CUS-QA: Local-Knowledge-Oriented Open-Ended Question Answering Dataset

Jindřich Libovický[1]* **Jindřich Helcl**[2]* **Andrei-Alexandru Manea**[1] **Gianluca Vico**[1]

[1]Faculty of Mathematics and Physics, Charles University, Czech Republic
[2]Language Technology Group, University of Oslo, Norway
libovicky@ufal.mff.cuni.cz jindrich@ifi.uio.no

Abstract

We introduce CUS-QA, a benchmark for evaluation of open-ended regional question answering that encompasses both textual and visual modalities. We also provide strong baselines using state-of-the-art large language models (LLMs). Our dataset consists of manually curated questions and answers grounded in Wikipedia, created by native speakers from Czechia, Slovakia, and Ukraine, with accompanying English translations. It includes both purely textual questions and those requiring visual understanding. We evaluate state-of-the-art LLMs through prompting and add human judgments of answer correctness. Using these human evaluations, we analyze the reliability of existing automatic evaluation metrics. Our baseline results show that even the best open-weight LLMs achieve only over 40% accuracy on textual questions and below 30% on visual questions. LLM-based evaluation metrics show strong correlation with human judgment, while traditional string-overlap metrics perform surprisingly well due to the prevalence of named entities in answers.

1 Introduction

The ability to answer factual questions is one of the most frequently used and evaluated capabilities of current large language models (LLMs), both purely textual and multimodal. However, existing evaluation approaches have significant limitations that fail to capture realistic usage scenarios.

Most widely used QA benchmarks focus on globally known facts and use multiple-choice formats for easy evaluation and cross-lingual comparison (Hendrycks et al., 2021; Clark et al., 2020; Rajpurkar et al., 2016). This setup is unrealistic in two important ways. First, when people use LLMs to get answers to their questions, they ask open-ended questions without providing multiple answer choices. Second, the questions that matter to users vary significantly across regions and cultures. A question about local geography, politics, or cultural figures that is highly relevant to speakers of one language may be completely irrelevant to speakers of another language.

Regional knowledge presents a particularly challenging test case for LLMs. Unlike globally known facts, regional information requires models to have learned culturally specific knowledge during training, often from less-represented languages and sources. Moreover, question-answering performance of LLMs varies wildly depending on the language of the question (Singh et al., 2025). This makes region-specific QA an ideal probe for understanding the knowledge gaps and cultural biases in current models.

Open-ended question answering poses an additional evaluation challenge beyond the regional knowledge gap. Unlike the multiple-choice setup, open-ended responses cannot be evaluated through exact matching, as there are typically multiple valid ways to express the correct answer. Furthermore, evaluation of open-ended QA systems remains an underexplored area, with not enough publicly available resources that can be used to validate the evaluation metrics.

We introduce CUS-QA (Czech-Ukrainian-Slovak Question Answering), a dataset that addresses both limitations simultaneously. We collect questions about regional knowledge from native speakers of Czech, Slovak, and Ukrainian, three Slavic languages with varying resource levels and speaker populations. Our questions are grounded in local Wikipedias and cover both textual and visual modalities and require models to show knowledge about local geography, culture, history, and politics that is well-known within each country but largely unknown outside it.

*Equal contribution

Transactions of the Association for Computational Linguistics, vol. 14, pp. 1482–1509, 2026. https://doi.org/10.1162/TACL.a.737
Action Editor: Daisuke Kawahara. Submission batch: 8/2025; Revision batch: 1/2026; Published 7/2026.

Image	Question	Answer
	Kdo nechal postavit známý český hrad na obrázku?	Karel IV.
	Who built the famous Czech castle in the picture?	*Charles IV.*

Table 1: Visual sample of collected Czech questions and answers. Image: *Hrad Karlštejn, pohled od jihu*, Lukáš Kalista, CC BY-SA 3.0.

	# Annotators	Hours paid	Collected counts		After filtering	
			Textual	Visual	Textual	Visual
CZ	9	89	1,242	596	1,080	456
SK	8	120	1,203	329	972	238
UA	9	110	889	740	755	403

Table 2: Basic descriptive stats of the data collection process.

To establish baselines and understand the challenges posed by our dataset, we evaluate several state-of-the-art open-weight LLMs and complement automatic evaluation with human judgment. Our results reveal significant gaps in regional knowledge: even the best models achieve only over 40% accuracy on textual questions and below 30% on visual questions. We also find substantial cross-lingual inconsistencies, where models sometimes perform better at answering regional questions in English rather than the local language (§ 5.3).

Beyond establishing performance baselines, we conduct a human evaluation to analyze the reliability of automatic evaluation metrics for open-ended QA. This addresses the critical need in the field for human-labeled datasets to develop metrics for reference-based evaluation of open-ended generation.

The paper is structured as follows: Section 2 describes our data collection process, followed by a detailed dataset description in Section 3. Section 4 presents our evaluation methodology and baseline approaches. Baseline results, correlation analysis between automatic metrics and human judgment, cross-lingual differences, analysis of robustness to prompt variations, and results of retrieval-augmented generation experiments are described in Section 5. We discuss implications and future work in Section 6, give an overview of related work on multilingual QA and evaluation metrics in Section 7, and conclude in Section 8.

	Number of seed Wikipedia pages				% of seed	
	Single-language	Match location	Total	% of Wiki	Text	Visual
CZ	111k	12k	123k	21.4	31.9	20.6
SK	36k	4k	40k	15.7	32.6	15.3
UA	187k	10k	197k	14.2	25.2	18.1

Table 3: Number of seed Wikipedia pages per language in the pool. Single-language are pages that exist only in the local language, Match location are pages about entities located in the respective country, and % of Wikipedia shows percentage of the local Wikipedia the pool represents. "% of seed" is the percentage of dataset items created from seed articles per modality.

2 Data Collection

In this section, we describe our data collection methodology that relies on native speakers to identify regional-specific facts that are well-known within specific countries (Czechia, Slovakia, and Ukraine) but largely unknown outside these regions. The basic statistics of the data we collected are shown in Table 2.

The annotators were asked to find facts specific to their country, i.e., facts that are commonly known locally but largely unknown outside the country. To avoid overly obscure facts while maintaining regional specificity, we provided annotators with a guideline to consider whether the fact would be known by a substantial portion of the population (we suggested thinking of at least 10,000 people as a rough threshold). Annotators ultimately relied on their judgment of what constitutes regionally common knowledge.

We developed a web interface for collecting the data. The annotators were presented with a Wikipedia page randomly selected from a predefined pool. The tool allows annotators to browse Wikipedia content, including clicking links navigating different pages or skipping the current page, redirecting the annotator to a random page from the seed pool. The article pool consisted of Wikipedia pages about (1) entities that are located in the respective country in DBPedia (Mendes et al., 2012), and (2) pages that only existed in the given language. Table 3 shows the number of Wikipedia pages per language that were available in the pool. Each annotator could only create a single annotation for each Wikipedia page. The code for the annotation interface is available on GitHub.[1]

[1] https://github.com/ufal/regional-qa-annotation

		Number of instances		Avg. # chars				Ans. is a full sent.?
				Local		English		
		Dev	Test	Q	A	Q	A	
Textual	CZ	530	550	65	19	72	22	4%
	SK	493	479	57	19	62	19	7%
	UA	385	370	65	32	69	34	14%
Visual	CZ	226	230	31	18	38	21	2%
	SK	118	120	38	20	43	22	9%
	UA	204	199	40	34	45	36	18%

Table 4: Basic descriptive statistics of the dataset: dev and test split size, the average length of questions and answers in the number of characters, and proportion of answers that are a full sentence.

If a suitable fact was found on a page, the annotators wrote a question and an answer in their language (Czech, Slovak, Ukrainian) and provided an English translation of both. If the Wikipedia page had a photograph, the annotators were encouraged to also write down a question and an answer about the photograph (including the English translations).

All annotators are native speakers of the respective languages with sufficient knowledge of English (have language certificate and/or passed a compulsory university English exam), who were raised in the countries where these languages are spoken. However, all annotators are currently university students living in the Czech Republic. The annotators were compensated at a rate of 300 CZK (≈ 12€) per hour, which is 2.4× the minimum wage and 1.2× the median wage in the Czech Republic as of 2024.

We manually verified *all* questions and answers to ensure they adhered to the annotation guidelines. Questions were rejected if they: (1) concerned globally known facts, (2) were too obscure, (3) contained factual errors, or (4) had unclear phrasing. During this process, we corrected 8% of the items (mostly typos and grammatical issues, in both languages) and discarded 22% of the questions. All filtering decisions were made by research team members with domain expertise. Our manual inspection did not find any major issues in the quality of the English translations provided by the annotators.

3 Dataset Description

Table 4 lists characteristics of the dataset. For each country and question type, the dataset is divided into a development and test split. Each example in the dataset consists of the following, manually annotated features:

- the Wikipedia page URL,
- a question in the local language,
- a answer in the local language,
- the same question translated into English,
- the corresponding answer in English, and
- the image data (visual modality only).

We show examples of the collected data in the visual modality part in Table 7, with additional examples from both modalities in Tables 19 and 20 in

		Textual			Visual		
		CZ	SK	UA	CZ	SK	UA
Categories	Geography	39	42	32	60	50	39
	Culture	27	23	19	21	24	20
	Politics	5	9	12	3	9	9
	History	21	12	26	10	6	20
	Sports	5	8	5	2	3	5
	Other	3	6	4	4	8	5
Basic NER	Location	94	87	85	76	60	57
	Person	46	44	50	32	39	43
	Organization	33	41	45	11	19	31
	Date/Year	8	7	11	1	0	5
Fine-grained named entities	River	9	8	7	4	3	1
	Mountain	14	17	6	10	8	6
	City/Town	56	48	65	43	40	42
	Village	11	13	5	5	9	2
	Movie	3	1	0	0	0	0
	TV show	5	7	3	0	0	1
	Book	7	3	3	2	0	1
	Song	1	2	3	0	0	1
	Actor	6	3	3	1	1	1
	Writer	9	6	8	3	3	5
	Musician	3	6	4	2	10	1
	Other artist	5	6	8	2	5	4
	Politician	13	13	24	3	11	14
	Sportsperson	2	5	2	1	5	2
	Other person	32	24	32	29	27	33
	Sports event	5	9	5	0	1	3
	Political party	3	5	5	0	0	1
	Other org.	21	27	36	6	12	22
	Product/Brand	11	8	13	8	2	14
	Company	4	2	5	0	1	3

Table 5: Percentage of questions belonging to six mutually exclusive main categories (upper part) and a proportion of question-answer pairs containing named entities of the listed types with different granularity (middle and lower parts).

Location	content								has text	medium		
	portrait	building	city	nature	statue	food	technology	other		color photo	BW photo	painting
CZ	23	57	30	8	8	2	1	4	18	86	10	4
SK	31	52	15	5	4	1	0	6	25	92	7	2
UA	25	36	27	5	4	3	1	21	23	79	10	10

Table 6: Percentage of images in the dataset belonging to the listed categories based on the visual content.

the appendix. We release the dataset on Hugging Face Hub.[2]

The average length of questions in the textual modality part is around 60 characters in all languages; the English translations are slightly longer. The questions in the visual modality part are shorter because some context is already provided by the image. In both modalities, the answers are, on average, 20 characters long in Czech and Slovak but over 30 in Ukrainian. Most of the answers are short noun phrases. In Czech, only a small portion of the answers are full sentences; it is slightly more in Slovak, and almost one-sixth of the answers in Ukrainian.

Besides the manually created features, we augment each example with automatic translations into the other two local languages, using Claude 3.5 Sonnet (best scoring system in WMT 2024; Kocmi et al., 2024), which we later use for cross-lingual comparison of the models (§ 5.3). To verify the translation quality, we randomly sampled 100 question-answer pairs in three translation directions and compared Google Translation, GPT-4o, and Claude 3.5 Sonet (Table 8). Because the questions and answers are quite simple texts, the translation was perfect in the vast majority of cases, with only minor issues in the rest.

For a quantitative summary of the dataset content, we use LLaMA 3.3 70B (Dubey et al., 2024) to assign one of the following categories to each example: Geography, Culture, Politics, History, Sports, and Other. We treat these categories as mutually exclusive, even though sometimes, the classification may be ambiguous (e.g., History vs. Politics). The category statistics are shown in Table 5. In both textual and visual questions, geography was the most frequent category, followed by culture and history.

[2] https://huggingface.co/datasets/ufal/cus-qa

Additionally, we use GliNER (Zaratiana et al., 2024) on the English translation of the question-answer pairs and we include the extracted statistics in Table 5. GliNER allows defining custom entities based on prompts, so besides the basic entities (location, person, organization, date), we defined our own set of fine-grained entities that better describe the content of the dataset. The overview of the named entities shows that the vast majority of question-answer pairs contain a location, with city names being its most frequent subtype. Roughly 40% of the questions contain the name of a person. According to GliNER, the most frequent subcategory of names belongs to politicians. However, this might be inaccurate, as we apply an English-language recognizer to non-English names.

Similarly for the visual part of the dataset, we used LLaMA 3.2 11B Vision to categorize the images based on whether they primarily capture faces of people (portraits), a building, a town or city, natural scenery, a statue, food, or technological artifacts (such as cars and tools). Photos that do not contain any of these are categorized as other. Further, we detect if there is text in the photo. Finally, we classify whether it is a color photo, black-and-white photo or a different medium (in most cases a drawing or a painting). The statistics are in Table 6.

CUS-QA is designed for evaluation purposes only. The dataset size (1,536 Czech, 1,210 Slovak, and 1,158 Ukrainian question-answer pairs) is intentionally limited to serve as a test set for assessing model capabilities on regional knowledge rather than as training data.

4 Baselines & Evaluation

We evaluate several state-of-the-art open-weight models on CUS-QA to establish baseline performance and examine cross-lingual consistency. Our experiments test how well current LLMs perform on regional questions across both textual and visual modalities, and whether models maintain consistent performance when questions are asked in different languages.

We use human evaluation to analyze the reliability of automatic metrics for open-ended QA. We compare various evaluation approaches with human judgments to determine which metrics best capture answer quality across different languages and modalities.

Image	Question	Answer
	Jak se jmenuje zpěvák na obrázku? *What is the name of the singer in the picture?*	Karel Kryl *Karel Kryl*
	Ktorá slovenská jaskyňa je na obrázku? *Which Slovak cave is in the picture?*	Demänovská ľadová jaskyňa *Demänovská Ice Cave*
	Хто зображений на фотографії? *Who is shown in the photo?*	На фотографії зображена постать Гетьмана Павла Скоропадського. *The photo shows the figure of Hetman Pavel Skoropadskyi.*

Table 7: Examples of collected Czech, Slovak and Ukrainian questions and answers. More examples from both modalities in Tables 19 and 20 in the Appendix. Images: *Karel Kryl na koncertě v Sušici v roce 1990*, Josef Karas, CC BY-SA 4.0. *Demanova Ice Cave 23*, Jojo, CC BY-SA 3.0. *Pavlo Skoropadsky portrait, colorized by Ruslan Habanets*, Pavlo Shtelmakh, CC BY-SA 3.0.

Translator	sk-cs	uk-cs	cs-sk	uk-sk	cs-uk	sk-uk
Google Translate	78	65	57	75	73	76
GPT-4o	98	86	94	82	88	87
Claude 3.5 Sonet	98	88	97	84	90	90

Table 8: Comparison of Google Translation, GPT-4o and Claude 3.5 Sonet. Sentences were evaluated by a single annotator, indicating whether the translation can be considered perfect. Low performance of Google Translation is due to pivoting via English, which leads to incorrect transliteration and English names in the translations.

4.1 Baselines

As baselines, we test zero-shot generation performance of several widely used open-weight LLMs in each modality setting. In the text-only setup, we evaluate the following models: LLaMA 3.1 8B, LLaMA 3.3 70B (Dubey et al., 2024), Mistral v0.3 7B (Jiang et al., 2023) and EuroLLM 7B (Martins et al., 2024). For visual QA, we test mBLIP (Geigle et al., 2023), LLaMA 3.2 and 4 Vision, Maya (Alam et al., 2024), Gemma 3 (Team, 2025) and idefics (Laurençon et al., 2023).

Based on preliminary experiments, we decided to use an English system prompt and instruct the model to answer in the language of the questions. See Appendix A for the exact prompt formulation. For decoding, we used nucleus sampling with a nucleus of 0.9 in all setups, and we only sampled one answer per question and model.

4.2 Human Evaluation

We manually evaluate the model outputs using the following binary criteria: correctness, truthfulness, relevance, and coherence. Our aim was to design a small set of error categories that best describe what we observed in the generated answers.

Correctness. In the main evaluation criterion, the annotators decided if the generated answer is correct, given the reference answer, using their subjective judgment. If the answer contained minor hallucinations or misleading information that do not substantially change the overall meaning of the answer, it is still considered correct.

Truthfulness. The annotators verified whether the entire answer is objectively true. This is separate from whether the answer addresses the question. An answer can be factually true but incorrect if it does not answer the question. We also allow saying that the answer is correct if it contains irrelevant, untrue statements. If the model refuses to answer, it is annotated as not true.

Relevance. Here, the annotators judged whether the answer contains only the information the question asks for and nothing else, is appropriately specific, and is not too general.

Coherence. A coherent answer is grammatically correct and in the correct language. This is judged regardless of the factuality or relevance.

These categories allow us to define different levels of strictness in how we judge the answers: In the permissive case, we only consider the correctness criterion, and in the strict case, we want all criteria to be fulfilled. Each answer was independently evaluated by two annotators. We consider a criterion to be fulfilled when both annotators agree.

The annotators in the evaluation campaign received the same compensation as in the data collection phase (§ 2). Some of the annotators for data collection were also involved in human evaluation. All of the annotators for this phase are native speakers of the relevant language. The total time spent on manual evaluation was 106 hours.

Inter-annotator agreement (Table 9), measured by Cohen's κ, shows generally strong consistency between annotators across most evaluation criteria. The agreement is highest for correctness and truthfulness (average $\kappa = 0.84$ and 0.86 respectively). It is lower for relevance ($\kappa = 0.70$) and coherence ($\kappa = 0.80$), likely reflecting the more subjective nature of these judgments.

4.3 Automatic Metrics

We evaluate the generated answers using the following metrics: BLEU (Papineni et al., 2002), chrF (Popović, 2015), ROUGE-L (Lin, 2004), BERTScore (Zhang et al., 2020) with XLM-R (Conneau et al., 2020) as an underlying model, BLEURT (Sellam et al., 2020).

Additionally, we use LLM as a judge (Zheng et al., 2023). We use LLaMA 3.3 70B, a general-purpose model, and M-Prometheus that was fine-tuned to be used as a judge in multilingual scenarios. The judge prompt contains the question, the reference answer, the model output, and the question of whether the answer is correct or not (see Appendix A for the full prompt). The LLM-based judges produce a binary value (correct/incorrect) and we report the ratio of correct answers in the dataset.

4.4 Retrieval Augmented Generation

Because the dataset design ensures that all answers are sourced from the local Wikipedia, we conduct additional experiments on the textual part of the dataset using retrieval-augmented generation (RAG).

			Correct	True	Relevant	Coherent
Text	CZ	cs	.91	.90	.84	.76
		en	.90	.86	.72	.84
	SK	sk	.90	.88	.72	.69
		en	.90	.85	.61	.85
	UA	uk	.84	.83	.79	.88
		en	.78	.79	.58	.91
Visual	CZ	cs	.65	.83	.55	.64
		en	.58	.85	.47	.81
	SK	sk	.90	.91	.83	.66
		en	.91	.91	.88	.83
	UA	uk	.91	.83	.75	.83
		en	.90	.82	.65	.93
Average κ			.84	.86	.70	.80

Table 9: Inter-annotator agreement for all evaluation criteria for two annotators measured by Cohen's κ.

We used the FineWiki dataset (Penedo, 2025) to build the RAG index for each local Wikipedia. We split each document into overlapping chunks, use the Multilingual-E5-Large text embeddings (Wang et al., 2024) to get the vector representations of the chunks and store the vectors in the index. In all experiments, we set the chunk size to 512 tokens, with 10% overlap between neighboring chunks of the same document. We use the Faiss library for building the index and retrieval (Douze et al., 2026).

During decoding, we embed the question using the same model, and retrieve the top 3 most similar vectors and the respective chunks from the index, and include the retrieved data in the prompt (see Appendix A for the exact prompt formulation).

5 Results

We present the performance of the baseline models on open-ended regional question answering on CUS-QA (§ 5.1), analyze the correlation between automatic metrics and human judgments (§ 5.2), and examine cross-lingual consistency in model capabilities (§ 5.3). We include assessment of the robustness of the results to prompt variations (§ 5.4), and proof-of-concept experiments with retrieval-augmented generation (§ 5.5).

5.1 Baseline Results

We report the strongest baseline results in Table 10, including results of manual evaluation and best-correlating automatic metrics (chrF, LLM as a judge) on the development and test data in both modalities. Table 11 shows LLM-as-a-judge evalu-

			Human dev		Auto. dev		Auto. test	
			OK	Perf.	chrF	LLM	chrF	LLM
Textual	CZ	cs	58.7	53.4	42.4	57.4	43.7	60.0
		en	52.5	44.7	37.8	56.0	37.5	63.1
	SK	sk	46.2	40.8	39.2	45.4	35.2	47.7
		en	45.8	40.4	32.5	47.5	32.6	48.3
	UA	uk	51.9	46.2	42.4	57.4	42.6	52.2
		en	47.8	43.1	37.8	56.0	37.0	54.9
Visual	CZ	cs	14.6	11.1	21.3	14.6	21.3	14.6
		en	13.3	8.4	15.5	15.9	15.5	15.9
	SK	sk	23.7	20.3	24.4	18.6	24.4	18.6
		en	19.5	12.7	16.6	16.9	20.4	17.5
	UA	uk	29.9	29.9	21.1	26.0	20.7	15.1
		en	17.2	17.2	19.1	20.1	21.7	23.6

Table 10: Results of the strongest baselines for textual (LLaMA 3.3 70B Instruct) and visual (LLaMA 4 Scout 17B 16E Instruct) QA tasks. Human evaluation includes factual correctness (OK) and full correctness that also includes truthfulness, relevance, and coherence (Perfect). Additionally, we report the chrF score and LLaMA 3.3 70B as a judge for development and test data. Complete results of all tested models are in the Appendix in Tables 21 and 22.

Model	CZ		SK		UA	
	cs	en	sk	en	uk	en
LLaMA 3.3 70B Ins.	60.0	63.1	47.7	48.3	52.2	54.9
LLaMA 3.1 8B Ins.	29.6	36.5	19.6	25.6	25.7	34.1
EuroLLM 9B Ins.	39.3	43.1	30.8	30.2	26.8	28.4
Mistral 7B Ins. v0.3	22.9	30.4	14.4	22.3	24.6	27.8
LLaMA 4 Scout Ins.	41.8	45.6	34.6	34.4	45.7	45.4
Gemini 2.5 Flash	78.0	75.8	64.2	61.3	68.9	63.0
Gemini 2.5 Fl.-Lite	60.0	54.2	47.9	43.3	54.3	51.9
GPT-5-2025-08-07	83.1	80.7	73.3	70.2	73.2	72.2
GPT-5-mini-2025-08-07	67.5	70.7	52.3	50.4	61.4	59.7
LLaMA 3.2 11B Ins.	12.6	14.3	11.7	15.0	11.6	12.1
LLaMA 4 Scout Ins.	18.7	18.3	15.0	17.5	15.1	23.6
maya	1.7	3.0	1.7	2.5	2.5	4.0
idefics	4.8	7.4	6.7	9.2	6.0	8.5
gemma3	15.2	13.9	11.7	13.3	16.6	12.1
Gemini 2.5 Flash	39.6	37.0	38.3	34.2	41.2	40.7
Gemini 2.5 Fl.-Lite	30.9	27.4	26.7	24.2	27.1	27.6
GPT-5-2025-08-07	46.5	42.6	44.2	41.7	37.2	36.2
GPT-5-mini-2025-08-07	33.0	38.7	24.2	26.7	26.1	31.7

Table 11: Test data results for all models (including commercial ones) evaluated by LLaMA 3.3 70B as a judge. Textual question are in the upper part of the table, visual questions are in the lower part. Commercial models are in italics.

ation for all tested models, including recent commercial systems.

Textual QA. The best-scoring model for textual QA was LLaMA 3.3, which outperformed the other models by a large margin, often 15–20 percentage points in accuracy compared to the others, and performed consistently well across languages and regions. There is a gap between focusing on correctness only, versus considering all evaluation criteria. This shows that models often add irrelevant or hallucinated information to their answers. Bigger commercial models (GPT-5 and Gemini 2.5 Flash) outperform the best open-weight model by a large margin. Faster versions have a similar score to the best open-weight model.

Results of all textual models, including all metrics and human evaluation breakdown, are presented in Table 21 in the Appendix. All models show differences in performance across languages. Mistral performs much worse in Slovak than in other languages. LLaMA 3.1 also struggles with Slovak, and like Mistral, it answers questions about Slovakia better in English than in Slovak. EuroLLM, which focuses on European languages, outperforms both models of similar size (Mistral and LLaMA 3.1) in all settings except asking about Ukraine in English, where LLaMA 3.1 is better.

Breakdown per question category (Table 23 in the Appendix) shows that models score the highest on geography, followed by history and culture.

Visual QA. As shown in Table 10, the results of the visual subtask are much lower. Whereas in textual QA, the accuracy of the best-scoring model consistently exceeds 40% in the strictest setting, with visual QA, it does not exceed 30% in any of the languages.

Results of all vision models are presented in Table 22. By a large margin, the best-scoring model is LLaMA 4 Scout, followed by Gemma 3 and LLaMA 3.2. Idefics and Maya score very poorly, with Maya achieving near-zero performance across all languages and conditions. The visual QA results show even greater regional variation than textual QA. For instance, Llama 4 Scout performs notably better for Ukraine compared to Czechia or Slovakia. Commercial models are by a large margin better than open-weight models. Note, however, that whilst in the textual scenario, the performance of LLaMA 3.3 was comparable to the lightweight closed-source models (Gemini 2.5 Flash-Lite, GPT-5-mini), in case of visual QA, we did not have access to a open-weight vision-language model of comparable size. The question category breakdown

	Textual							Visual						
	BLEU	chrF	RG	BScr	BLRT	LLM	M-P	BLEU	chrF	RG	BScr	BLRT	LLM	M-P
Correct	.733	.867	.833	.817	.883	.950	950	.667	.550	.761	-.267	.817	1.000	.962
+ true	.753	.879	.845	.829	.879	.970	970	.667	.550	.761	-.267	.817	1.000	.962
+ relevant	.783	.917	.883	.867	.917	.967	967	.667	.550	.761	-.267	.817	1.000	.962
+ coherent	.783	.917	.883	.867	.917	.967	967	.629	.695	.773	-.215	.829	.920	.858

Table 12: System-level Spearman correlations of the automatic evaluation metrics (BLEU, chrF, ROUGE-L, BERTScore, BLEURT, LLaMA 3.3 70B as a judge, M-Prometheus as a judge) with increasing levels of strictness of human evaluation. The presented numbers are averages of Spearman correlations computed separately for each country and language. The breakout for specific languages is in Table 26 in the Appendix. Pearson correlations are in Table 25 in the Appendix.

	Textual					Visual				
	chrF	RG	BLRT	LLM	M-P	chrF	RG	BLRT	LLM	M-P
Correct	.594	.477	.552	.790	.707	.409	.370	.450	.807	.699
+ true	.585	.472	.543	.775	.703	.408	.370	.449	.804	.696
+ relevant	.567	.457	.524	.743	.683	.410	.370	.452	.797	.715
+ coherent	.548	.448	.511	.715	.665	.402	.376	.423	.710	.654

Table 13: Answer-level point-wise biserial correlation of human judgment with automatic metrics, break-out per language is in Table 27 in the Appendix.

(Table 23 in the Appendix) does not show consistent patterns across countries and languages.

When we consider the visual content of the images (see Table 24), we observe a difference in questions about images with faces in Czechia and Slovakia on the one side, and Ukraine on the other. Whereas for Czechia and Slovakia, the models often refuse to answer (even though people in the images are public figures), in Ukrainian, images with faces are the best-scoring image category. Questions about images with cities score relatively well across languages and locations, followed by pictures of buildings. The presence of text increases the likelihood that the answer will be correct, even in cases where the text is not part of the answer but still provides some clues. Questions with color photos receive, on average, better answers than questions about black-and-white images. However, the confounding factor might be that questions about black-and-white images ask about older facts.

5.2 Metrics and Human Judgment

We examine the correlation between automatic evaluation metrics and human judgments by comparing system-level (Table 12) and answer-level (Table 13) performance.

System-level correlation. For textual QA, we find very high correlations between most metrics and human judgment. Traditional string-overlap metrics like BLEU, chrF, and ROUGE-L achieve correlations above 0.85 with human correctness judgments. This strong performance likely stems from the high proportion of named entities in the answers, where exact string matches are more meaningful than in other generation tasks like machine translation or summarization. LLM-based evaluation metrics show even stronger correlations, with LLaMA 3.3 70B achieving near-perfect system-level correlation ($r > 0.95$) across most conditions. M-Prometheus, specifically designed for multilingual evaluation, performs similarly well. Interestingly, BERTScore shows lower correlations than simpler string-overlap metrics.

This changes considerably for visual QA, where correlations drop substantially across all metrics except for LLM as a judge. String-overlap metrics maintain reasonable performance (correlations around 0.6–0.8), but BERTScore shows negative correlations in some cases.

Answer-level correlation. At the answer level, correlations are generally lower but follow similar patterns. LLM-based metrics consistently outperform other metrics, achieving correlations around 0.6–0.8 with human judgments for both modalities. The difference between system-level and answer-level correlations shows that, while automatic metrics can reliably rank systems on average, they are less effective at judging individual answer quality.

Legend		Average		Textual	CZ cs		CZ en		SK sk		SK en		UA uk		UA en		Visual	CZ cs		CZ en		SK sk		SK en		UA uk		UA en	
TP	FN	29	4		34	3	30	5	26	3	25	6	32	3	28	7		7	1	7	1	9	4	9	2	13	4	14	1
FP	TN	5	61		2	61	8	57	2	69	7	62	3	63	9	56		2	90	3	89	1	86	1	88	2	82	4	81

Table 14: Averaged confusion matrices comparing using LLaMA 3.3 70B Instruct as judge and human evaluation of correctness. The per-model breakout is in Table 28 in the appendix.

		Czechia				Slovakia				Ukraine			
		cs	sk	uk	en	cs	sk	uk	en	cs	sk	uk	en
Textual	EuroLLM-9B-Instruct	**35.5**	29.8	8.3	31.7	29.0	30.2	9.7	**32.9**	17.4	15.8	26.8	**29.1**
	Llama-3.1-8B-Instruct	**28.5**	17.0	8.9	31.3	21.5	18.3	7.9	**25.2**	18.7	15.1	23.4	**34.8**
	Llama-3.3-70B-Instruct	**57.5**	52.3	30.0	56.2	43.6	43.8	24.9	**47.1**	40.8	39.0	**54.5**	54.3
	Llama-4-Scout-17B-16E-Instruct	**38.3**	33.2	21.5	38.1	28.6	32.9	18.7	**33.5**	31.7	34.8	**47.3**	46.8
	Mistral-7B-Instruct-v0.3	22.6	11.9	11.7	**26.8**	17.2	13.6	8.9	**22.7**	17.4	12.2	21.8	**32.5**
Visual	gemma3	10.6	7.5	5.8	**12.8**	12.7	**13.6**	7.6	11.9	13.7	12.7	**22.1**	21.1
	idefics	**7.5**	3.5	4.4	6.6	6.8	7.6	8.5	**10.2**	7.8	4.9	8.3	**10.3**
	Llama-3.2-11B-Vision-Instruct	10.2	7.5	8.4	**13.7**	**11.9**	9.3	11.0	11.0	11.8	9.3	15.2	**19.1**
	Llama-4-Scout-17B-16E-Instruct	15.0	11.1	4.9	**15.9**	16.1	**18.6**	7.6	16.1	16.2	14.2	25.0	**29.9**
	maya	**1.8**	0.0	0.4	**1.8**	0.0	0.0	0.0	0.0	2.9	1.5	2.9	**5.9**

Table 15: Cross-lingual comparison of models: We report the accuracy by LLaMA 3.3 70B across translations of the questions and answers into other languages from the dev subset (with the local language highlighted). Note that the accuracies are only comparable within the country-specific blocks.

LLM as a judge. We further analyzed the LLaMA 3.3 70B as a judge. Table 14 shows the confusion matrices. The number of false positives and false negatives is relatively balanced. When exploring the data, we noticed that questions that get misclassified by the judge repeat across models, both as false positive and false negatives. However, we were not able to find a pattern explaining what they have in common E.g., in Czech, singer Karel Gott, football club Baník Ostrava or the Estates Theatre in Prague systematically confused the judge.

Our findings reveal that simple string-overlap metrics perform surprisingly well for textual QA, likely because named entities are more prevalent. However, LLM-based evaluation seems to be the most reliable approach across all conditions, particularly for the more challenging visual QA task.

5.3 Cross-Lingual Differences

The results of cross-lingual comparison (Table 15) show substantial performance differences depending on the language of the question.

For textual QA, models tend to perform well when answering questions in the local language. However, the performance gaps vary across models and countries. For questions about Czechia, the largest model (LLaMA 3.3 70B) maintains stable performance for Czech, Slovak, and English. Smaller models exhibit greater variability, often performing better in English than in the translated languages. When questions about Slovakia are translated into Czech, the performance remains roughly the same, whilst when translated into Ukrainian, the performance deteriorates. Finally, answering in English yields better results than answering in Slovak itself. For questions about Ukraine, the cross-lingual performance pattern differs: models answer questions about Czechia and Slovakia with comparable accuracy in Czech and Slovak, but performance decreases when these questions are asked in Ukrainian. Likewise, when answering questions about Ukraine, models perform poorly if the questions are translated into Czech or Slovak.

Visual QA shows even bigger variations. However, relatively low accuracy scores might affect the reliability of the statistics. Whereas Czech is generally the best-performing language for questions about Czechia, it is better to query the models in English for questions about Slovakia and Ukraine. Surprisingly, for questions about Slovakia, the LLaMA 3 models worked better in Czech than in Slovak.

5.4 Prompt Variation

To verify that the presented results are not an artifact of the prompts we selected, we conduct additional experiments to assess the dataset's robust-

Model	CZ		SK		UA	
	cs	en	sk	en	uk	en
EuroLLM 9B Ins.	1.6	2.4	0.7	0.7	1.9	1.5
Llama 3.1 8B Ins.	1.9	2.3	1.3	3.2	1.4	2.2
Llama 3.3 70B Ins.	1.0	1.6	1.4	1.8	1.0	0.6
Llama 4 Scout Ins.	1.0	1.3	1.2	0.7	1.5	2.7
Mistral 7B Ins. v0.3	1.3	0.4	0.7	0.7	1.0	0.7

Table 16: Standard deviation of the LLM-as-a-judge score when running the models with different prompts.

Model	CZ		SK		UA	
	cs	en	sk	en	uk	en
EuroLLM 9B Ins.	37.7	33.0	68.5	60.2	27.3	19.2
Llama 3.1 8B Ins.	52.3	47.0	69.8	47.0	44.2	33.7
Llama 3.3 70B Ins.	65.9	59.8	76.9	73.4	58.7	51.2
Mistral 7B Ins. v0.3	50.4	47.2	71.2	63.3	46.0	33.5

Table 17: Accuracy of retrieval-augmented generation using local Wikipedia evaluated with LLaMA 3.3 70B as a judge on the development set.

ness to prompt variations. For each location, we tested additional 4 prompts in the local language and 4 in English, varying in wording and style (see Appendix A).

Table 16 shows standard deviations of scores given by LLM as a judge, evaluated on the development set. In most cases, the deviation is around one percentage point, which is smaller than the differences between models. The responses to more colloquial prompts differ in style, but rarely in the factual content. The average chrF and LLM-as-a-judge scores are in Table 29 in the Appendix.

5.5 Retrieval Augmented Generation

The results are presented in Table 17. As expected, having access to the retrieved context leads to superior performance in terms of LLM as a judge score. LLaMA 3.3 70B is still the best performing model. The results of the Mistral model are especially noteworthy, as its performance increased dramatically in the local language settings. EuroLLM seem not to benefit from the additional context, which suggests weaker long context handling abilities.

All models perform better with RAG in all setups, except when asking about Ukraine in English. Upon manual inspection of the outputs, we noticed that the low scores in English questions about Ukraine are caused by the models often answering in Ukrainian, despite the instructions. The biggest scores were achieved on Slovak questions. This is likely caused by smaller size of the Slovak Wikipedia and shorter article lengths, which makes the retrieval more accurate and the prompts shorter.

6 Discussion and Future Work

Our experiments show that current evaluation metrics work reasonably well for open-ended regional QA. This finding suggests that the focus on multiple-choice QA in many existing datasets may be unnecessarily restrictive. The high proportion of named entities in factual answers makes string-overlap metrics more effective than in other generation tasks.

The dataset can serve multiple research purposes beyond basic QA evaluation. Since all information was available on Wikipedia as of late 2024, researchers can use it to evaluate RAG systems or cross-lingual RAG approaches. This provides a controlled setting where the knowledge source is known and accessible.

The question of what knowledge should be built into models versus retrieved from external sources remains open. Regional knowledge presents an interesting test case for this trade-off. While one could argue that such specific information should be left to external knowledge sources, we believe that broader factual knowledge generally improves a model's ability to search for and contextualize additional information. Models that perform better on our dataset are likely to provide more culturally appropriate responses across different regions.

CUS-QA is designed as an evaluation benchmark to measure model performance on regional knowledge. The dataset supports assessment across both textual QA and visual QA, with visual QA proving significantly more difficult. The dataset also enables studies of cross-lingual consistency that allow researchers to examine how well models transfer knowledge across related languages.

To make the benchmark more accessible, we host it on Codabench (Xu et al., 2022), where users can test their models. Participants can download the full dev set and the test set without the reference answers. The platform will then score the prediction on the full test set, with results broken by language and modality. In this way, participants can score partial submissions and focus only on a subset of modalities and languages. On Codabench, we provide all the metrics with the exception of LLM as a judge with M-Prometheus due to hardware limitations.

Our human evaluation provides a resource for developing and validating automatic metrics for open-ended QA. Since human-labeled datasets for this task are rare, our annotations offer a benchmark for future research on reference-based evaluation and metric reliability.

Finally, the dataset can serve as a seed for automatically generating additional training data from Wikipedia or other resources. Researchers can extend our approach to create larger-scale datasets covering more regions and languages, or use similar collection methods for other knowledge domains.

7 Related Work

We review existing datasets for multilingual question answering (§ 7.1), visual question answering (§ 7.2), dataset targeting our languages of interest (§ 7.3), and evaluation metrics (§ 7.4).

7.1 Textual Question Answering

With the rise of decoder-based generative models, multiple-choice QA has become a common evaluation method. The most frequently used English MMLU dataset (Hendrycks et al., 2021) covers 57 diverse topics but is English-only and assumes US-centric knowledge in some areas.

Several datasets extend QA evaluation to other languages with regional content. Etxaniz et al. (2024) introduce parallel trivia questions in English and Basque, with a subset on Basque culture. KazQAD (Yeshpanov et al., 2024) combines Kazakh local knowledge with general subjects from school exams. Kostiuk et al. (2025) compile a Lithuanian history dataset, showing that prompting in Lithuanian improves performance. Etori et al. (2025) localize MMLU into Latvian and Giriama.

Larger-scale multilingual efforts include INCLUDE (Romanou et al., 2025), which collects multiple-choice questions from local exam sources in 44 languages, and Global MMLU (Singh et al., 2025), which covers 42 languages with both culture-agnostic and culture-specific questions. Food-focused datasets use cuisine as a cultural proxy: Zhou et al. (2025) presents questions on local ingredients, while Lavrouk et al. (2025) covers dishes from former Soviet states, revealing that LLMs often confuse post-Soviet nations.

Cross-lingual performance gaps have been documented in several studies. Rohera et al. (2025) find that LLMs often perform better in English even for questions from Indic contexts. Hupkes and Bogoychev (2025) show more nuanced results with MultiLoKo, a 31-language benchmark with 500 locally-sourced questions per language. Goldman et al. (2025) present ECLeKTic, which uses Wikipedia article presence across 12 languages to identify language-specific facts and evaluate cross-lingual knowledge transfer, showing that current models struggle to share knowledge across languages even when they can answer in the source language.

7.2 Visual Question Answering

Initial formulations of Visual Question Answering (VQA; Antol et al., 2015; Hudson and Manning, 2019) assumed a closed output vocabulary to evaluate object understanding, an approach carried into multilingual extensions (Pfeiffer et al., 2022).

ALM-Bench (Vayani et al., 2025) is most similar to our work, covering cultural elements from 73 countries in both local languages and English, and including several question types: short and long open-ended, binary, and multiple-choice. CVQA (Romero et al., 2024) covers 30 countries, but it uses a multiple-choice format and does not overlap with our target regions. Nayak et al. (2024) introduce an English-only benchmark for open-ended VQA on traditions, food, and clothing. Wang et al. (2025) propose a multiple-choice VQA set on tourism. Other multimodal benchmarks incorporate exam questions across languages (Zhang et al., 2023; Das et al., 2024), but focus on visual elements, such as graph comprehension, rather than cultural knowledge.

7.3 QA in Czech, Slovak, and Ukrainian

Table 18 summarizes existing datasets covering our target languages.

MultiLoKo (Hupkes and Bogoychev, 2025) uses a similar Wikipedia-based methodology but covers only Czech with 500 questions. INCLUDE (Romanou et al., 2025) provides school exam questions in all three languages, but with a limited size and inconsistent domains. Global MMLU (Singh et al., 2025) does not include Slovak and relies mostly on unverified machine translation for Czech and Ukrainian.

For single-language resources, BenCzechMark (Fajcik et al., 2025) aggregates Czech datasets, including school exams with some regional content. The ZNO dataset (Paniv et al., 2025) comprises 3.7k Ukrainian national exam questions covering also regional history and literature.

Dataset name	Citation	Languages			Modality		Regional questions	Human eval.	Size in cs+sk+uk
		cs	sk	uk	Text	Vis.			
MultiLoKo	Hupkes and Bogoychev (2025)	✓	✗	✗	✓	✗	✓	✗	500
INCLUDE	Romanou et al. (2025)	✓	✓	✓	✓	✗	✗	✗	50+31+182
Global MLLU	Singh et al. (2025)	✓	✗	✓	✓	✗	✗	✗	14k
BenCzechMark	Fajcik et al. (2025)	✓	✗	✗	✓	✗	◌	✗	
ZNO	Paniv et al. (2025)	✗	✗	✓	✓	✗	◌	✗	3.7k
ALM-bench	Vayani et al. (2025)	✓	✓	✓	✓	✓	◌	✗	269+128+269
WorldCuisines	Winata et al. (2025)	✓	✗	✗	✓	✓	✗	✗	1.5k
CUS-QA		✓	✓	✓	✓	✓	✓	✓	1536+1210+1158

Table 18: Comparison of CUS-QA with other existing QA datasets for Czech, Slovak and Ukrainian. The symbol ◌ in the Regional question column indicate that the dataset contains some regional questions, but it was not the focus of the dataset.

Two datasets include visual QA: ALM-bench (Vayani et al., 2025) covers all three languages with culturally-focused questions that were machine-translated and post-edited. WorldCuisines (Winata et al., 2025) includes Czech but focuses narrowly on cuisine with minimal coverage of our target regions.

CUS-QA provides broader coverage across both modalities, an explicit regional focus, and human evaluation annotations.

7.4 Evaluation Metrics

Traditional metrics measure textual overlap using word or character n-grams (e.g., BLEU: Papineni et al., 2002; ROUGE: Lin, 2004; chrF: Popović, 2015). More recent approaches leverage embedding similarity (e.g., BERTScore: Zhang et al., 2020, BLEURT: Sellam et al., 2020) or fully trained metrics like COMET (Rei et al., 2020) that are targeted at machine translation.

Recently, LLMs have often been used as judges, either directly assessing the quality (Zheng et al., 2023) or as part of more structured pipelines, such as FactScore (Min et al., 2023), which first breaks the generated text into atomic facts and evaluates them individually.

The standard approach to validating evaluation metrics is to measure their correlation with human judgments, a practice best established in machine translation (e.g., Freitag et al., 2024). However, human-labeled datasets for other tasks are limited, with exceptions like image captioning (Flickr8k-Expert Hodosh et al., 2013; Nebula Matsuda et al., 2024; Polaris Wada et al., 2024), text summarization (ROSE: Liu et al., 2023), and MOCHA for reading comprehension (Chen et al., 2020), mostly in English.

8 Conclusions

We introduce CUS-QA, a regional knowledge dataset covering textual and visual question answering in Czech, Slovak, and Ukrainian. The dataset contains manually curated questions and answers from native speakers, grounded in Wikipedia and focused on local knowledge that is well-known within each country but largely unknown outside it.

Our baseline experiments reveal significant gaps in the regional knowledge capabilities of the current LLMs. The best textual QA model achieves above 40% accuracy, while visual QA is much more challenging with an accuracy below 30%. We observe substantial cross-lingual inconsistencies, where models sometimes perform better answering regional questions in English rather than the local language.

Through human evaluation, we find that LLM-based metrics correlate well with human judgment for this task, while traditional string-overlap metrics perform surprisingly well due to the high number of named entities in the dataset, especially in the textual QA part. These findings suggest that open-ended evaluation of factual QA might be more feasible than previously thought.

The dataset addresses two key research needs: evaluating regional knowledge in LLMs, including cross-lingual generation consistency, and validating automatic evaluation metrics for open-ended QA. As an evaluation benchmark, CUS-QA enables researchers to measure model capabilities on culturally specific knowledge without training on the test distribution. We release both the dataset and human evaluation results to support future research in these areas.

Limitations

Our dataset has several limitations. First, the dataset size is relatively small compared to major benchmarks. Second, all annotators were university students, which does not reflect the demography of the respective countries. Third, our focus is narrowed on three closely related Slavic languages.

The dataset is designed for evaluation only. With approximately 1,200–1,500 examples per language, it is too small for fine-tuning but appropriately sized for assessing model performance on regional knowledge.

The dataset reflects knowledge as of late 2024. Our reliance on Wikipedia as the grounding source introduces potential biases toward formal, institutional knowledge. Some overlap exists between Czech and Slovak questions due to shared history, potentially inflating cross-lingual performance comparisons.

The human evaluation was limited to five LLMs, which may not capture the full range of model outputs and error patterns that could emerge in future models. The binary evaluation criteria achieved good inter-annotator agreement but may not capture subtle quality differences between generated responses

Our evaluation focused solely on factual accuracy and excluded other important factors like cultural sensitivity, tone, and political awareness that could affect real-world performance.

Acknowledgments

We would like to thank Michal Novák and Tymur Kotkov for helping with the Slovak and Ukrainian prompts and for helping with the annotation coordination.

This research was supported by the Charles University project PRIMUS/23/SCI/023 and project CZ.02.01.01/00/23_020/0008518 of the Ministry of Education, Youth and Sports of the Czech Republic, and by the European Union Digital Europe project OpenEuroLLM, number 101195233.

References

Nahid Alam, Karthik Reddy Kanjula, Surya Guthikonda, Timothy Chung, Bala Krishna S. Vegesna, Abhipsha Das, Anthony Susevski, Ryan Sze-Yin Chan, S. M. Iftekhar Uddin, Shayekh Bin Islam, Roshan Santhosh, Snegha A, Drishti Sharma, Chen Liu, Isha Chaturvedi, Genta Indra Winata, Ashvanth. S, Snehanshu Mukherjee, and Alham Fikri Aji. 2024. Maya: An Instruction Finetuned Multilingual Multimodal Model. *CoRR*, abs/2412.07112.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *2015 IEEE International Conference on Computer Vision, ICCV 2015*, pages 2425–2433, Santiago, Chile. IEEE Computer Society.

Anthony Chen, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2020. MOCHA: A dataset for training and evaluating generative reading comprehension metrics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6521–6532, Online. Association for Computational Linguistics.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Rocktim Das, Simeon Hristov, Haonan Li, Dimitar Dimitrov, Ivan Koychev, and Preslav Nakov. 2024. EXAMS-V: A multi-discipline multilingual multimodal exam benchmark for evaluating vision language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7768–7791, Bangkok, Thailand. Association for Computational Linguistics.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hos-

seini, and Hervé Jégou. 2026. The Faiss Library. *IEEE Trans. Big Data*, 12(2):346–361.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, et al. 2024. The Llama 3 Herd of Models. *CoRR*, abs/2407.21783.

Naome A. Etori, Arturs Kanepajs, Kevin Lu, and Randu Karisa. 2025. LAG-MMLU: benchmarking frontier LLM understanding in Latvian and Giriama. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 109–120, Tallinn, Estonia. University of Tartu Library.

Julen Etxaniz, Gorka Azkune, Aitor Soroa, Oier Lopez de Lacalle, and Mikel Artetxe. 2024. BertaQA: How Much Do Language Models Know About Local Culture? In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC*, Canada.

Martin Fajcik, Martin Docekal, Jan Dolezal, Karel Ondrej, Karel Beneš, Jan Kapsa, Pavel Smrz, Alexander Polok, Michal Hradis, Zuzana Neverilova, Ales Horak, Radoslav Sabol, Michal Stefanik, Adam Jirkovsky, David Adamczyk, Petr Hyner, Jan Hula, and Hynek Kydlicek. 2025. BenCzechMark: A Czech-centric multitask and multimetric benchmark for large language models with duel scoring mechanism. *Transactions of the Association for Computational Linguistics*, 13:1068–1095.

Markus Freitag, Nitika Mathur, Daniel Deutsch, Chi-Kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, David Ifeoluwa Adelani, Marianna Buchicchio, Chrysoula Zerva, and Alon Lavie. 2024. Are LLMs breaking MT metrics? results of the WMT24 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 47–81, Miami, Florida, USA. Association for Computational Linguistics.

Gregor Geigle, Abhay Jain, Radu Timofte, and Goran Glavas. 2023. mBLIP: Efficient Bootstrapping of Multilingual Vision-LLMs. *CoRR*, abs/2307.06930.

Omer Goldman, Uri Shaham, Dan Malkin, Sivan Eiger, Avinatan Hassidim, Yossi Matias, Joshua Maynez, Adi Mayrav Gilady, Jason Riesa, Shruti Rijhwani, Laura Rimell, Idan Szpektor, Reut Tsarfaty, and Matan Eyal. 2025. ECLeKTic: a Novel Challenge Set for Evaluation of Cross-lingual Knowledge Transfer. *CoRR*, abs/2502.21228.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. In *9th International Conference on Learning Representations, ICLR 2021*, Virtual Event, Austria. OpenReview.net.

Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing Image Description as a

Ranking Task: Data, Models and Evaluation Metrics. *J. Artif. Intell. Res.*, 47:853–899.

Drew A. Hudson and Christopher D. Manning. 2019. GQA: A New Dataset for Real-world Visual Reasoning and Compositional Question Answering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019*, pages 6700–6709, Long Beach, CA, USA. Computer Vision Foundation / IEEE.

Dieuwke Hupkes and Nikolay Bogoychev. 2025. MultiLoKo: a multilingual local knowledge benchmark for LLMs spanning 31 languages. *CoRR*, abs/2504.10356.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *CoRR*, abs/2310.06825.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, Mariya Shmatova, Steinthór Steingrímsson, and Vilém Zouhar. 2024. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA. Association for Computational Linguistics.

Yevhen Kostiuk, Oxana Vitman, Lukasz Gagala, and Artur Kiulian. 2025. Towards Multilingual LLM Evaluation for Baltic and Nordic languages: A study on Lithuanian History. *CoRR*, abs/2501.09154.

Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M. Rush, Douwe Kiela, Matthieu Cord, and Victor Sanh. 2023. OBELICS: An Open Web-scale Filtered Dataset of Interleaved Image-text Documents. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023*, New Orleans, LA, USA.

Anton Lavrouk, Tarek Naous, Alan Ritter, and Wei Xu. 2025. What are foundation models cooking in the post-soviet world? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20687–20709, Suzhou, China. Association for Computational Linguistics.

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.

Yixin Liu, Alex Fabbri, Pengfei Liu, Yilun Zhao, Linyong Nan, Ruilin Han, Simeng Han, Shafiq Joty, Chien-Sheng Wu, Caiming Xiong, and Dragomir Radev. 2023. Revisiting the gold standard: Grounding summarization evaluation with robust human evaluation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4140–4170, Toronto, Canada. Association for Computational Linguistics.

Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno Miguel Guerreiro, Ricardo Rei, Duarte M. Alves, José Pombal, M. Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, Pierre Colombo, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. 2024. EuroLLM: Multilingual Language Models for Europe. *CoRR*, abs/2409.16235.

Kazuki Matsuda, Yuiga Wada, and Komei Sugiura. 2024. Deneb: A Hallucination-robust Automatic Evaluation Metric for Image Captioning. In *Computer Vision - ACCV 2024 - 17th Asian Conference on Computer Vision*, Lecture Notes in Computer Science, pages 166–182, Hanoi, Vietnam. Springer.

Pablo Mendes, Max Jakob, and Christian Bizer. 2012. DBpedia: A multilingual cross-domain knowledge base. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1813–1817, Istanbul, Turkey. European Language Resources Association (ELRA).

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

Shravan Nayak, Kanishk Jain, Rabiul Awal, Siva Reddy, Sjoerd Van Steenkiste, Lisa Anne Hendricks, Karolina Stanczak, and Aishwarya Agrawal. 2024. Benchmarking vision language models for cultural understanding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5769–5790, Miami, Florida, USA. Association for Computational Linguistics.

Yurii Paniv, Artur Kiulian, Dmytro Chaplynskyi, Mykola Khandoga, Anton Polishko, Tetiana Bas, and Guillermo Gabrielli. 2025. Benchmarking Multimodal Models for Ukrainian Language Understanding Across Academic and Cultural Domains. In *Proceedings of the Fourth Ukrainian Natural Language Processing Workshop (UNLP 2025)*, pages 14–26, Vienna, Austria (online). Association for Computational Linguistics.

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.

Guilherme Penedo. 2025. Finewiki. Source: Wikimedia Enterprise Snapshot API (https://api.enterprise.wikimedia.com/v2/snapshots). Text licensed under CC BY-SA 4.0 with attribution to Wikipedia contributors.

Jonas Pfeiffer, Gregor Geigle, Aishwarya Kamath, Jan-Martin O. Steitz, Stefan Roth, Ivan Vulić, and Iryna Gurevych. 2022. xGQA: Cross-lingual visual question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2497–2511, Dublin, Ireland. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. 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.

Pritika Rohera, Chaitrali Ginimav, Gayatri Sawant, and Raviraj Joshi. 2025. Better to ask in english? evaluating factual accuracy of multilingual llms in english and low-resource languages. *CoRR*, abs/2504.20022.

Angelika Romanou, Negar Foroutan, Anna Sotnikova, Zeming Chen, Sree Harsha Nelaturu, Shivalika Singh, Rishabh Maheshwary, Micol Altomare, Mohamed A. Haggag, Imanol Schlag, Marzieh Fadaee, Sara Hooker, Antoine Bosselut, Snegha A, Alfonso Amayuelas, Azril Hafizi Amirudin, Viraat Aryabumi, Danylo Boiko, Michael Chang, Jenny Chim, Gal Cohen, Aditya Kumar Dalmia, Abraham Diress, Sharad Duwal, Daniil Dzenhaliou, Daniel Fernando Erazo Florez, Fabian Farestam, Joseph Marvin Imperial, Shayekh Bin Islam, Perttu Isotalo, Maral Jabbarishiviari, Börje F. Karlsson, Eldar Khalilov, Christopher Klamm, Fajri Koto, Dominik Krzeminski, Gabriel Adriano de Melo, Syrielle Montariol, Yiyang Nan, Joel Niklaus, Jekaterina Novikova, Johan Samir Obando-Ceron, Debjit Paul, Esther Ploeger, Jebish Purbey, Swati Rajwal, Selvan Sunitha Ravi, Sara Rydell, Roshan Santhosh, Drishti Sharma, Marjana Prifti Skenduli, Arshia Soltani Moakhar, Bardia Soltani Moakhar, Ran Tamir, Ayush Kumar Tarun, Azmine Toushik Wasi, Thenuka Ovin Weerasinghe, Serhan Yilmaz, and Mike Zhang. 2025. INCLUDE: Evaluating Multilingual Language Understanding with Regional Knowledge.

In *The Thirteenth International Conference on Learning Representations, ICLR 2025*, Singapore. OpenReview.net.

David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, Bontu Fufa Balcha, Chenxi Whitehouse, Christian Salamea, Dan John Velasco, David Ifeoluwa Adelani, David Le Meur, Emilio Villa-Cueva, Fajri Koto, Fauzan Farooqui, Frederico Belcavello, Ganzorig Batnasan, Gisela Vallejo, Grainne Caulfield, Guido Ivetta, Haiyue Song, Henok Biadglign Ademtew, Hernán Maina, Holy Lovenia, Israel Abebe Azime, Jan Christian Blaise Cruz, Jay Gala, Jiahui Geng, Jesus-German Ortiz-Barajas, Jinheon Baek, Jocelyn Dunstan, Laura Alonso Alemany, Kumaranage Ravindu Yasas Nagasinghe, Luciana Benotti, Luis Fernando D'Haro, Marcelo Viridiano, Marcos Estecha-Garitagoitia, Maria Camila Buitrago Cabrera, Mario Rodríguez-Cantelar, Mélanie Jouitteau, Mihail Mihaylov, Mohamed Fazli Mohamed Imam, Muhammad Farid Adilazuarda, Munkhjargal Gochoo, Munkh-Erdene Otgonbold, Naome Etori, Olivier Niyomugisha, Paula Mónica Silva, Pranjal Chitale, Raj Dabre, Rendi Chevi, Ruochen Zhang, Ryandito Diandaru, Samuel Cahyawijaya, Santiago Góngora, Soyeong Jeong, Sukannya Purkayastha, Tatsuki Kuribayashi, Teresa Clifford, Thanmay Jayakumar, Tiago Timponi Torrent, Toqeer Ehsan, Vladimir Araujo, Yova Kementchedjhieva, Zara Burzo, Zheng Wei Lim, Zheng Xin Yong, Oana Ignat, Joan Nwatu, Rada Mihalcea, Thamar Solorio, and Alham Fikri Aji. 2024. Cvqa: Culturally-diverse multilingual visual question answering benchmark.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Sebastian Ruder, Wei-Yin Ko, Antoine Bosselut, Alice Oh, Andre Martins, Leshem Choshen, Daphne Ippolito, Enzo Ferrante, Marzieh Fadaee, Beyza Ermis, and Sara Hooker. 2025. Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18761–18799, Vienna, Austria. Association for Computational Linguistics.

Gemma Team. 2025. Gemma 3 Technical Report. *CoRR*, abs/2503.19786.

Ashmal Vayani, Dinura Dissanayake, Hasindri Watawana, Noor Ahsan, Nevasini Sasikumar, Omkar Thawakar, Henok Biadglign Ademtew, Yahya Hmaiti, Amandeep Kumar, Kartik Kuckreja, Mykola Maslych, Wafa Al Ghallabi, Mihail Minkov Mihaylov, Chao Qin, Abdelrahman M. Shaker, Mike Zhang, Mahardika Krisna Ihsani, Amiel Gian Esplana, Monil Gokani, Shachar Mirkin, Harsh Singh, Ashay Srivastava, Endre Hamerlik, Fathinah Asma Izzati, Fadillah Adamsyah Maani, Sebastian Cavada, Jenny Chim, Rohit Gupta, Sanjay Manjunath, Kamila Zhumakhanova, Feno Heriniaina Rabevohitra, Azril Hafizi Amirudin, Muhammad Ridzuan, Daniya Najiha Abdul Kareem, Ketan Pravin More, Kunyang Li, Pramesh Shakya, Muhammad Saad, Amirpouya Ghasemaghaei, Amirbek Djanibekov, Dilshod Azizov, Branislava Jankovic, Naman Bhatia, Alvaro Cabrera, Johan S. Obando-Ceron, Olympiah Otieno, Fabian Farestam, Muztoba Rabbani, Sanoojan Baliah, Santosh Sanjeev, Abduragim Shtanchaev, Maheen Fatima, Thao Nguyen, Amrin Kareem, Toluwani Aremu, Nathan Augusto Zacarias Xavier, Amit Bhatkal, Hawau Olamide Toyin, Aman Chadha, Hisham Cholakkal, Rao Muhammad Anwer, Michael Felsberg, Jorma Laaksonen, Thamar Solorio, Monojit Choudhury, Ivan Laptev, Mubarak Shah, Salman H. Khan, and Fahad Shahbaz Khan. 2025. All Languages Matter: Evaluating LMMs on Culturally Diverse 100 Languages. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025*, pages 19565–19575, Nashville, TN, USA. Computer Vision Foundation / IEEE.

Yuiga Wada, Kanta Kaneda, Daichi Saito, and Komei Sugiura. 2024. Polos: Multimodal Metric Learning from Human Feedback for Image

Captioning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024*, pages 13559–13568, Seattle, WA, USA. IEEE.

Hao Wang, Pinzhi Huang, Jihan Yang, Saining Xie, and Daisuke Kawahara. 2025. Traveling Across Languages: Benchmarking Cross-lingual Consistency in Multimodal LLMs.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual E5 Text Embeddings: A Technical Report. *CoRR*, abs/2402.05672.

Genta Indra Winata, Frederikus Hudi, Patrick Amadeus Irawan, David Anugraha, Rifki Afina Putri, Wang Yutong, Adam Nohejl, Ubaidillah Ariq Prathama, Nedjma Ousidhoum, Afifa Amriani, Anar Rzayev, Anirban Das, Ashmari Pramodya, Aulia Adila, Bryan Wilie, Candy Olivia Mawalim, Cheng Ching Lam, Daud Abolade, Emmanuele Chersoni, Enrico Santus, Fariz Ikhwantri, Garry Kuwanto, Hanyang Zhao, Haryo Akbarianto Wibowo, Holy Lovenia, Jan Christian Blaise Cruz, Jan Wira Gotama Putra, Junho Myung, Lucky Susanto, Maria Angelica Riera Machin, Marina Zhukova, Michael Anugraha, Muhammad Farid Adilazuarda, Natasha Christabelle Santosa, Peerat Limkonchotiwat, Raj Dabre, Rio Alexander Audino, Samuel Cahyawijaya, Shi-Xiong Zhang, Stephanie Yulia Salim, Yi Zhou, Yinxuan Gui, David Ifeoluwa Adelani, En-Shiun Annie Lee, Shogo Okada, Ayu Purwarianti, Alham Fikri Aji, Taro Watanabe, Derry Tanti Wijaya, Alice Oh, and Chong-Wah Ngo. 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.

Zhen Xu, Sergio Escalera, Adrien Pavão, Magali Richard, Wei-Wei Tu, Quanming Yao, Huan Zhao, and Isabelle Guyon. 2022. Codabench: Flexible, easy-to-use, and reproducible meta-benchmark platform. *Patterns*, 3(7):100543.

Rustem Yeshpanov, Pavel Efimov, Leonid Boytsov, Ardak Shalkarbayuli, and Pavel Braslavski. 2024. KazQAD: Kazakh open-domain question answering dataset. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9645–9656, Torino, Italia. ELRA and ICCL.

Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. 2024. GLiNER: Generalist model for named entity recognition using bidirectional transformer. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5364–5376, Mexico City, Mexico. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020*, Addis Ababa, Ethiopia. OpenReview.net.

Wenxuan Zhang, Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. 2023. M3Exam: A Multilingual, Multimodal, Multilevel Benchmark for Examining Large Language Models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023*, New Orleans, LA, USA.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and Chatbot Arena. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023*, New Orleans, LA, USA.

Li Zhou, Taelin Karidi, Wanlong Liu, Nicolas Garneau, Yong Cao, Wenyu Chen, Haizhou Li, and Daniel Hershcovich. 2025. Does mapo tofu contain coffee? probing LLMs for food-related cultural knowledge. In *Proceedings of the 2025 Conference of the Nations of the Americas Chap-*

ter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 9840–9867, Albuquerque, New Mexico. Association for Computational Linguistics.

A Prompts

We use all of the following prompts in testing the textual models, while for the visual only the 6th prompt was involved. All of them are followed by an image token (if the question is visual) and the question.

1. You are an experienced trivia game contestant. Provide a short and correct answer!
2. Odpověz pravdivě a krátce na následující otázku:
3. Teraz odpovedajte na nasledujúcu otázku pravdivo a stručne:
4. Відповідайте на наступне питання правдиво і коротко:
5. Jsi zkušený účastník vědomostních soutěží. Odpovídej správně a stručně!
6. You are an experienced trivia game contestant. Provide a short and correct answer in the same language as the question!

For **LLM as a judge**, we used the following prompts:

- System prompt: You are a judge in a trivia game and you are supposed to tell if the contestant answered the question correctly based on the question and the reference answer.
- User prompt: The question is: {question} The contestant's answer is {model_answer}. The correct answer is {correct_answer}. Answer 'Yes' or 'No' and nothing else.

For the **prompt variation study**, we used the following prompts:

(cs 1) Jsi zkušený účastník vědomostních kvízů. Odpověz na následující otázky správně a stručně.

(cs 2) Jsi odborník s neuvěřitelně širokým rozhledem a často vyhráváš populární vědomostní soutěže v rozhlase a televizi. Odpovídej na následující otázku stručně, správně a úplně.

(cs 3) Odpověz na následující otázku stručně a pravdivě!

(cs 4) Vědomostní soutěže jsou pro tebe úplně všechno. Roky jsi intenzivně studoval, abys teď uspěl ve vědomostních soutěžích. Všechno v tvém životě směřovalo k tomu, aby jsi teď bezchybně a stručně odpověděl na následující otázku.

(sk 1) Si skúsený účastník kvízov. Odpovedz na nasledujúce otázky stručne!

(sk 2) Odpovedz na nasledujúce vedomostné otázky správne a čo najstručnejšie.

(sk 3) Si odborník na vedomostné kvízy. Správne a stručne zodpovedaj na otázky, ktoré dostaneš.

(sk 4) V škole si mal samé jednotky, mama ti pred spaním čítala encyklopédie, vyhral si Milionára a doma máš plnú policu fliaš s chľastom z pub kvízov. Zodpovedať nasledujúce otázky tak pre teba musí byť hračka. Správne a stručne.

(uk 1) Ви досвідчений учасник вікторин. Надайте коротку і правильну відповідь!

(uk 2) Ти досвідчений гравець у вікторини. Дай коротку та точну відповідь!

(uk 3) Ти експерт у вікторинах. Відповідай коротко і правильно!

(uk 4) Досвідчений учасник вікторин. Коротка правильна відповідь!

(en 1) You have extensive general knowledge. Answer the following question correctly and concisely.

(en 2) You're a trivia expert with an incredibly broad knowledge base and frequently win popular quiz competitions. Answer the following question briefly, correctly, and completely.

(en 3) Answer the following question concisely and truthfully!

(en 4) Trivia is your life. You've spent years devouring facts, competing in quiz leagues, and crushing every pub quiz in town. Your brain is a finely-tuned knowledge machine. Now prove it by answering the following question correctly and briefly.

We use the following prompts for **retrieval augmented generation**.

- You are an experienced trivia game contestant. Provide a short and correct answer to the question! Attached to the question is a snippet from the local Wikipedia. You may use the information in the context as you see fit. Do not mention the context snippet in your reply. Reply in the original language of the question.

Czechia	
Kdo byl jediným českým králem, který nepocházel z panovnické dynastie? *Who was the only Czech king who did not come from a ruling dynasty?*	Jiří z Poděbrad. *George of Poděbrady*
V kterém roce proběhla bitva na Bílé hoře? *In what year did the Battle of White Mountain take place?*	1620 *1620*
Ve kterém městě sídlí Tatra? *In which city is the Tatra factory located?*	V Kopřivnici. *In Kopřivnice.*
Jak se jmenuje zřícenina hradu založená Karlem IV. jižně od Starého Plzence, která je kulturní památkou? *What is the name of the ruins of the castle founded by Charles IV south of Stary Plzenec, which is a cultural monument?*	Radyně *Radyně*
Který lihovar vyrábí OMG gin? *Which distillery produces OMFG gin?*	Žufánek *Žufánek*
Která linka pražského metra je nejstarší? *Which Prague metro line is the oldest?*	Linka C *The C line*
Slovakia	
Ktorá kapela naspievala album Nikdy nebolo lepšie? *Which band sang the album Never Been Better?*	Hex *Hex*
Ako sa volá najvyšší vrch Belianskych Tatier? *What is the name of the highest peak of the Beliany Tatras?*	Havran *Havran*
Kto založil Medzinárodný maratón mieru v Košiciach? *Who founded the International Peace Marathon in Košice?*	Vojtech Bukovský. *Vojtech Bukovský.*
Ktorá televízna stanica vysiela šou s názvom Tvoja tvár znie povedome? *Which television station airs a show called Tvoja tvár znie povedome?*	televízia Markíza *television Markíza*
Ako sa volal politický spor o názov Česko-Slovenska na začiatku roka 1990? *What was the name of the political conflict regarding the name of Czechoslovakia at the beginning of the year 1990?*	pomlčková vojna *The Hyphen War*
Ktorá diaľnica prechádza nedaľeko Nového Mesta nad Váhom? *Which highway passes near Nové Mesto nad Váhom?*	D1 *D1*
Ukraine	
На якій вулиці знаходиться Будинок з химерами у Харкові? *What street is the House with Chimeras located on in Kharkiv?*	На вулиці Чернишевській *On Chernyshevskaya Street*
Який оператор мобільного зв'язку в 2015 викупив право надання послуг у компанії МТС? *Which mobile operator bought the right to provide services from MTS in 2015?*	Vodafone *Vodafone*
Які кольори футбольного клубу Нива? *What are the colors of Niva football club?*	Жовтий і зелений *Yellow and green*
Яку міжнародну премію зміг отримати Іван Данилович Низовий? *What international award was Ivan Danylovych Nyzovyi able to receive?*	За найкращий музичний твір. *For the best musical composition.*
Як називається головна вулиця Києва? *What is the name of the main street in Kyiv?*	Хрещатик *Khreshchatyk*
Хто очолював УНР до квітня 1918 року? *Who led the UPR until April 1918?*	Михайло Грушевський *Mykhailo Hrushevsky*

Table 19: Examples of text-only questions and answers.

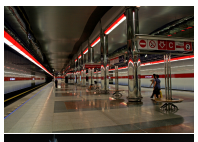
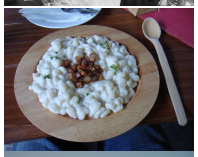
	Image	Question	Answer
Czechia		Která stanice pražského metra je na obrázku? *Which Prague metro station is in the picture?*	Letňany *Letňany*
		Kdo je autorem této sochy? *Who is the author of this statue?*	Josef Václav Myslbek. *Josef Václav Myslbek.*
		Jaké značky je tento autobus? *What brand is this bus?*	Karosa. *Karosa.*
Slovakia		Ktorý bývalý slovenský tenista sa nachádza na fotke? *Which former Slovak tennis player is in the picture?*	Miloslav Mečíř. *Miloslav Mečíř.*
		Čo je na obrázku? *What is in the picture?*	Bryndzové halušky. *Bryndzové halušky.*
		Aká budova je na obrázku? *What building is in the picture?*	Nová budova Slovenského národného divadla *New building of the Slovak National Theatre*
Ukraine		Яку українську страву зображено на фото? *What Ukrainian dish is depicted in the photo?*	Банош Banosh
		Як називається цей стадіон? *What is the name of this stadium?*	«Авангард» “Avangard”
		Де знаходиться зображений на картинці замок? *Where is the castle in the picture located?*	Ужгород *Uzhhorod*

Table 20: Examples of collected questions and answers from all regions, in both the local language and in English. Images: *Metro Letňany 1*, Karel Jakubec, Public domain, *Prag Wenzelsplatz Wenzelsdenkmal Nationalmuseum bei Nacht*, Andreas Praefcke, CC BY 3.0, *Brno, Starý Lískovec, zastávka Nemocnice Bohunice, Karosa B 732 č. 7396*, Harold, CC BY-SA 3.0, *Tennis, Melkhuisje open kampioenschappen winnaar Mecir met beker, Bestanddeelnr 934-0434*, Rob Croes / Anefo, Public domain, *Bryndzové halušky so slaninou*, Gregory finster, Public domain, *Bratislava*, Wizard, Public Domain, *Banusz (banosz) w restauracji-piwowarni "Kumpel" we Lwowie (ul. Volodymyra Vynnychenka 6)*, Trzewik, CC BY-SA 4.0, *Avanhard Stadium, Lutsk*, FC Volyn Lutsk, CC BY-SA 1.0, *Ужгородський замок*, Irkap, CC BY-SA 4.0.

Textual QA: Czechia		Manual evaluation				Automatic evaluation						
		Correct	+ True	+ Relevant	+ Coherent	BLEU	chrF	ROUGE	BERTScore	BLEURT	LLaMA	M-P
cs	EuroLLM-9B-Instruct	35.5	34.3	31.1	27.9	2.4	26.2	.186	.221	.382	35.8	30.4
	Llama-3.1-8B-Instruct	28.7	28.3	27.4	23.6	3.1	26.5	.212	.275	.374	28.7	24.0
	Llama-3.3-70B-Instruct	58.7	57.4	56.4	53.4	4.7	42.4	.365	.366	.510	57.4	48.9
	Llama-4-Scout-17B-16E-Instruct	39.2	37.2	35.5	30.6	3.8	32.6	.279	.323	.451	39.1	30.9
	Mistral-7B-Instruct-v0.3	23.0	22.3	20.0	14.3	1.4	20.2	.139	.177	.328	22.8	17.5
en	EuroLLM-9B-Instruct	31.5	28.7	25.3	23.6	1.4	20.4	.142	.075	.271	31.5	26.8
	Llama-3.1-8B-Instruct	29.4	28.5	27.4	26.0	1.4	24.1	.191	.162	.317	31.3	26.0
	Llama-3.3-70B-Instruct	52.5	50.4	47.5	44.7	3.4	37.8	.350	.340	.439	56.0	45.5
	Llama-4-Scout-17B-16E-Instruct	37.0	34.2	30.4	28.7	1.3	23.2	.175	.093	.313	38.3	32.1
	Mistral-7B-Instruct-v0.3	24.3	20.9	9.6	9.2	0.7	14.5	.087	-.003	.277	26.8	20.2

Textual QA: Slovakia		Manual evaluation				Automatic evaluation						
		Correct	+ True	+ Relevant	+ Coherent	BLEU	chrF	ROUGE	BERTScore	BLEURT	LLaMA	M-P
sk	EuroLLM-9B-Instruct	32.0	30.8	29.4	25.4	2.6	24.2	.169	.677	.370	30.6	27.8
	Llama-3.1-8B-Instruct	18.9	18.1	17.0	15.0	2.7	21.2	.152	.689	.335	18.5	13.8
	Llama-3.3-70B-Instruct	46.2	45.8	43.4	40.8	7.4	39.2	.354	.755	.485	45.4	38.5
	Llama-4-Scout-17B-16E-Instruct	34.1	33.5	31.8	30.6	3.9	29.2	.251	.706	.425	32.3	28.8
	Mistral-7B-Instruct-v0.3	12.8	12.2	11.4	6.5	0.8	15.5	.090	.632	.284	14.2	12.2
en	EuroLLM-9B-Instruct	30.8	28.4	26.2	25.2	2.0	21.4	.158	.090	.284	33.1	26.8
	Llama-3.1-8B-Instruct	24.5	24.1	22.9	22.3	2.0	22.6	.186	.145	.309	25.4	19.5
	Llama-3.3-70B-Instruct	45.8	44.8	42.2	40.4	3.2	32.5	.298	.269	.397	47.5	38.1
	Llama-4-Scout-17B-16E-Instruct	32.9	30.8	27.8	27.0	1.8	22.6	.180	.094	.311	33.7	27.6
	Mistral-7B-Instruct-v0.3	20.9	17.2	11.8	11.8	1.0	14.9	.101	-.002	.280	22.7	18.5

Textual QA: Ukraine		Manual evaluation				Automatic evaluation						
		Correct	+ True	+ Relevant	+ Coherent	BLEU	chrF	ROUGE	BERTScore	BLEURT	LLaMA	M-P
uk	EuroLLM-9B-Instruct	26.8	26.2	25.7	22.6	2.4	26.2	.186	.221	.382	35.8	30.4
	Llama-3.1-8B-Instruct	22.6	21.6	21.3	20.3	3.1	26.5	.212	.275	.374	28.7	24.0
	Llama-3.3-70B-Instruct	51.9	50.9	47.3	46.2	4.7	42.4	.365	.366	.510	57.4	48.9
	Llama-4-Scout-17B-16E-Instruct	49.6	48.1	46.5	44.9	3.8	32.6	.279	.323	.451	39.1	30.9
	Mistral-7B-Instruct-v0.3	21.3	20.0	18.7	15.1	1.4	20.2	.139	.177	.328	22.8	17.5
en	EuroLLM-9B-Instruct	26.8	25.5	22.9	21.3	1.4	20.4	.142	.075	.271	31.5	26.8
	Llama-3.1-8B-Instruct	30.4	29.4	27.3	27.3	1.4	24.1	.191	.162	.317	31.3	26.0
	Llama-3.3-70B-Instruct	47.8	46.0	43.6	43.1	3.4	37.8	.350	.340	.439	56.0	45.5
	Llama-4-Scout-17B-16E-Instruct	40.8	39.2	35.6	35.3	1.3	23.2	.175	.093	.313	38.3	32.1
	Mistral-7B-Instruct-v0.3	29.4	25.5	15.1	14.8	0.7	14.5	.087	-.003	.277	26.8	20.2

Table 21: Detailed results of textual QA on the development data, including strictness levels of manual evaluation and all tested automatic metrics.

Visual QA: Czechia		Manual evaluation				Automatic evaluation						
		Correct	+ True	+ Relevant	+ Coherent	BLEU	chrF	ROUGE	BERTScore	BLEURT	LLaMA	M-P
cs	gemma3	12.4	12.4	11.5	8.8	2.7	20.2	.213	.325	.359	11.1	7.5
	idefics	5.8	5.8	4.4	3.5	0.9	10.0	.082	.221	.209	7.5	4.4
	Llama-3.2-11B-Vision-Instruct	9.7	9.7	9.3	1.8	0.1	9.1	.053	.093	.208	9.7	8.8
	Llama-4-Scout-17B-16E-Instruct	14.6	13.7	11.9	11.1	2.7	21.3	.218	.146	.356	14.6	8.8
	maya	0.9	0.9	0.9	0.9	0.1	11.6	.082	.153	.239	1.3	1.8
en	gemma3	10.2	10.2	8.4	4.4	3.1	15.9	.115	.116	.259	13.3	4.0
	idefics	3.1	3.1	2.7	1.3	0.3	10.4	.064	.137	.202	7.1	3.1
	Llama-3.2-11B-Vision-Instruct	10.6	10.6	8.8	8.0	0.7	14.9	.100	.024	.245	13.7	8.0
	Llama-4-Scout-17B-16E-Instruct	13.3	12.8	11.5	8.4	1.5	15.5	.131	-.077	.275	15.9	8.4
	maya	0.0	0.0	0.0	0.0	0.2	12.3	.069	.038	.186	2.2	0.0

Visual QA: Slovakia		Manual evaluation				Automatic evaluation						
		Correct	+ True	+ Relevant	+ Coherent	BLEU	chrF	ROUGE	BERTScore	BLEURT	LLaMA	M-P
sk	gemma3	18.6	18.6	18.6	17.8	9.7	24.9	.202	.741	.399	13.6	8.5
	idefics	9.3	9.3	8.5	6.8	3.2	15.7	.116	.699	.252	8.5	4.2
	Llama-3.2-11B-Vision-Instruct	12.7	12.7	12.7	0.8	0.6	13.4	.049	.652	.200	10.2	8.5
	Llama-4-Scout-17B-16E-Instruct	23.7	23.7	23.7	20.3	4.8	24.4	.214	.666	.412	18.6	10.2
	maya	0.8	0.8	0.8	0.8	0.5	15.6	.087	.667	.240	0.0	0.0
en	gemma3	13.6	13.6	13.6	10.2	2.8	15.6	.126	.152	.306	12.7	10.2
	idefics	8.5	8.5	7.6	5.1	0.4	14.5	.115	.178	.257	10.2	5.9
	Llama-3.2-11B-Vision-Instruct	12.7	12.7	12.7	11.0	1.4	18.2	.104	.043	.251	11.0	10.2
	Llama-4-Scout-17B-16E-Instruct	19.5	19.5	19.5	12.7	0.7	16.6	.149	.005	.327	16.9	14.4
	maya	0.8	0.8	0.8	0.8	1.1	15.7	.090	.082	.184	0.0	0.0

Visual QA: Ukraine		Manual evaluation				Automatic evaluation						
		Correct	+ True	+ Relevant	+ Coherent	BLEU	chrF	ROUGE	BERTScore	BLEURT	LLaMA	M-P
uk	gemma3	25.5	25.5	25.5	24.0	4.2	19.9	.026	.721	.338	21.6	14.7
	idefics	7.8	7.8	7.8	7.4	0.8	4.4	.005	.666	.156	7.8	3.9
	Llama-3.2-11B-Vision-Instruct	16.2	16.2	16.2	8.3	0.3	6.5	.005	.647	.197	15.2	10.3
	Llama-4-Scout-17B-16E-Instruct	29.9	29.9	29.9	29.9	2.1	21.1	.023	.663	.357	26.0	15.2
	maya	3.4	3.4	3.4	2.5	0.2	12.3	.010	.656	.171	2.9	2.0
en	gemma3	19.1	19.1	19.1	15.2	3.2	16.9	.155	.175	.294	21.6	15.7
	idefics	8.8	8.8	8.8	8.3	0.9	9.1	.099	.233	.215	11.3	4.9
	Llama-3.2-11B-Vision-Instruct	17.2	17.2	17.2	17.2	1.5	19.1	.114	.110	.276	20.1	12.7
	Llama-4-Scout-17B-16E-Instruct	25.5	25.0	25.0	21.1	2.4	19.4	.157	.023	.306	30.4	20.1
	maya	4.4	4.4	4.4	3.9	1.9	17.0	.085	.131	.201	5.4	2.9

Table 22: Detailed results on the visual QA on development data, including strictness levels of manual evaluation and all tested automatic metrics.

	Model	Czechia (cs)				Czechia (en)				Slovakia (sk)				Slovakia (en)				Ukraine (uk)				Ukraine (en)			
		geo	cul	his	oth	geo	cul	his	oth	geo	cul	his	oth	geo	cul	his	oth	geo	cul	his	oth	geo	cul	his	oth
Textual	EuroLLM 9B Ins.	40	14	33	12	33	12	29	12	30	16	37	20	30	15	35	22	30	11	24	21	23	14	26	19
	LLaMA 3.1 8B Ins.	32	9	31	15	34	13	37	14	15	8	15	23	27	10	21	27	25	8	23	21	31	19	25	31
	LLaMA 3.3 70B Ins.	65	41	57	38	58	34	44	28	49	29	44	35	49	24	44	38	52	41	42	46	49	34	39	46
	LLaMA 4 Scout Ins.	44	14	33	23	39	12	35	25	35	12	37	39	38	11	21	25	57	34	34	49	45	21	36	32
	Mistral 7B Ins. v0.3	15	8	20	15	16	5	7	3	9	3	8	5	17	8	4	8	17	10	13	19	20	5	12	19
Visual	gemma3	6	14	15	11	3	6	5	11	15	10	29	33	12	6	14	10	20	13	28	36	15	13	14	19
	idefics	3	6	0	5	1	0	5	5	3	10	0	14	5	0	0	14	8	5	8	7	9	8	10	5
	LLaMA 3.2 11B Vis. Ins.	2	0	0	5	10	4	0	11	0	3	0	0	15	0	0	19	7	5	10	12	14	13	24	19
	LLaMA 4 Scout Ins.	12	8	15	11	8	10	0	16	25	10	0	29	17	0	14	19	27	24	32	38	15	16	28	29
	maya	1	0	0	0	0	0	0	0	0	0	0	5	0	0	0	5	5	0	0	2	4	0	4	7

Table 23: Per-category breakdown how human evaluation scores (on the strictest level) fro evaluted model. We considered geography, culture and history. Politics, sports and other were put in the same category because they would be too sparse.

Loc.	Model	Local language										English									
		content					text		medium			content					text		medium		
		portrait	building	city	nature	other	text	no text	color photo	BW photo	painting	portrait	building	city	nature	other	text	no text	color photo	BW photo	painting
Czechia	gemma3	8	6	9	0	28	15	8	9	9	12	0	5	3	0	16	8	4	5	0	12
	idefics	2	2	4	6	12	15	1	3	4	12	2	1	1	0	4	5	1	1	4	0
	LLaMA 3.2 11B Vis. Ins.	0	2	1	0	8	5	1	2	0	0	0	8	12	12	16	10	8	9	4	0
	LLaMA 4 Scout Ins.	4	10	12	6	28	20	9	12	9	0	4	9	15	0	16	15	7	9	4	0
	maya	0	1	3	0	0	5	0	1	0	0	0	0	0	0	0	0	0	0	0	0
Slovakia	gemma3	11	21	33	0	18	41	10	18	12	50	6	11	17	0	9	24	6	10	0	50
	idefics	6	10	6	0	0	21	2	7	0	0	3	5	17	0	0	10	3	6	0	0
	LLaMA 3.2 11B Vis. Ins.	0	2	0	0	0	3	0	1	0	0	3	16	22	0	0	21	8	12	0	0
	LLaMA 4 Scout Ins.	3	30	44	17	18	41	13	22	0	0	8	16	22	0	9	31	7	12	12	50
	maya	3	0	0	0	0	3	0	1	0	0	0	2	0	0	0	0	1	1	0	0
Ukraine	gemma3	31	14	24	10	31	35	21	23	14	38	15	8	9	10	25	13	16	16	10	14
	idefics	8	5	7	0	12	7	8	8	5	5	8	5	11	0	12	7	9	9	5	5
	LLaMA 3.2 11B Vis. Ins.	8	4	7	0	15	15	6	10	0	5	15	16	20	10	21	26	15	18	10	19
	LLaMA 4 Scout Ins.	33	25	29	30	33	39	27	31	29	24	21	21	16	10	25	35	17	21	14	29
	maya	0	1	4	10	4	2	3	3	0	0	0	3	2	10	8	2	4	4	0	5

Table 24: Breakdown of the model accuracy in human evaluation based on the visual content of the images. Categories of food and technology are merged into others due to a low number of examples.

	Textual							Visual						
	BLEU	chrF	RG	BScr	BLRT	LLM	M-P	BLEU	chrF	RG	BScr	BLRT	LLM	M-P
Correct	.884	.927	.917	.857	.931	.963	.951	.577	.670	.842	-.108	.892	.990	.961
+ true	.897	.943	.933	.877	.937	.964	.955	.577	.668	.839	-.098	.890	.989	.962
+ relevant	.904	.954	.941	.897	.920	.945	.943	.571	.668	.835	-.114	.888	.981	.965
+ coherent	.907	.956	.944	.905	.923	.942	.940	.605	.773	.882	-.039	.932	.908	.848

Table 25: System-level Pearson correlations of the automatic evaluation metrics with different levels of strictness of human evaluation. The presented numbers are averages of Pearson correlations computed separately for each country and language.

Textual QA

CZ: cs	BL-EU	chrF	ROU-GE	BERT-Score	BLE-URT	LLa-MA	M-P
Correct	.90	.90	.90	.90	1.00	1.00	1.00
+ true	.90	.90	.90	.90	1.00	1.00	1.00
+ relevant	.90	.90	.90	.90	1.00	1.00	1.00
+ coherent	.90	.90	.90	.90	1.00	1.00	1.00

SK: sk	BL-EU	chrF	ROU-GE	BERT-Score	BLE-URT	LLa-MA	M-P
Correct	.90	1.00	1.00	.90	1.00	1.00	1.00
+ true	.90	1.00	1.00	.90	1.00	1.00	1.00
+ relevant	.90	1.00	1.00	.90	1.00	1.00	1.00
+ coherent	.90	1.00	1.00	.90	1.00	1.00	1.00

UA: uk	BL-EU	chrF	ROU-GE	BERT-Score	BLE-URT	LLa-MA	M-P
Correct	.90	.90	.90	.90	1.00	1.00	1.00
+ true	.90	.90	.90	.90	1.00	1.00	1.00
+ relevant	.90	.90	.90	.90	1.00	1.00	1.00
+ coherent	.90	.90	.90	.90	1.00	1.00	1.00

CZ: en	BL-EU	chrF	ROU-GE	BERT-Score	BLE-URT	LLa-MA	M-P
Correct	.60	.70	.70	.70	.50	1.00	1.00
+ true	.60	.70	.70	.70	.50	1.00	1.00
+ relevant	.70	.90	.90	.90	.80	.90	.90
+ coherent	.70	.90	.90	.90	.80	.90	.90

SK: en	BL-EU	chrF	ROU-GE	BERT-Score	BLE-URT	LLa-MA	M-P
Correct	.60	.90	.70	.70	.90	1.00	1.00
+ true	.60	.90	.70	.70	.90	1.00	1.00
+ relevant	.60	.90	.70	.70	.90	1.00	1.00
+ coherent	.60	.90	.70	.70	.90	1.00	1.00

UA: en	BL-EU	chrF	ROU-GE	BERT-Score	BLE-URT	LLa-MA	M-P
Correct	.50	.80	.80	.80	.90	.70	.70
+ true	.62	.87	.87	.87	.87	.82	.82
+ relevant	.70	.90	.90	.90	.80	.90	.90
+ coherent	.70	.90	.90	.90	.80	.90	.90

Visual QA

CZ: cs	BL-EU	chrF	ROU-GE	BERT-Score	BLE-URT	LLa-MA	M-P
Correct	.90	.60	.60	-.10	.50	1.00	.82
+ true	.90	.60	.60	-.10	.50	1.00	.82
+ relevant	.90	.60	.60	-.10	.50	1.00	.82
+ coherent	1.00	.70	.70	.20	.60	.90	.56

SK: sk	BL-EU	chrF	ROU-GE	BERT-Score	BLE-URT	LLa-MA	M-P
Correct	.80	.60	.70	-.10	.70	1.00	.97
+ true	.80	.60	.70	-.10	.70	1.00	.97
+ relevant	.80	.60	.70	-.10	.70	1.00	.97
+ coherent	.87	.87	.97	.31	.97	.82	.71

UA: uk	BL-EU	chrF	ROU-GE	BERT-Score	BLE-URT	LLa-MA	M-P
Correct	.80	.70	.56	.30	.90	1.00	1.00
+ true	.80	.70	.56	.30	.90	1.00	1.00
+ relevant	.80	.70	.56	.30	.90	1.00	1.00
+ coherent	.80	.70	.56	.30	.90	1.00	1.00

CZ: en	BL-EU	chrF	ROU-GE	BERT-Score	BLE-URT	LLa-MA	M-P
Correct	.70	.60	.80	-.70	.90	1.00	1.00
+ true	.70	.60	.80	-.70	.90	1.00	1.00
+ relevant	.70	.60	.80	-.70	.90	1.00	1.00
+ coherent	.70	.60	.80	-.70	.90	1.00	1.00

SK: en	BL-EU	chrF	ROU-GE	BERT-Score	BLE-URT	LLa-MA	M-P
Correct	.20	.30	.90	-.50	.90	1.00	.97
+ true	.20	.30	.90	-.50	.90	1.00	.97
+ relevant	.20	.30	.90	-.50	.90	1.00	.97
+ coherent	.10	.60	.70	-.70	.70	.90	.97

UA: en	BL-EU	chrF	ROU-GE	BERT-Score	BLE-URT	LLa-MA	M-P
Correct	.60	.50	1.00	-.50	1.00	1.00	1.00
+ true	.60	.50	1.00	-.50	1.00	1.00	1.00
+ relevant	.60	.50	1.00	-.50	1.00	1.00	1.00
+ coherent	.30	.70	.90	-.70	.90	.90	.90

Table 26: System-level Spearman correlation of automatic metrics and human judgment split over countries and languages. Note that each of the tables is only based on a comparison of 5 systems.

Textual QA

CZ: cs	chrF	ROU-GE	BLE-URT	LLa-MA	M-P
Correct	.716	.656	.682	.893	.795
+ true	.707	.649	.671	.879	.793
+ relevant	.705	.647	.658	.859	.793
+ coherent	.664	.629	.622	.801	.756

SK: sk	chrF	ROU-GE	BLE-URT	LLa-MA	M-P
Correct	.649	.575	.562	.867	.827
+ true	.645	.578	.564	.854	.820
+ relevant	.621	.554	.542	.832	.803
+ coherent	.591	.543	.530	.797	.789

UA: uk	chrF	ROU-GE	BLE-URT	LLa-MA	M-P
Correct	.625	.147	.577	.884	.788
+ true	.617	.145	.565	.877	.791
+ relevant	.616	.144	.557	.856	.780
+ coherent	.601	.138	.544	.815	.751

CZ: en	chrF	ROU-GE	BLE-URT	LLa-MA	M-P
Correct	.571	.534	.554	.714	.641
+ true	.558	.528	.535	.690	.639
+ relevant	.527	.506	.506	.646	.608
+ coherent	.514	.495	.494	.628	.594

SK: en	chrF	ROU-GE	BLE-URT	LLa-MA	M-P
Correct	.515	.520	.489	.712	.637
+ true	.509	.517	.480	.695	.625
+ relevant	.475	.489	.461	.659	.600
+ coherent	.465	.482	.454	.647	.595

UA: en	chrF	ROU-GE	BLE-URT	LLa-MA	M-P
Correct	.485	.430	.450	.671	.556
+ true	.475	.417	.442	.652	.552
+ relevant	.460	.401	.422	.606	.512
+ coherent	.454	.400	.419	.599	.506

Visual QA

CZ: cs	chrF	ROU-GE	BLE-URT	LLa-MA	M-P
Correct	.374	.334	.363	.807	.638
+ true	.376	.335	.359	.795	.631
+ relevant	.404	.359	.390	.799	.672
+ coherent	.365	.378	.349	.679	.602

SK: sk	chrF	ROU-GE	BLE-URT	LLa-MA	M-P
Correct	.492	.456	.480	.802	.606
+ true	.492	.456	.480	.802	.606
+ relevant	.488	.449	.478	.808	.610
+ coherent	.435	.462	.466	.721	.536

UA: uk	chrF	ROU-GE	BLE-URT	LLa-MA	M-P
Correct	.399	.223	.487	.805	.666
+ true	.399	.223	.487	.805	.666
+ relevant	.399	.223	.487	.805	.666
+ coherent	.463	.205	.509	.752	.644

CZ: en	chrF	ROU-GE	BLE-URT	LLa-MA	M-P
Correct	.392	.397	.405	.760	.711
+ true	.389	.396	.402	.754	.700
+ relevant	.381	.383	.399	.693	.763
+ coherent	.322	.331	.331	.608	.718

SK: en	chrF	ROU-GE	BLE-URT	LLa-MA	M-P
Correct	.410	.428	.486	.856	.786
+ true	.410	.428	.486	.856	.786
+ relevant	.405	.423	.482	.864	.793
+ coherent	.379	.453	.420	.730	.691

UA: en	chrF	ROU-GE	BLE-URT	LLa-MA	M-P
Correct	.384	.381	.479	.810	.784
+ true	.384	.381	.477	.813	.787
+ relevant	.384	.381	.477	.813	.787
+ coherent	.448	.426	.463	.767	.730

Table 27: Answer-level point biserial correlation of automatic metrics and human judgment split over countries and languages. Note that each of the tables is only based on a comparison of 5 systems.

	Model	Czechia				Slovakia				Ukraine			
		cs		en		sk		en		uk		en	
Textual	EuroLLM-9B-Instruct	32	3	25	7	27	5	21	10	24	3	17	10
		3	62	11	57	3	65	9	60	2	71	10	64
	Llama-3.1-8B-Instruct	26	2	23	7	17	2	20	5	22	1	22	8
		2	69	8	62	1	80	6	70	2	76	7	63
	Llama-3.3-70B-Instruct	55	3	49	4	42	4	41	4	49	3	44	3
		2	39	11	37	2	52	7	47	5	43	10	42
	Llama-4-Scout-17B-16E-Instruct	37	3	33	4	31	3	28	5	45	4	36	4
		2	59	6	57	2	64	6	61	2	49	9	50
	Mistral-7B-Instruct-v0.3	21	2	20	5	11	2	16	5	19	3	22	8
		2	75	5	71	3	85	5	74	2	76	7	63
Visual	gemma3	10	2	9	1	14	5	11	3	21	5	18	1
		0	87	4	86	0	81	1	86	1	73	3	77
	idefics	4	1	3	0	6	3	8	1	6	2	8	1
		4	91	4	93	3	88	3	89	2	90	3	88
	Llama-3.2-11B-Vision-Instruct	9	1	10	0	9	3	11	2	13	3	16	1
		1	89	4	86	1	86	0	87	2	82	4	79
	Llama-4-Scout-17B-16E-Instruct	12	3	12	2	19	5	15	4	23	7	23	3
		3	83	4	82	0	76	0	81	2	68	8	66
	maya	1	0	0	0	0	1	0	1	3	0	4	0
		0	99	2	98	0	99	0	99	0	97	0	95

Legend: TP FN / FP TN — TP/TN color scale: 0 10 20 30 40 50 60 70 80 90 100 — FP/FN color scale: 0 10 20 30 40 50 60 70 80 90 100

Table 28: Confusion matrices of LLaMA 3.3 70B Instruct as a judge compared to manually evaluated correctness.

Model	CZ		SK		UA	
	cs	en	sk	en	uk	en
chrF						
EuroLLM 9B Ins.	19.9±3.7	19.0±1.4	21.2±2.0	19.7±1.7	20.6±2.2	21.0±1.5
Llama 3.1 8B Ins.	23.3±0.8	20.0±2.5	21.4±1.2	19.4±2.0	25.2±0.6	23.0±1.5
Llama 3.3 70B Ins.	40.2±2.9	30.3±5.3	36.7±1.3	26.2±4.3	38.7±2.6	29.5±2.7
Llama 4 Scout Ins.	28.9±1.9	19.2±2.2	28.4±1.0	19.3±1.9	34.6±2.8	22.9±3.5
Mistral 7B Ins. v0.3	16.9±2.0	13.4±1.6	17.1±1.4	13.7±1.9	24.5±1.2	16.5±1.8
LLM as a Judge						
EuroLLM 9B Ins.	42.8±1.6	35.4±2.4	33.7±0.7	32.5±0.7	27.4±1.9	30.8±1.5
Llama 3.1 8B Ins.	30.5±1.9	31.4±2.3	20.6±1.3	25.4±3.2	25.0±1.4	32.5±2.2
Llama 3.3 70B Ins.	58.0±1.0	58.7±1.6	46.6±1.4	49.5±1.8	53.1±1.0	54.0±0.6
Llama 4 Scout Ins.	41.2±1.0	38.8±1.3	32.5±1.2	33.5±0.7	47.5±1.5	44.3±2.7
Mistral 7B Ins. v0.3	24.4±1.3	27.4±0.4	17.2±0.7	23.2±0.7	23.4±1.0	32.0±0.7

Table 29: Average chrF and LLM as a judge scores and their standard deviation for prompt variation.