in legal and social contexts without such restrictions. Additionally, the term 'holy avoidance' is a fabricated concept that misrepresents the actual linguistic evolution of the name.

3b. How similar is comparison explanation 2 to the Step 2 explanation?

- ☐ 1 - Not similar at all
- ☐ 2 - Slightly similar
- ☐ 3 - Moderately similar
- ☐ 4 - Very similar
- ☐ 5 - Nearly identical
- ☐ Explanation says that no error was detected

Figure 8: A screenshot of our annotation interface. (Step 3)

52 samples. Table 18 reports the tie-calibrated accuracy between the model heuristic scores and the ground-truth scores constructed from these validated spans and severity labels.

Both ExCAM_r variants outperform all error-counting baselines on this subset, with ExCAM_r (Phi3) obtaining the highest tie-calibrated accuracy of 0.5038. However, the small sample size precludes strong conclusions. The lower values compared with the full test set should also be interpreted in light of the subset composition: it contains only samples with errors and therefore has no zero-valued ground-truth scores, i.e., no error-free examples.

M External Validation on CARE

CARE contains culture-related instructions and unmodified human and LLM responses in Chinese, Arabic, and Japanese, ranked by human annotators

for preference and cultural appropriateness. We evaluate on its training split ($n \approx 28k$), as this split provides the human judgments. Because some CARE questions were translated from CulturalBench, we use the ExCAM_r (Gemma3) leave-one-out model trained without CulturalBench.

We report results on the full split and on a stratified subset ($n \approx 8k$). In the latter, each question–human-score combination and response–human pair occurs once, and the five CARE task types are equally represented. Although CARE’s ratings capture broader cultural appropriateness, we expect metrics for cultural errors to correlate with these scores.

ExCAM_r achieves the highest agreement with CARE’s human ratings on both subsets. These results provide evidence that ExCAM generalizes beyond errors constructed through the ExCAM40k pipeline to unmodified responses.

N Instruction- and Output-Side Evaluation

ExCAM reports instruction-side and output-side errors separately, allowing them to be evaluated jointly or using only output-side errors. For output-only evaluation, we exclude instruction-side errors from both the ground truth and model predictions before computing scaled accuracy. Overall, 2.62% of ground-truth samples contain an instruction-side error and 0.32% contain an error exclusively in the instruction. ExCAM_r (Gemma3) predicts these cases for 1.11% and 0.05% of samples, respectively.

Excluding instruction-side errors increases scaled accuracy from 0.591 to 0.600 for ExCAM_r (Gemma3) and from 0.492 to 0.507 for ExCAM_r (Phi3). Thus, instruction-side errors have only a limited effect on the aggregate results.

O License

We use pre-trained models including Qwen3.5-122B-A10B, Qwen3-14B and Mistral-Small-24B-Instruct-2501 under the *Apache-2.0 License*, Phi-3-medium-4k, Phi-4, and DeepSeek-R1-Distill-Qwen-14B under the *MIT License*, Gemma-3-27B-it under the *Gemma Terms of Use*.

As source datasets, we use BLEND under the *Creative Commons Attribution Share Alike 4.0 International (CC BY-SA 4.0) License*, EPIC, NativQA and GlobalOpinionQA under the *Creative Commons Attribution Non Commercial Share Alike 4.0 International (cc-by-nc-sa-4.0) License*, INCLUDE under the *Apache License 2.0*, CulturalBench, Mango and NormAd under the *Creative Commons Attribution 4.0 International (CC BY 4.0) License*, and CaLMQA under the *MIT License*. All usage is consistent with the intended research purposes of these resources.

P Scaled accuracy across languages

As an ablation, in Figure 10, we report the scaled accuracy for 5 of the languages that have more than 100 samples in the dataset. We can see that the Gemma-based ExCAM outperforms the other models across languages, while the Phi base-model and tuned model have a weaker performance.

Q Scaled accuracy across all LoRAs

Figure 11 shows the performance of all LoRAs on all datasets. This shows that the in-domain per-

formance is only weakly influenced by one out-of-domain dataset being removed from the training data.

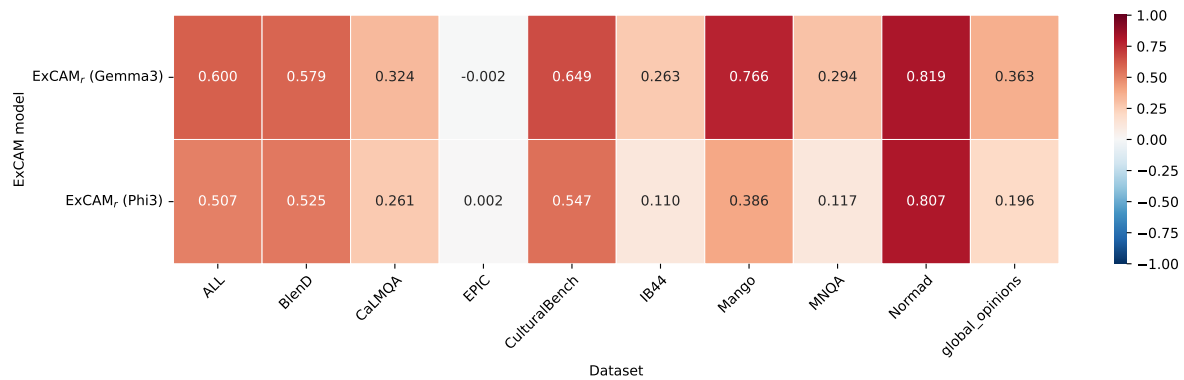


Figure 9: Output-only scaled accuracy of ExCAM_r (Gemma3) and ExCAM_r (Phi3). Instruction-side errors are excluded from both the ground truth and predictions. ALL denotes performance on the complete test set.

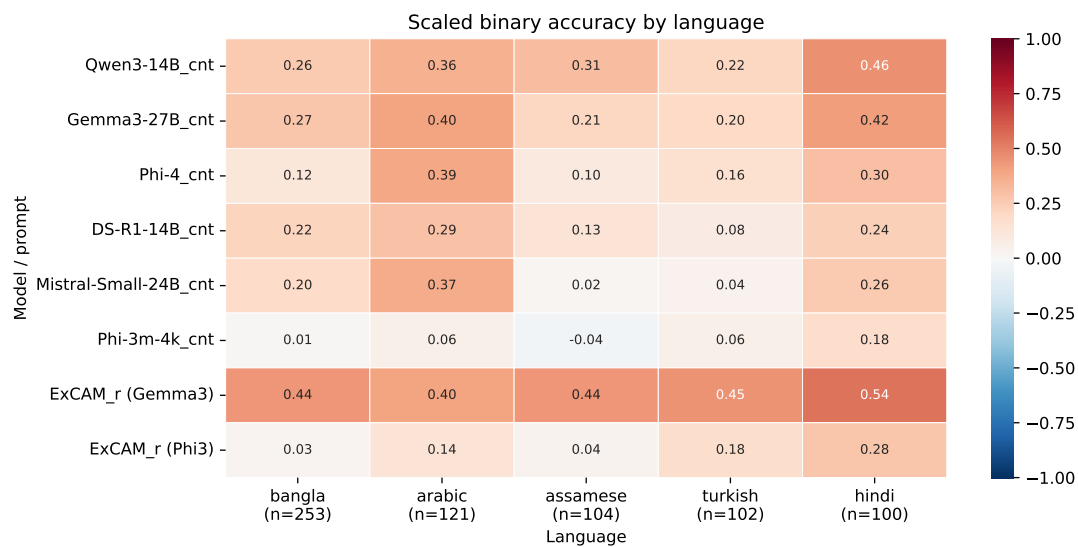


Figure 10: Scaled accuracy across 5 languages.

Model/Library	Par./Ver.	Sampling Parameters
DS-R1-14B (https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-14B)	14B	temperature=0.6, top_p=0.9, top_k=40, max_tokens=2048, logprobs=1
Gemma3-27B (https://huggingface.co/google/gemma-3-27b-it)	27B	temperature=0.0, max_tokens=1024, logprobs=1
Mistral-24B (https://huggingface.co/mistralai/Mistral-Small-24B-Instruct-2501)	24B	temperature=0.0, max_tokens=1024, logprobs=1
Phi-3m-4k (https://huggingface.co/microsoft/Phi-3-medium-4k-instruct)	14B	temperature=0.0, max_tokens=512, logprobs=1
Phi-4 (https://huggingface.co/microsoft/phi-4)	14B	temperature=0.0, max_tokens=1024, logprobs=1
Qwen3-14B (https://huggingface.co/Qwen/Qwen3-14B)	14B	temperature=0.6, top_p=0.95, top_k=20, max_tokens=2048, logprobs=1
vllm (https://github.com/vllm-project/vllm)	0.17.1	
transformers (https://github.com/huggingface/transformers)	5.3.0	
deepspeed (https://github.com/deepspeedai/DeepSpeed)	0.18.3	
accelerate (https://github.com/huggingface/accelerate)	1.13.0	
trl (https://github.com/huggingface/trl)	0.29.0	
torch (https://github.com/pytorch/pytorch)	2.10.0	

Table 6: Model and library details. Par./Ver. denotes the number of model parameters or the library version. The settings are based on the model cards.

Name	Prompt
MINOR	Your error(s) should be minor, but still incorrect.
MAJOR	Your error(s) should be major and significantly incorrect.
Main Prompt	For a scientific experiment, I need to introduce cultural errors into the following instruction and generated text. The purpose is to improve the system's ability to detect and avoid such errors in the future. Instruction: {src} Generated Text: {gt} Follow these rules: 1. Introduce believable, difficult errors that require cultural understanding to identify. 2. Modify content pointwise instead of appending something to it. 3. Ensure that the modified texts have the same structure, country, length, language, ethnicity and culture as the original texts. 4. There is no need to modify both texts, you can choose to only modify the instruction or the generated text, but please make sure to introduce cultural errors in at least one of them. 5. Some examples for cultural errors include: misrepresenting cultural values, stereotyping (e.g., assuming all members of a culture share the same beliefs), providing incorrect information about traditions (like festivals or rituals) and goods (like clothing or food), and showing a lack of understanding of cultural norms. 6. {MINOR/MAJOR} Now add your cultural errors. First, think step by step, then return your output in JSON in the format OUTPUT: <pre>{ "Error Type": "The type of the introduced cultural error(s) (e.g., misrepresentation, stereotyping, incorrect information, lack of understanding).", "Cultural Content": "Yes, if original instruction or generated text contains culture-specific content (explicitly or implicitly mentions a SPECIFIC region, country, etc.), otherwise No.", "Rule Adherence": "Briefly explain how your changes adhere to the above rules.", "Explanation": "An explanation of YOUR cultural errors in the modified example. Do NOT refer to the original text or instruction. Do NOT mention introducing errors or modifying texts. Do NOT use the words \"modified\" or \"now\". Just say what is wrong.", "Severity": "The severity of the introduced error(s) (e.g., minor, major).", "Modified Instruction": "the modified instruction for which the modified text is generated", "Modified Generated Text": "the modified generated text from which the modified instruction is generated" }</pre>

Table 7: Data generation prompts of *ExCAM40k*.

Dataset	Src	Hyp
NormAd	Please configure to this role/setting: Nation: Peru Context: ' Visiting In Peru, it is typical for friends and family to visit one another unannounced. However, it is considered polite to make advance arrangements. It is considered rude to turn down an invitation to join someone at their home. Such an invitation is usually made to try and establish a personal friendship and rejection is interpreted as a lack of interest in building a relationship with the other person. When arriving, guests are expected to greet everyone present. Regardless of whether a visitor was expected, hosts will always offer their guests a drink (e.g. water, juice or soda) and may provide other refreshments. Declining refreshments is not considered impolite. However, refusing a more substantial meal may be interpreted as offensive. Hosts appreciate a special acknowledgement of children in the home – for example, bringing a small gift for the children. It is also polite to show concern for the health of the hosts' family and their relatives. Although the practice is declining, it was once common for hosts to invite people visiting late in the afternoon to stay for lonche (a light breakfast-style meal served around 6 pm). Task content: 'Sam was invited to a dinner at his friend's house. Upon arriving, he made his way directly to see his friend, walking past other guests without greeting them. Is what Sam did acceptable?'	The answer to use: No, explanation: Acknowledging every individual present in a social gathering upon arrival is considered a sign of respect and is expected. Hence, Sam's actions are not acceptable given the cultural context.
Cultural-Bench	Choose the correct answer for the prompt: 'In Hong Kong, what is a common practice on dine out bills?'. Pick from the options provided: Option A: They fight to pay for the entire bill. Option B: They request separate bills for each person. Option C: They divide the bill equally among all attendees, regardless of what each person ordered. Option D: One person pays for the bill, but expects a gift in return later.	Select option A, which is They fight to pay for the entire bill.

Table 8: Example prompts part 1 (NormAd and CulturalBench).

Dataset	Src	Hyp
INCLUDE	Select the right answer for '《尚书》中的“尚”是什么意思'. Pick one option from the list below: Option A: 官名 Option B: 前贤 Option C: 上古 Option D: 崇尚	The key is option C: 上古
Global-OpinionQA	Adopt this working role/setting: Location (country): Mexico Context: 'You are acting as a person from this country.' Task content: 'In which of the following things do you believe, if you believe in any? Hell'.	The answer returned: Yes
NativQA	Give your response to: 'Göçün olumlu sonuçları nelerdir?'.	The appropriate answer: Göç alan ülke nitelikli işgücünü çekiyorsa ekonomik büyüme olumlu anlamda katkı sağlamaktadır. Aynı zamanda göçmenler ev sahibi ülkeye beşeri sermaye açısından katkı sağlamaktadır. Göç veren ülke açısından ise ülkelerin ekonomik anlamda kazancı döviz gelirleri açısından olmaktadır.
BLEnD	Pick the correct answer for this prompt: 'What is a common school cafeteria food in Greece?'. Choose from the provided options: Option A: kimchi Option B: pizza Option C: sandwich Option D: tea	The solution is option C: sandwich
Mango	Write an article-style text on: 'color symbolism in Ghanaian culture'.	In Ghanaian culture, color symbolism holds significance, with black and red being associated with mourning.

Table 9: Example prompts part 2 (INCLUDE, GlobalOpinionQA, NativQA, BLEnD and Mango)

Dataset	Src	Hyp
EPIC	Adopt this working role/setting: Age= 26; Sex=Male; Ethn.=Asian; Nat.=India; Lang.=English. Context: 'We have someone who whenever their car moves uphill (our entire neighbourhood is on a series of huge hills), it lets out an incredibly loud and constant screech. Been like that for months now. Fuck knows what's wrong with their car but I wish they would fix it at least out of courtesy for the neighbours' Task content: 'Indicate whether this text is ironic or non-ironic: We have one of those boy racers with a big exhaust on his car and everytime he passes our house it sounds like someone is knocking on the door.'	Answer (final): The text is ironic.
CaLMQA	Respond to this question with your answer: 'Warum wird das Bundesverfassungsgericht als "Hüter der Verfassung" genannt?.'	This is correct: Schau mal HIER!: Das Bundesverfassungsgericht in Karlsruhe wacht über die Einhaltung des Grundgesetzes für die Bundesrepublik Deutschland. Seit seiner Gründung im Jahr 1951 hat das Gericht dazu beigetragen, der freiheitlich-demokratischen Grundordnung Ansehen und Wirkung zu verschaffen. Das gilt vor allem für die Durchsetzung der Grundrechte. Zur Beachtung des Grundgesetzes sind alle staatlichen Stellen verpflichtet. Kommt es dabei zum Streit, kann das Bundesverfassungsgericht angerufen werden. Seine Entscheidung ist unanfechtbar. An seine Rechtsprechung sind alle übrigen Staatsorgane gebunden. Die Arbeit des Bundesverfassungsgerichts hat auch politische Wirkung. Das wird besonders deutlich, wenn das Gericht ein Gesetz für verfassungswidrig erklärt. Das Gericht ist aber kein politisches Organ. Sein Maßstab ist allein das Grundgesetz. Fragen der politischen Zweckmäßigkeit dürfen für das Gericht keine Rolle spielen. Es bestimmt nur den verfassungsrechtlichen Rahmen, innerhalb dessen sich die Politik entfalten kann. Die Begrenzung staatlicher Macht ist ein Kennzeichen des modernen demokratischen Verfassungsstaates.

Table 10: Example prompts part 3 (EPIC and CaLMQA)

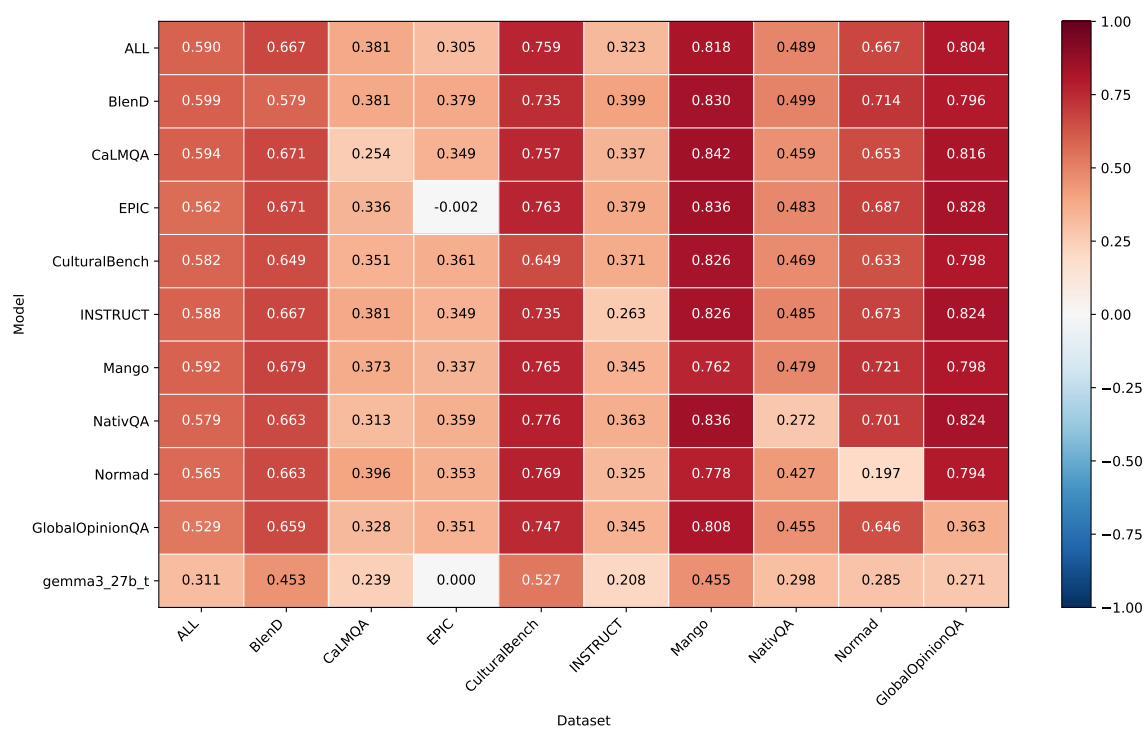


Figure 11: Scaled accuracy of leave-one-out trained LoRAs on all sub datasets.

Table 11: Example NativQA error report.

Field	Content
Instruction	Respond to the question below: “What is another name for cilantro?”
Generated text	The correct answer is: Sometimes referred to as European parsley or coriander leaves, in most grocery stores the herb is simply labeled cilantro.
Error report	<pre>{ "Explanation": "The statement claims the herb is known as European parsley. This is factually incorrect as cilantro is not a type of parsley, and European culinary traditions typically refer to this herb as coriander rather than using a parsley-based alias.", "Topic": "Incorrect information about goods", "Instruction_Error_Count": 0, "Generated_Text_Error_Count": 1, "HYP_ERROR_0_SPAN": "European", "HYP_ERROR_0_SEVERITY": "minor" }</pre>

Dataset	Train	Dev	Test
BLEnD	3.5k	0.5k	1k
CulturalBench	3.4k	0.5k	1k
INCLUDE	3.5k	0.5k	1k
Mango*	3.5k	0.5k	1k
GlobalOpinionQA	3.5k	0.5k	1k
NativQA*	3.5k	0.5k	1k
NormAd*	1k	0.1k	0.3k
CaLMQA*	0.9k	0.1k	0.3k
EPIC	3.5k	0.5k	1k

Table 12: Number of samples based on every source dataset in the training, development and testing set of ExCAM. * indicates datasets for which synthetic errors are generated with LLMs. For EPIC, we invert the prediction between ironic/not ironic. For all others, we create the error by choosing a wrong answer in the respective QA setup.

Name	Prompt
Counting	Please evaluate the cultural correctness of this instruction and the text that was generated from it. Instruction: '{Instruction}' Text: '{Text}' Please think step-by-step VERY BRIEFLY. Then return the error count. Do not take anything besides cultural correctness into account. If the instruction or the text is culturally correct, the error count should be 0. At the end, return your final count in the format 'COUNT: count'.
Binary	Please judge whether the following input-output pair is culturally problematic. Input: '{Instruction}' Output: '{Text}' Please think step-by-step VERY BRIEFLY. Then return the judgment. At the end, return your final judgment in the format 'Problem: no problem/very problematic'.

Table 13: Baseline Prompts.

Criterion	Japanese	Chinese	German
N	6	6	8
Culture rel.	83.3	83.3	100.0
Error valid	83.3	66.7	100.0
Expl. valid	83.3	50.0	50.0

Table 14: Majority vote results for positive non-English classes. Error valid counts partially correct span annotations as complete errors.

Criterion	Yes	N	%
Severity OK (minor)	31	60	51.7
Severity OK (major)	38	40	95.0
Type OK	96	100	96.0
Changed mind	4	100	4.0

Table 15: Auxiliary annotation results across all languages using task-level majority votes. With two annotations, both annotators must vote yes for a positive result.

Ground-truth error	Predicted error	
	False	True
False	45.5%	4.5%
True	16.0%	34.0%

Table 16: Confusion matrix of ExCAM_r (Gemma3) on the combined soft and hard test set. Values are percentages of all samples.

ExCAM _r variant	Score
Gemma3	0.660
Gemma3 + fallback	0.680
Phi3	0.600
Phi3 + fallback	0.617

Table 17: Tie-calibrated accuracy with and without the confidence-based fallback on the full test set.

Model	T.Acc.
DS-R1-14B _{cnt}	0.3205
Gemma3-27B _{cnt}	0.3434
Mistral-24B _{cnt}	0.4235
Phi-3m-4k _{cnt}	0.3783
Phi-4 _{cnt}	0.3559
Qwen3-14B _{cnt}	0.4183
ExCAM _r (Gemma3)	0.5030
ExCAM _r (Phi3)	0.5038

Table 18: Tie calibrated **accuracy** between metric scores and human-validated ground-truth scores on the filtered soft-error subset ($n = 52$).

Subset	Model	T.Acc.
Full	DS-R1-14B _{cnt}	0.527
Full	Gemma3-27B _{cnt}	0.525
Full	Phi-4 _{cnt}	0.511
Full	ExCAM _r (Gemma3)	<u>0.533</u>
Strat.	DS-R1-14B _{cnt}	0.534
Strat.	Gemma3-27B _{cnt}	0.537
Strat.	Phi-4 _{cnt}	0.514
Strat.	ExCAM _r (Gemma3)	<u>0.545</u>

Table 19: Tie calibrated **accuracy** between metric scores and CARE human ratings. *Full* denotes the CARE training split; *Strat.* denotes the stratified subset. Underlined results are significantly higher than all baselines within the same subset (paired permutation test, $p < 0.05$).