🧘 CaLMQA: Exploring culturally specific long-form question answering across 23 languages

Shane Arora[♣*] **Marzena Karpinska**[♡*] **Hung-Ting Chen**[♢]
Ipsita Bhattacharjee[♡] **Mohit Iyyer**[♡†] **Eunsol Choi**[♢†]
The University of Texas at Austin[♣]
New York University[♢]
University of Massachusetts Amherst[♡]
shane.arora@utexas.edu {hung-ting.chen, eunsol}@nyu.edu
{mkarpinska, ibhattacharj, miyyer}@umass.edu

Abstract

Despite rising global usage of large language models (LLMs), their ability to generate *long-form* answers to *culturally specific* questions remains unexplored in many languages. To fill this gap, we perform the first study of textual multilingual long-form QA by creating CaLMQA, a dataset of **51.7K** culturally specific questions across **23** different languages. We define culturally specific questions as those that refer to concepts unique to one or a few cultures, or have different answers depending on the cultural or regional context. We obtain these questions by crawling naturally-occurring questions from community web forums in high-resource languages, and by hiring native speakers to write questions in under-resourced, rarely-studied languages such as Fijian and Kirundi. Our data collection methodologies are translation-free, enabling the collection of culturally unique questions like 'Kuber iki umwami wa mbere w'uburundi yitwa Ntare?" (Kirundi; English translation: "Why was the first king of Burundi called Ntare (Lion)?"). We evaluate factuality, relevance and surface-level quality of LLM-generated long-form answers, finding that (1) for many languages, even the best models make critical surface-level errors (e.g., answering in the wrong language, repetition), especially for low-resource languages; and (2) answers to culturally specific questions contain more factual errors than answers to culturally agnostic questions – questions that have consistent meaning and answer across many cultures. We release CaLMQA to facilitate future research in cultural and multilingual long-form QA.

github.com/2015aroras/CaLMQA
🤗 hf.co/datasets/shanearora/CaLMQA

[*†]These authors contributed equally to this work.

1 Introduction

While large language models (LLMs) are increasingly used by people across the world, most NLP efforts are focused on English and western cultures. Growing evidence reveals significant gaps in their performance across languages (Qiu et al., 2023; Guerreiro et al., 2023) and their understanding of diverse cultures (Tao et al., 2024; Li et al., 2024), as well as a persistent bias toward Western-centric perspectives (Palta and Rudinger, 2023; Durmus et al., 2024; AlKhamissi et al., 2024; Naous et al., 2024). Existing research of multilingual QA largely focuses on assets derived from English resources (Singh et al., 2024; Zhang et al., 2023; Lai et al., 2023), limiting their coverage of culturally unique concepts especially in low-resource languages. While some prior work collects short-answer and multiple-choice questions in non-English languages (Myung et al., 2025; Clark et al., 2020; Liu et al., 2019), multilingual long-form QA, a task more aligned with real-world applications, remains unexplored.

In this work, we develop a translation-free multilingual QA dataset of long-form culturally specific questions: **C**ultur**a**l **L**ong-form **M**ultilingual **Q**uestion **A**nswering (CaLMQA). Questions are posed in the language of the target culture and demand nuanced, long-form responses. We only collect ***culturally specific*** questions that (1) refer to concepts unique to one or a few cultures, such as "Kuber iki umwami wa mbere w'uburundi yitwa Ntare?" (Kirundi),[1] or (2) have different answers depending on the cultural or regional context, as in "बंदूक का लाइसेंस कैसे बनता है?" (Hindi).[2] We contrast the quality of an LLM's answers to these questions with its answers to ***culturally agnostic*** questions that have consistent meaning and

[1]English translation: "Why was the first king of Burundi called Ntare (Lion)?"

[2]English translation: "How do you get a gun license?"

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 11772–11817
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

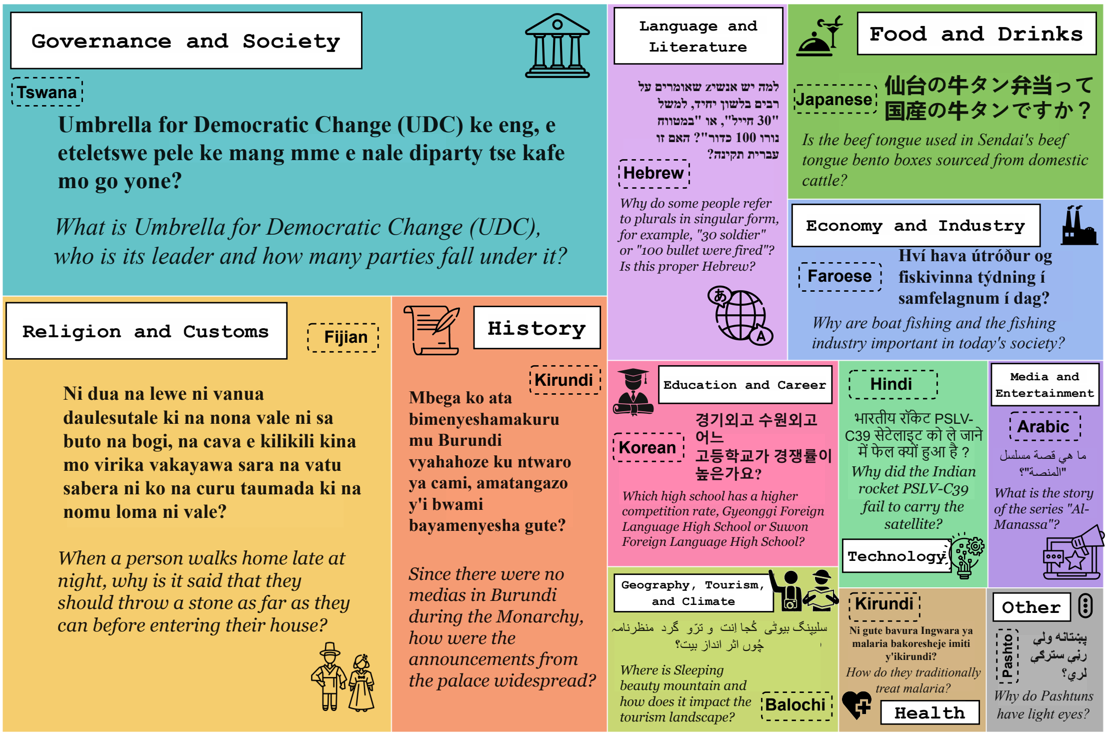

Figure 1: Distribution of topics in CALMQA, with box size indicating the frequency of each topic. Each topic is accompanied by an example and its English translation. Table 12 contains descriptions of the topics, and §B describes our topic classification method.

answer across many cultures (e.g., "Why is smoking bad for the heart?"), which are prevalent in many translation-centric multilingual QA works.

Evaluation of multilingual long-form QA is challenging: lexical metrics for short-form QA do not correlate with human preferences in long-form QA (Krishna et al., 2021; Xu et al., 2023) or transfer from English to other languages (Kang et al., 2024; Koto et al., 2021; Min et al., 2023; Song et al., 2024). We thus adopt a multi-aspect evaluation protocol including (1) surface-level measures of language identification and repetition; (2) automatic factuality and relevance metrics run on translated answers; and (3) human evaluations from native speakers. To distinguish the effects of culture and language on model performance, we use a baseline set of parallel culturally agnostic questions created by translating a seed set of 51 English questions into the 22 other languages, following common practice in prior work (Vayani et al., 2024; Artetxe et al., 2020; Lewis et al., 2020; Alonso et al., 2024).

We show that seven popular LLMs, including closed models such as CLAUDE-3-OPUS, GEMINI-1.5-PRO and GPT-4O, suffer from basic surface-level issues, especially on low-resource languages (e.g., none of them reliably generate text in Afar). Also, open-weight models such as MIXTRAL-8X22B and LLAMA-3-70B often apologize instead of providing an answer or generate text in English when prompted with non-English questions. We observe that the factuality and relevance of LLM-generated culturally specific answers is significantly lower than that of culturally agnostic answers, underscoring the importance of studying culturally specific questions. Factuality and relevance drop considerably on low-resource languages, with GPT-4-TURBO and GPT-4O performing best.

We conduct a human evaluation on a subset of the data (spanning five languages) for the best-performing models. Native speakers rate and rank answers from different LLMs, and an analysis of their annotations reveals that omissions and factuality issues are strong predictors of answer quality ratings. This human evaluation also supports our automatic factuality and relevance evaluations in that culturally agnostic questions are twice as likely to receive higher ratings than culturally specific questions, regardless of the generation model.

Overall, our work establishes a foundation for studying multilingual long-form question answering by releasing CALMQA – the first textual multilingual long-form question answering (LFQA) dataset, with 51.7K questions across 23 languages derived from culturally specific sources.

2 CaLMQA: Cultural Long-form Multilingual Question Answering

Each of the 51.7K examples in CALMQA consists of (1) a *culturally specific* question written in one of 23 languages, (2) an optional human-written English translation (for low-resource languages), and (3) an optional human-written reference answer (for high- and mid-resource languages). We detail CALMQA's collection process and statistics below.

2.1 What questions are culturally specific?

Culture is a multifaceted and abstract concept that eludes a simple definition (Adilazuarda et al., 2024; Liu et al., 2024). We define culturally specific questions as questions that (1) refer to topics, concepts, objects, entities or events that are unique to one or a few cultures, or (2) have different answers depending on the cultural or regional context. Our notion of culturally specific questions is based on Liu et al. (2024): "1) basic concepts that are 'configured' differently, reflecting the cultural- specific way of thinking, and 2) concepts that are unique to a culture"; our definition embeds the former by including questions with answers dependent on culture, and the latter by including questions that refer to concepts related to culture. Liu et al. (2024) taxonomizes cultural NLP works into 10 categories including values, norms and morals, and knowledge; we collect that cultural knowledge in CALMQA.

2.2 Data Collection

We collect our dataset through two processes. For high- and mid-resource languages, we follow prior work (Fan et al., 2019) and collect questions from community Q&A forums. For low resource languages where such web content is scarce, we hire freelancers to write culturally specific questions.

Culturally specific questions for high- and mid-resource languages: Many countries have their own community forums where people can exchange information, similar to Quora, Reddit or StackExchange in English. We collect culturally specific questions from these websites via a crowdsourcing process that we scale with LLM assistance: first, we ask English-proficient Prolific[3] crowdworkers from different countries to provide a link to a community web forum in their language that contains many complex questions that cover a diverse range of topics. Next, we ask workers to collect culturally specific questions and real users' answers from the identified websites, for $0.65-1.33 USD per question. We manually review all provided examples and websites. Our workers yielded 923 questions across 11 languages with answers at a cost of $1427 USD (Table 4, left). Refer to §A.2 for more details.

We scale our question collection process by automating the collection and verification of questions. We obtain around 10k questions for each language. For English, Chinese and Russian, we use existing Hugging Face datasets containing questions scraped from our chosen websites (Gao et al., 2021; Wang, 2023; its5Q, 2022). For the remaining high-resource languages (except Hebrew, for which we were unsuccessful), we implement and utilize website-specific question extraction scripts. We do not collect answers due to the challenges of extracting them. We filter our questions using GPT-4O-MINI, with two model passes that assess each question's cultural specificity and general quality, retaining 52% of questions (prompts in Table 5 and Table 6). We apply these filters on the worker-collected questions too, retaining $>90\%$ of questions. This procedure yielded 50,227 additional questions at a cost of $34 USD.

Culturally specific questions for low-resource languages: Unlike existing LFQA datasets, CALMQA also includes twelve *low-resource languages* (Table 4, right). We choose languages with scarce online resources that are not well-studied in prior work, but for which we can also find at least one annotator bilingual in English. We hire 29 native speakers (1-3 annotators per language, depending on their availability) on Upwork,[4] each of whom receives guidelines, takes a paid ($7 USD) comprehension test, and then writes culturally specific questions with English translations for $0.65-1.00 USD per question. As having them write answers for all of these languages is prohibitively expensive, we collect answers and their English translations only for Kirundi ($2 USD per question, $106 USD total). This process yielded a total of 548 questions with English translations at a cost of $833 USD. The protocol was reviewed and deemed *exempt* by an Institutional Review Board. Please refer to §A.3 for more details.

Quality control: We screened crowdworkers through a qualification task to ensure understand-

[3] https://www.prolific.com/

[4] https://www.upwork.com/

ing of culturally specific, long-form questions. Each question curated by workers was manually reviewed by at least one author; workers provided clarifications or replaced unsuitable questions as needed. In the case of low-resource languages, the questions were also checked by another annotator of that language. See §A.2 for detailed guidelines.

2.3 Dataset Analysis

Table 1 and Table 4 summarize the statistics CALMQA's 51.7K culturally specific questions. We measure the length of questions with bytes (Clark et al., 2020) as token count is not comparable across languages due to different compression rates (Ahia et al., 2023). High- and mid-resource language questions are generally longer than low-resource language questions, except for Arabic and Balochi. This can be largely attributed to different collection method (gathered from community forums vs. manually written by crowdworkers); see Table 9 for examples.

Finally, we categorize CALMQA's questions based on their topic by first manually curating a set of categories and developing GPT-4-TURBO-based pipeline. Figure 1 shows a treemap of the question categories with examples. We find that the distribution of categories of culturally specific questions is similar between different languages. See §B for details.

3 Evaluating LLMs on CALMQA

We evaluate answers from seven state-of-the-art LLMs using automatic metrics for surface quality, relevance and factuality, combining these into a unified metric. We supplement this with human evaluation of LLM answers across five languages.

Models: We evaluate four closed-source LLMs (CLAUDE-3-OPUS, GEMINI-1.5-PRO, GPT-4-TURBO, GPT-4O (Anthropic, 2024; Gemini Team, 2024; OpenAI, 2024a,b) and three open-weights LLMs (AYA-EXPANSE-32B, LLAMA-3-70B, MIXTRAL-8X22B (Dang et al., 2024; AI@Meta, 2024; Mistral AI, 2024). Model details are in Appendix Table 15.

Inference Setting: Each model is prompted with a question from our dataset in a zero-shot setup without instructions. We use greedy decoding and limit outputs to 2048 tokens. The total cost of API calls is $530 USD.[5]

Data: For controlled comparison of LLM performance on questions with and without cultural knowledge requirements, we assemble an evaluation set of 3,644 questions from three sources: (1) all 1471 human-collected culturally specific questions, (2) 100 randomly sampled automatically collected questions per language, and (3) 51 culturally agnostic questions from `r/explainlikeimfive` translated into 22 languages using GPT-4-TURBO, which has demonstrated superior translation performance (Yan et al., 2024; Jiao et al., 2023). For Balochi, Fijian, and Kirundi, where translation quality was poor, we hire native speakers. This subset allows comprehensive evaluation while managing computational costs compared to using our full dataset of 51.7K questions.

3.1 Automatic Evaluation Metrics

Since common QA metrics like BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004a) do not correlate well with human judgement for long-form QA (Xu et al., 2023; Krishna et al., 2021), we (1) identify answers with surface level issues (e.g. incorrect language), (2) measure factuality and relevance of the remaining answers using the VeriScore pipeline of Song et al. (2024) and LLM-as-a-Judge (Zheng et al., 2023) with GPT-4O respectively, and (3) combine our individual measures to produce a single metric of answer quality.

Identifying surface-level issues ($S_{surf} \in \{0, 1\}$): Useful answers must be in the correct language (i.e., the language of the question) and free from word or phrase repetition. We start by detecting answers in the wrong language using a pipeline that combines `polyglot`[6] and `langid`[7], which yields optimal results for most languages (see Table 13 for accuracy). Balochi, Kirundi, Papiamento, and Hiligaynon are excluded due to low language identification accuracy. Then, we identify responses with repetitions by employing `tiktoken`[8] with the `o200_base` encoding and flagging any answers in which a sequence of 20 tokens is repeated four or more times.[9] See §C.1 for details. We assign a score of 1 if there is no surface issue, 0 otherwise.

[5]We note the total cost of calls for each model as follows: GEMINI-1.5-PRO $17 USD, GPT-4O $40 USD, GPT-4-TURBO $80, LLAMA-3-70B and MIXTRAL-8X22B $4 USD, and CLAUDE-3-OPUS $390 USD.

[6]https://pypi.org/project/polyglot/

[7]https://pypi.org/project/py3langid/

[8]https://github.com/openai/tiktoken

[9]GEMINI-1.5-PRO often returned an API error for questions in low-resource languages; we mark such answer as invalid.

Language	# Q	# A	Q. Bytes (avg/std)	A. Bytes (avg/std)
English	2617	78	205.1 / 209.4	674.1 / 475.9
Arabic	5300	85	127.0 / 77.2	2105.0 / 2378.6
Chinese	5901	75	69.0 / 49.2	588.8 / 939.7
German	4091	96	427.8 / 451.9	1169.0 / 744.7
Hebrew	96	96	142.5 / 84.2	2043.6 / 1934.9
Hindi	6404	91	133.7 / 46.5	3618.8 / 1867.1
Hungarian	3843	75	366.0 / 441.8	379.3 / 333.2
Japanese	6466	75	814.2 / 696.7	920.6 / 637.1
Korean	5875	75	248.8 / 198.5	1008.6 / 936.3
Russian	5403	75	291.0 / 487.7	4546.7 / 5067.9
Spanish	5058	102	547.3 / 544.0	852.0 / 817.9
Total	51150	923	152.8 / 140.8	1640.8 / 2291.3

Language	# Q	Q. Bytes (avg/std)	A. Bytes (avg/std)
Afar	25	43.7 / 16.5	N/A
Balochi	65	122.7 / 52.4	N/A
Faroese	30	47.8 / 16.6	N/A
Fijian	75	75.0 / 36.9	N/A
Hiligaynon	65	93.4 / 39.1	N/A
Kirundi	53	64.6 / 21.2	557.2 / 160.9
Papiamento	10	66.8 / 28.5	N/A
Pashto	75	64.8 / 26.9	N/A
Samoan	25	51.2 / 19.3	N/A
Tongan	10	81.2 / 19.2	N/A
Tswana	65	87.2 / 43.4	N/A
Wolof	50	45.3 / 18.9	N/A
Total	548	75.1 / 41.3	557.2 / 160.9

Table 1: Data statistics of high- & mid-resource language (left) and low-resource language (right) culturally specific questions. We report the number of bytes in the UTF-8 encoding as token counts will significantly vary between the languages. For high- & mid-resource languages, answers were only obtained for a subset of questions collected by crowdworkers, due to challenges with extracting and ranking answers automatically. For low-resource languages, we collect answers for Kirundi only. See Table 4 for culturally agnostic questions.

We only evaluate factuality and relevance for answers without surface-level issues.

Evaluating factuality ($S_{fact} \in [0, 1]$): To evaluate factuality of long-form texts, FactScore (Min et al., 2023) verifies automatically extracted claims against retrieved evidence, and recent work expands this to multilingual texts by translating the non-English responses into English (Shafayat et al., 2024). Following this, we translate our questions and answers into English using GPT-4o. Then, we apply the claim extraction and verification pipeline introduced in Song et al. (2024), which improves the robustness of FactScore by focusing exclusively on verifiable, non-trivial claims and using Google Search to obtain evidence.[10][11] Finally, for every valid answer (i.e., answer without surface-level issues), we obtain a list of claims, the corresponding top 10 search results, and faithfulness labels (supported or unsupported); see Figure 7 for more details. The S_{fact}score will be the fraction of claims that are deemed supported, or 0% if there are no verifiable claims.

Evaluating relevance ($S_{rel} \in \{0, 1\}$): LLM prompting has been shown to have reasonable agreement with human annotations in English and multilingual settings (Hada et al., 2023; Hu et al., 2024). Hence, to evaluate the relevance of long-form answers to their questions, we employ LLM-as-a-Judge (Zheng et al., 2023) using GPT-4o. That is, we prompt GPT-4o to decide whether each answer is relevant to its question, using the prompt in Figure 15 with the English translation of the question and answer from our factuality evaluation, at a total cost of $120 USD.

Overall performance: We combine three metrics to measure the overall quality of the general answer. We obtain the overall quality score at the instance level S by multiplying the surface-level quality, factuality and relevance scores ($S = S_{surf} * S_{fact} * S_{rel}$).

3.2 Results of automatic evaluation

Table 2 reports micro-averaged automatic metrics of each model on culturally agnostic and culturally specific sets, respectively.

Answers to culturally agnostic questions are more factual: Generated answers to culturally agnostic questions tend to be more factual (64%–71%) than answers to culturally specific questions (45%–52%).[12] By contrast, surface issues and relevance are relatively consistent between culturally specific and culturally agnostic questions.

Open-weight models perform much worse than closed-weight models in low-resource languages:

[10] We use Google Search via the Serper API at a total cost of $510 USD.

[11] VeriScore's claim extraction and verification open-source models were run on 1xA40 GPU for 48h.

[12] Models generate a similar number of factual claims on average for both culturally specific and culturally agnostic questions, with the former yielding slightly lower mean claim counts (see Figure 8).

	Surface Level			Fine-Grained		Overall
	Wrong Lang. (%) ↓	Repetitions (%) ↓	W/o Issues (%) ↑	Factual Pr. (%) ↑	Relevance (%) ↑	Overall Score S ↑
GPT-4o	2.7 / 1.2	7.5 / 1.4	90.4 / 97.4	69.6 / **52.2**	88.2 / 95.7	**56.9 / 49.2**
GPT-4-Turbo	3.6 / 1.6	3.0 / 0.5	93.4 / 97.9	69.9 / 51.9	85.0 / 94.4	**56.9** / 48.7
Claude-3-Opus	4.3 / 1.2	0.6 / **0.1**	**95.0 / 98.7**	63.6 / 45.5	84.4 / 93.6	52.9 / 42.6
Aya-Expanse-32B	19.8 / 6.7	7.9 / 1.8	73.4 / 91.7	63.8 / 45.6	84.7 / 91.9	43.4 / 39.5
Gemini-1.5-Pro	**0.3***/ **0.3***	**0.1***/ **0.1***	58.2 / 82.9	**71.1** / 48.7	**98.2** / 96.5	40.9 / 46.6
Mixtral-8x22B	33.2 / 11.2	10.7 / 7.7	57.2 / 81.3	64.0 / 46.2	95.5 / 93.0	35.6 / 35.7
Llama-3-70B	76.0 / 70.0	0.5 / 0.6	23.5 / 29.5	66.6 / 46.7	97.4 / **97.5**	15.3 / 13.5

Table 2: Model performance aggregated across languages. Each cell reports values for culturally agnostic / culturally specific portions. Due to language identification errors, we exclude Balochi, Kirundi, Papiamento, and Hiligaynon from the aggregation. Fine-grained metrics are only computed over answers that lack surface-level issues. *Gemini-1.5-Pro returned API errors for 41.4% (agnostic) / 16.7% (specific) of answers, which likely obscures surface-level errors that it makes.

Figure 2 shows the overall scores for each model by language. Open-weight models are comparable to their closed counterparts on high- and mid-resource languages, with Aya-Expanse-32B outperforming Claude-3-Opus in 8 of these languages on culturally agnostic questions. The closed models significantly outperform the open models on the low-resource languages, scoring mostly 22 – 66 while the open models mostly score below 10. This gap is attributed to surface-level issues, which are present in as high as 70% for Llama-3-70B (see Table 2). The exception is Gemini-1.5-Pro, which throwing API errors when prompted in low-resource languages.

We show specific categories of answer deficiencies detected by our surface-level issue metrics in Table 16, and provide further analysis in Appendix C.2.

4 Human Evaluation

Given the limitations of automatic metrics, we supplement our evaluation with native speaker judgments across five languages: Kirundi, Fijian, Hindi, German, and English.

Evaluation setup: We evaluate Claude-3-Opus, GPT-4-Turbo, and Mixtral-8x22B. For each language we sampled 10 culturally specific and 10 culturally agnostic questions.[13]

We recruit native speakers via Prolific and Upwork, all of whom participated in the question collection process, paying $7.50 USD per question and an additional $8.00 USD for reviewing the guidelines, totaling $720 USD. Annotators are presented with a question, reference answer (if available), and answers generated by the three models in random order. For each candidate answer, they are asked to: (1) identify whether it is in the correct language, (2) mark minor and major errors,[14] (3) evaluate factuality, (4) note significant omissions, (5) comment on the answer's overall quality (Figure 3), and (6) rate it on a 5-point scale (excellent, good, average, poor, unusable). Finally, annotators rank the three answers from best to worst and provide a free-form explanation for their ranking. We provide details of the workflow in Figure 17 and §D. The study was reviewed by the Institutional Review Board and received a *non-human subject* determination.

4.1 Results of human evaluation

Looking at the overall answer ratings, human annotators prefer GPT-4-Turbo's answers, followed by Claude-3-Opus's and then lastly Mixtral-8x22B's (Figure 4). To confirm, we fit a cumulative link mixed model (`clmm()`) for predicting ratings from models (Table 18), with annotators nested within language included as a random effect.[15] We find that a Mixtral-8x22B answer has an 88% chance of having a lower rating than a Claude-3-Opus answer (p<.001) and a 94% chance of having a lower rating than a GPT-4-Turbo answer (p<.001). Also, a Claude-3-Opus answer has a 30% chance of having a lower rating than a GPT-4-Turbo answer (p<.001).

[13]For culturally specific questions, annotators selected 10 questions they were confident in answering accurately. For culturally agnostic questions, we supplied annotators with bullet-point answers in English.

[14]This step was included to help the annotators visualize any issues with the answer and encourage them to read the entire answer. Hence, we did not require annotators to classify errors beyond a simple minor vs major distinction.

[15]We use `clmm` from the `ordinal` package (Christensen, 2023) because of the ordinal nature of our response variable (ratings) and repeated measures, with annotators rating each model multiple times for different questions.

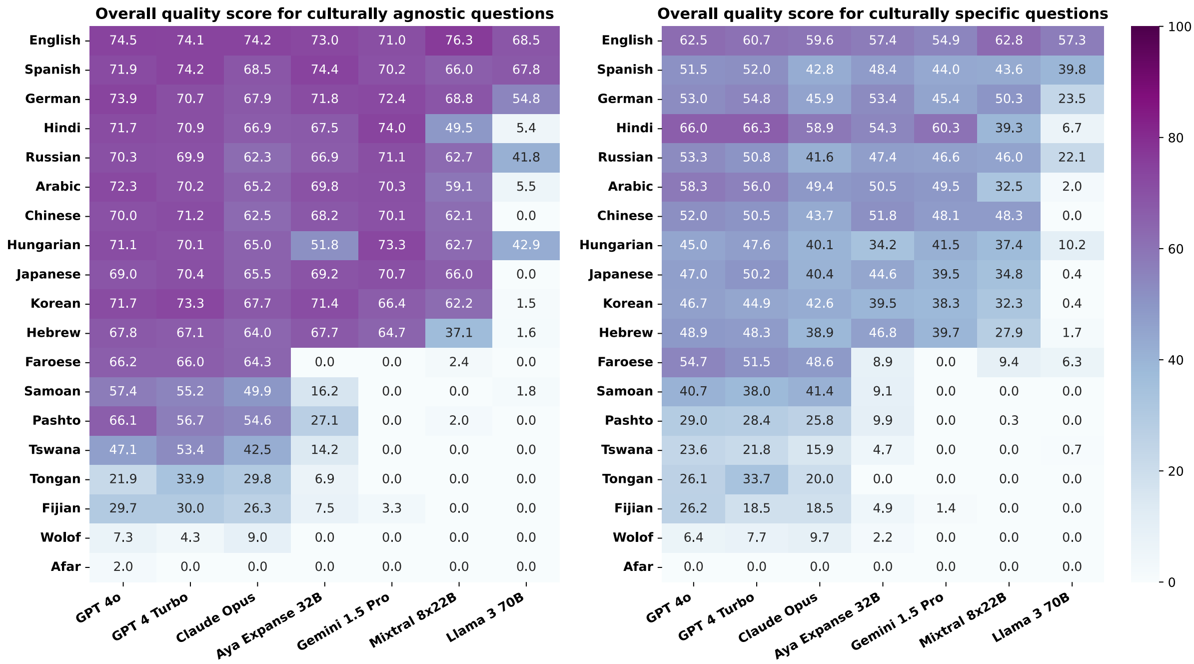

Figure 2: Answer scores S based on our quality criteria: surface issues, factuality and relevance. The left heatmap shows the results for culturally agnostic questions while the right heatmap shows the results for culturally specific questions. Closed- and open-weight models perform comparably on high- to mid- resource languages, while open-weight models are much worse on low-resource languages. Scores degrade on culturally specific questions due to factual imprecision (see Figure 10).

Answer ratings are lower for culturally specific questions: Figure 4 suggests that LLMs generate worse answers for culturally specific questions than for culturally agnostic questions. To check this, we fit a cumulative link mixed model for predicting ratings from question type (Table 20), with annotators nested within language included as a random effect. We see that an answer to a culturally agnostic question has a 67% chance of having a higher rating than an answer to a culturally specific question (p<.001). CLAUDE-3-OPUS's performance drop on culturally specific questions is notable: its answer to a culturally specific question has an 80% chance of receiving a lower rating compared to a culturally agnostic question (p<.001).

Factuality and omission issues are strong predictors of answer rating: To determine which variables of this experiment (e.g., model, question type, factuality issues, omissions) correlate with answer rating, we fit cumulative link mixed models for predicting the rating, with each variable being used as the sole predictor of a separate model. Table 23 shows R^2 values of these models. We observe high marginal R^2 for the factuality issues model (R^2=0.560) and the omissions model (R^2=0.740), indicating that these factors are strong predictors of answer rating.[16]

4.2 Analyzing annotator comments

We analyze annotators' comments to gain insights into answer quality. For each comment field, we iteratively develop and apply an annotation schema, linking the results to the corresponding ratings and scores. (See Appendix §D for details.)

Factuality errors are more frequent for culturally specific answers: All 12 issues regarding incorrect dates, entity and events (e.g., "It is mentioned that Nifty was launched in 1995 but it was actually launched in 1996.") occur in culturally specific answers, likely due to a greater prevalence of dates, entities and events in culturally specific questions about topics like HISTORY than culturally agnostic topics like HEALTH AND WELLNESS.

GPT-4-TURBO answers rank first due to content. We analyze the reasons mentioned for ranking each model's answers as best. Having good CONTENT (e.g. due to being complete; see Table 28 for a description) is a reason for GPT-4-TURBO being chosen as best answer 51% of the time (e.g. *"Answer 1 (*GPT-4-TURBO*) is the perfect answer and*

[16] In case of mixed effects models, *marginal* R^2 refers to the proportion of variance explained by the fixed effects (predictors) alone.

	Question	Generated Answer	Model	Comment
Kirundi	**Kubera iki cazira ko umwana arya mugahinga ku mwaka mushasha?** *Why was it forbidden for kids to eat at their neighbors' house at the new year?*	Mu muco nyarwanda, Kubera iki cazira ko umwana arya mugahinga ku mwaka mushasha bifite impamvu zikurikira:1. Mugahinga ni ibiryo by'abakuru: Mu muco nyarwanda, mugahinga ufatwa nk'ibiryo by'abakuru kubera ko ari ibiryo bikomeye (...)	Claude 3 Opus	The answer is not answering the question. It talks about something unrelated to the question asked. It's really off-topic. It is also written in Kinyarwanda and talks about Rwanda from the very beginning. The first three words: "Mu muco nyarwanda..." are translated as "In Rwandan culture...", which shows that it is purely about Rwanda, not Burundi.
Hindi	**जनरल और ओबीसी में क्या अंतर है?** *What is difference between general and OBC?*	General (G) और Obstetrics and Gynecology (OB/GYN) दोनों अलग-अलग प्रकार के चिकित्सक हैं और उनके बीच कई अंतर हैं। 1. विशेषता: General चिकित्सक सामान्य रोगों और बीमारियों के लिए प्रदर्शित होते हैं, जबकि OB/GYN चिकित्सक स्त्री के योनि संबंधी बीमारियों और गर्भावस्था से संबंधित बीमारियों के लिए प्रदर्शित होते हैं। (...)	Mixtral 8x22B	The entire answer is wrong. The question is to state the difference between General and OBC category of people (based on caste and society status), but the answer is about general and gynecology practitioners (doctors). It is irrelevant to the question. The answer is written partially in Hindi and English. It is not clear. It is fluent. It sounds artificial.
German	**Was ist Überhangmandat? Gegoogelt, aber nichts passendes gefunden. Ist zu schwer erklärt.** *What is an overhang seat? Googled it, but couldn't find anything suitable. It's explained too complicatedly.*	Ein Überhangmandat ist ein Begriff aus dem Wahlrecht, der vor allem in Verhältniswahlsystemen mit Direktkandidaten eine Rolle spielt, wie zum Beispiel in Deutschland bei Bundestagswahlen. Ich versuche es einfach zu erklären: (...)	GPT-4 Turbo	Very helpful, has a good mixture of theory and a well explained example. Complete, clear and fluent. Definitely it would go for "human-like".

Figure 3: Examples of comments on LLM-generated answers written by human annotators.

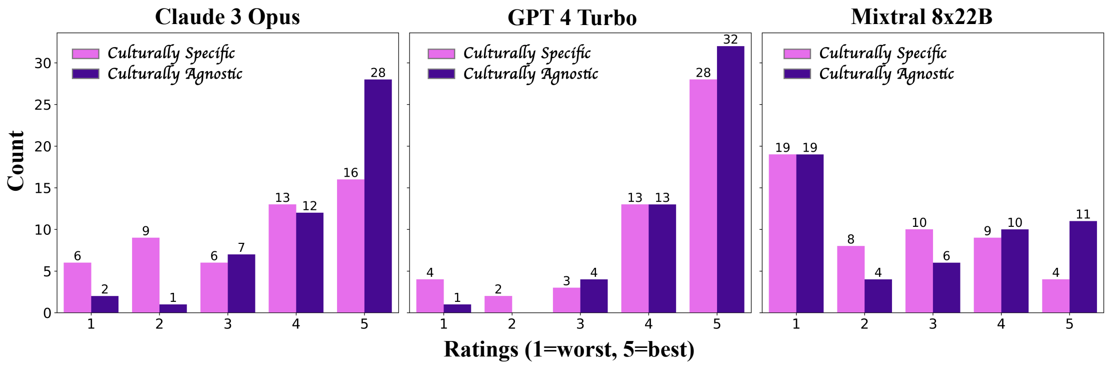

Figure 4: Distribution of human ratings of answers by model and question type. Each model generates 50 answers per question type. Humans give higher ratings for culturally agnostic answers, especially for CLAUDE-3-OPUS.

and explains all the points needed to understand how to play the game 'Teen Patti'."). In the culturally agnostic setting, where CLAUDE-3-OPUS and GPT-4-TURBO perform comparably, more GPT-4-TURBO wins (48%) are attributed to CONTENT than CLAUDE-3-OPUS wins (32%). The full result can be found in §D (Table 24).

5 Related Work

Cultural & Multilingual NLP: Cultural knowledge has been explored through the creation of knowledge bases (Fung et al., 2024; Nguyen et al., 2022) as well as datasets for tasks like probing (Keleg and Magdy, 2023; Yin et al., 2022; Zhou et al., 2024), short-form QA and visual QA. Short-form QA work for multilingual cultural knowledge includes MMLU (Hendrycks et al., 2020) translations or MMLU-style datasets (Singh et al., 2024; Lai et al., 2023; Kim et al., 2024; Koto et al., 2024a), common sense datasets (Myung et al., 2025; Wibowo et al., 2023; Koto et al., 2024b), and evaluations (Shen et al., 2024). Visual long-form QA (LVQA) is less explored and mostly monolingual (Yu et al., 2024; Alwajih et al., 2024), but the contemporaneous work Vayani et al. (2024) looks at LVQA in 100 languages. We are not aware of any textual LFQA datasets of cultural knowledge.

Some multilingual cultural works rely on translation for their multilinguality (Singh et al., 2024), potentially limiting their coverage of cultural concepts. Surveys (Adilazuarda et al., 2024; Liu et al., 2024) call out a lack of multilingual datasets that cover a diverse set of cultural concepts. Our work seeks to make progress in this gap of the literature.

Evaluation of Long-Form QA: Evaluating long-form QA (LFQA) remains challenging. Lexical metrics of text generation like ROUGE (Lin, 2004b) and some neural-based metrics like BERTScore (Zhang et al., 2019) and BLEURT (Sellam et al., 2020) show poor correlation with human ratings (Krishna et al., 2021; Xu et al., 2023; Cambazoglu et al., 2021; Chen et al., 2019). For most other model-based evaluations (Zheng et al., 2023; Fu et al., 2023; Zhong et al., 2022), correlation with human annotations is measured for tasks like instruction-following, summarization and machine translation but mostly not LFQA. Jiang et al. (2023) assess effectiveness of metrics for LFQA, however this is done only on GPT-4-created data.

6 Conclusion & Future Work

We introduce CALMQA, the first textual multilingual long-form QA dataset, which contains 51.7K culturally specific questions across 23 high- and low-resource languages. Our evaluation of seven state-of-the-art LLMs reveals that culturally specific questions are more difficult for models than culturally agnostic ones, evidenced by lower factuality and human ratings. Furthermore, we observe critical surface-level issues (wrong language, repetition) in all models, especially for low-resource languages. Our results stress the importance of diversifying pre- and post-training datasets to emphasize cultural knowledge acquisition, which can help improve culturally specific QA. Also, improving cross-lingual transfer to address data scarcity may help for underrepresented languages like Afar.

Limitations

While we strive to cover as many aspects of the cultures represented in CALMQA as possible, we acknowledge that it is not feasible to encompass every cultural nuance. Additionally, for low-resource languages, we employed workers to manually write questions, which impacts scalability. Finally, our culturally agnostic questions are translations from English performed by GPT-4-TURBO, and thus may not match the quality of human translations.

It would be ideal to have identical distributions of topics across language and type (culturally specific vs culturally agnostic). However, topics like religion, food & drinks, history and literature, among many others, are naturally bound to the culture, making it impossible to have similar distributions for culturally specific and culturally agnostic questions. Moreover, such topics may have different relative significance for different cultures. Consequently, collecting questions representative of the topics important to people conflicts with having identical distributions between languages. Nevertheless, we found that the topic distribution is similar between languages.

Our automatic evaluation relies on surface-level measures such as language detection and token repetitions. While this approach allows us to determine that current LLMs still struggle with producing outputs in the correct language and without repetitions, it does not assess the fluency or completeness of outputs that lack these surface-level issues. This underscores the need for comprehensive metrics to evaluate overall answer quality in multilingual LFQA, which we leave to future work.

We assess factuality of model generated answers by translating them into English and extracting verifiable claims and validating them against evidence retrieved through web searches. However, this evaluation is influenced by three factors: (1) the quality of translation, (2) the quantity of extracted claims and (3) the availability of relevant online evidence. Our relevance evaluation also depends on the quality of translation. While we do not observe any evident issues with our pipelines during data inspection, it is possible then these factors influenced the results.

Our human evaluation uses 100 questions across 5 languages to demonstrate that models struggle to generate well-written, factual, and complete answers in non-English languages. Large-scale human evaluation is time-consuming and prohibitively expensive, and finding workers proficient in low-resource languages presented a significant challenge, constraining our evaluation efforts. However, we have shown that we can statistically justify various insights about LLM multilingual capabilities with our scale of data.

Ethical consideration

The protocols for data collection and human evaluation described in this paper were reviewed and deemed *exempt* by the Institutional Review Board. All annotators provided informed consent for the use and publication of their annotations and collected questions. They were compensated fairly for their work, with their preferred rates respected for both the question collection and evaluation tasks.

Acknowledgments

We extend our gratitude to the annotators from Prolific and Upwork for their hard work and for sharing their expertise about their culture. This project was supported by NSF grant RI-2312948.

References

Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Singh, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. 2024. Towards measuring and modeling "culture" in llms: A survey. *Preprint*, arXiv:2403.15412.

Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Jungo Kasai, David R. Mortensen, Noah A. Smith, and Yulia Tsvetkov. 2023. Do all languages cost the same? tokenization in the era of commercial language models. *ArXiv*, abs/2305.13707.

AI@Meta. 2024. Llama 3 model card.

Badr AlKhamissi, Muhammad ElNokrashy, Mai Alkhamissi, and Mona Diab. 2024. Investigating cultural alignment of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12404–12422, Bangkok, Thailand. Association for Computational Linguistics.

Iñigo Alonso, Maite Oronoz, and Rodrigo Agerri. 2024. Medexpqa: Multilingual benchmarking of large language models for medical question answering. *Preprint*, arXiv:2404.05590.

Fakhraddin Alwajih, El Moatez Billah Nagoudi, Gagan Bhatia, Abdelrahman Mohamed, and Muhammad Abdul-Mageed. 2024. Peacock: A family of Arabic multimodal large language models and benchmarks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12753–12776, Bangkok, Thailand. Association for Computational Linguistics.

Anthropic. 2024. The Claude 3 Model Family: Opus, Sonnet, Haiku. Technical report, Anthropic. Accessed: 2024-05-23.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Berkant Barla Cambazoglu, Valeriia Bolotova-Baranova, Falk Scholer, Mark Sanderson, Leila Tavakoli, and W. Bruce Croft. 2021. Quantifying human-perceived answer utility in non-factoid question answering. *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*.

Anthony Chen, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. Evaluating question answering evaluation. In *Conference on Empirical Methods in Natural Language Processing*.

Rune H. B. Christensen. 2023. *ordinal—Regression Models for Ordinal Data*. R package version 2023.12-4.

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. *Preprint*, arXiv:2003.05002.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, Acyr Locatelli, Hangyu Lin, Dwarak Talupuru, Bharat Venkitesh, David Cairuz, Bowen Yang, Tim Chung, Wei-Yin Ko, Sylvie Shang Shi, Amir Shukayev, Sammie Bae, Aleksandra Piktus, Roman Castagné, Felipe Cruz-Salinas, Eddie Kim, Lucas Crawhall-Stein, Adrien Morisot, Sudip Roy, Phil Blunsom, Ivan Zhang, Aidan Gomez, Nick Frosst, Marzieh Fadaee, Beyza Ermis, Ahmet Üstün, and Sara Hooker. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *Preprint*, arXiv:2412.04261.

Esin Durmus, Karina Nguyen, Thomas I. Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. 2024. Towards measuring the representation of subjective global opinions in language models. *Preprint*, arXiv:2306.16388.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. Eli5: Long form question answering. *Preprint*, arXiv:1907.09190.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. In *North American Chapter of the Association for Computational Linguistics*.

Yi Fung, Ruining Zhao, Jae Doo, Chenkai Sun, and Heng Ji. 2024. Massively multi-cultural knowledge acquisition & lm benchmarking. *Preprint*, arXiv:2402.09369.

Jingsong Gao, Qingren Zhou, and Rui Qiu. 2021. ELI5-Category: a categorized open-domain qa dataset.

Gemini Team. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

Nuno M. Guerreiro, Duarte M. Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André Martins. 2023. Hallucinations in large multilingual translation models. *Transactions of the Association for Computational Linguistics*, 11:1500–1517.

Rishav Hada, Varun Gumma, Adrian de Wynter, Harshita Diddee, Mohamed Ahmed, Monojit Choudhury, Kalika Bali, and Sunayana Sitaram. 2023. Are large language model-based evaluators the solution to scaling up multilingual evaluation? In *Findings*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Xiaodong Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *ArXiv*, abs/2009.03300.

Xinyu Hu, Li Lin, Mingqi Gao, Xunjian Yin, and Xiaojun Wan. 2024. Themis: A reference-free nlg evaluation language model with flexibility and interpretability. In *Conference on Empirical Methods in Natural Language Processing*.

its5Q. 2022. its5q/yandex-q - datasets at hugging face.

Dongfu Jiang, Yishan Li, Ge Zhang, Wenhao Huang, Bill Yuchen Lin, and Wenhu Chen. 2023. Tigerscore: Towards building explainable metric for all text generation tasks. *Trans. Mach. Learn. Res.*, 2024.

Wenxiang Jiao, Wenxuan Wang, Jen tse Huang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023. Is chatgpt a good translator? yes with gpt-4 as the engine. *Preprint*, arXiv:2301.08745.

Haoqiang Kang, Terra Blevins, and Luke Zettlemoyer. 2024. Comparing hallucination detection metrics for multilingual generation. *Preprint*, arXiv:2402.10496.

Amr Keleg and Walid Magdy. 2023. Dlama: A framework for curating culturally diverse facts for probing the knowledge of pretrained language models. *Preprint*, arXiv:2306.05076.

Eunsu Kim, Juyoung Suk, Philhoon Oh, Haneul Yoo, James Thorne, and Alice Oh. 2024. Click: A benchmark dataset of cultural and linguistic intelligence in korean. *ArXiv*, abs/2403.06412.

Fajri Koto, Jey Han Lau, and Timothy Baldwin. 2021. Evaluating the efficacy of summarization evaluation across languages. *ArXiv*, abs/2106.01478.

Fajri Koto, Haonan Li, Sara Shatnawi, Jad Doughman, Abdelrahman Boda Sadallah, Aisha Alraeesi, Khalid Almubarak, Zaid Alyafeai, Neha Sengupta, Shady Shehata, Nizar Habash, Preslav Nakov, and Timothy Baldwin. 2024a. Arabicmmlu: Assessing massive multitask language understanding in arabic. *ArXiv*, abs/2402.12840.

Fajri Koto, Rahmad Mahendra, Nurul Aisyah, and Timothy Baldwin. 2024b. Indoculture: Exploring geographically-influenced cultural commonsense reasoning across eleven indonesian provinces. *ArXiv*, abs/2404.01854.

Kalpesh Krishna, Aurko Roy, and Mohit Iyyer. 2021. Hurdles to progress in long-form question answering. *Preprint*, arXiv:2103.06332.

Viet Dac Lai, Chien Van Nguyen, Nghia Trung Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. 2023. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. *Preprint*, arXiv:2307.16039.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330, Online. Association for Computational Linguistics.

Huihan Li, Liwei Jiang, Nouha Dziri, Xiang Ren, and Yejin Choi. 2024. Culture-gen: Revealing global cultural perception in language models through natural language prompting. *Preprint*, arXiv:2404.10199.

Chin-Yew Lin. 2004a. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Chin-Yew Lin. 2004b. Rouge: A package for automatic evaluation of summaries. In *Annual Meeting of the Association for Computational Linguistics*.

Chen Cecilia Liu, Iryna Gurevych, and Anna Korhonen. 2024. Culturally aware and adapted nlp: A taxonomy and a survey of the state of the art.

Jiahua Liu, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2019. XQA: A cross-lingual open-domain question answering dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2358–2368, Florence, Italy. Association for Computational Linguistics.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *Preprint*, arXiv:2305.14251.

Mistral AI. 2024. Cheaper, Better, Faster, Stronger. Technical report, Mistral AI. Accessed: 2024-06-19.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Afina Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, Víctor Gutiérrez-Basulto, Yazmín Ibáñez-García, Hwaran Lee, Shamsuddeen Hassan Muhammad, Kiwoong Park, Anar Sabuhi Rzayev, Nina White, Seid Muhie Yimam, Mohammad Taher Pilehvar, Nedjma Ousidhoum, Jose Camacho-Collados, and Alice Oh. 2025. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. *Preprint*, arXiv:2406.09948.

Tarek Naous, Michael Ryan, Alan Ritter, and Wei Xu. 2024. Having beer after prayer? measuring cultural bias in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16366–16393, Bangkok, Thailand. Association for Computational Linguistics.

Tuan-Phong Nguyen, Simon Razniewski, Aparna S. Varde, and Gerhard Weikum. 2022. Extracting cultural commonsense knowledge at scale. *Proceedings of the ACM Web Conference 2023*.

OpenAI. 2024a. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

OpenAI. 2024b. Model release blog: GPT-4o. Technical report, OpenAI. Accessed: 2024-05-23.

Shramay Palta and Rachel Rudinger. 2023. FORK: A bite-sized test set for probing culinary cultural biases in commonsense reasoning models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9952–9962, Toronto, Canada. Association for Computational Linguistics.

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.

Yifu Qiu, Yftah Ziser, Anna Korhonen, E. Ponti, and Shay B. Cohen. 2023. Detecting and mitigating hallucinations in multilingual summarisation. *ArXiv*, abs/2305.13632.

Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. Bleurt: Learning robust metrics for text generation. In *Annual Meeting of the Association for Computational Linguistics*.

Sheikh Shafayat, Eunsu Kim, Juhyun Oh, and Alice Oh. 2024. Multi-fact: Assessing factuality of multilingual llms using factscore. *Preprint*, arXiv:2402.18045.

Siqi Shen, Lajanugen Logeswaran, Moontae Lee, Honglak Lee, Soujanya Poria, and Rada Mihalcea. 2024. Understanding the capabilities and limitations of large language models for cultural commonsense. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5668–5680, Mexico City, Mexico. Association for Computational Linguistics.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David I. Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Wei-Yin Ko, Madeline Smith, Antoine Bosselut, Alice Oh, Andre F. T. Martins, Leshem Choshen, Daphne Ippolito, Enzo Ferrante, Marzieh Fadaee, Beyza Ermis, and Sara Hooker. 2024. Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation. *Preprint*, arXiv:2412.03304.

Yixiao Song, Yekyung Kim, and Mohit Iyyer. 2024. Veriscore: Evaluating the factuality of verifiable claims in long-form text generation. *Preprint*, arXiv:2406.19276.

Yan Tao, Olga Viberg, Ryan S. Baker, and Rene F. Kizilcec. 2024. Cultural bias and cultural alignment of large language models. *Preprint*, arXiv:2311.14096.

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020-2022. Label Studio: Data labeling software. Open source software available from https://github.com/heartexlabs/label-studio.

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 Mihaylov, Chao Qin, Abdelrahman M Shaker, Mike Zhang, Mahardika Krisna Ihsani, Amiel 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 Amirudin, Muhammad Ridzuan, Daniya Kareem, Ketan More, Kunyang Li, Pramesh Shakya, Muhammad Saad, Amirpouya Ghasemaghaei, Amirbek Djanibekov, Dilshod Azizov, Branislava Jankovic, Naman Bhatia, Alvaro Cabrera, Johan Obando-Ceron, Olympiah Otieno, Fabian Farestam, Muztoba Rabbani, Sanoojan Baliah, Santosh Sanjeev, Abduragim Shtanchaev, Maheen Fatima, Thao Nguyen, Amrin Kareem, Toluwani Aremu, Nathan Xavier, Amit Bhatkal, Hawau Toyin, Aman Chadha, Hisham Cholakkal, Rao Muhammad Anwer, Michael Felsberg, Jorma Laaksonen, Thamar Solorio, Monojit Choudhury, Ivan Laptev, Mubarak Shah, Salman Khan, and Fahad Khan. 2024. All languages matter: Evaluating lmms on culturally diverse 100 languages. *Preprint*, arXiv:2411.16508.

Ray Wang. 2023. wangrui6/zhihu-kol - datasets at hugging face.

Haryo Akbarianto Wibowo, Erland Hilman Fuadi, Made Nindyatama Nityasya, Radityo Eko Prasojo, and Alham Fikri Aji. 2023. Copal-id: Indonesian language reasoning with local culture and nuances. *ArXiv*, abs/2311.01012.

Fangyuan Xu, Yixiao Song, Mohit Iyyer, and Eunsol Choi. 2023. A critical evaluation of evaluations for long-form question answering. *Preprint*, arXiv:2305.18201.

Jianhao Yan, Pingchuan Yan, Yulong Chen, Judy Li, Xianchao Zhu, and Yue Zhang. 2024. Gpt-4 vs. human translators: A comprehensive evaluation of translation quality across languages, domains, and expertise levels. *Preprint*, arXiv:2407.03658.

Da Yin, Hritik Bansal, Masoud Monajatipoor, Liunian Harold Li, and Kai-Wei Chang. 2022. GeoMLAMA: Geo-diverse commonsense probing on multilingual pre-trained language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2039–2055, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2024. Mm-vet: Evaluating large multimodal models for integrated capabilities. *Preprint*, arXiv:2308.02490.

Chen Zhang, L. F. D'Haro, Chengguang Tang, Ke Shi, Guohua Tang, and Haizhou Li. 2023. xdial-eval: A multilingual open-domain dialogue evaluation benchmark. In *Conference on Empirical Methods in Natural Language Processing*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *ArXiv*, abs/1904.09675.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Peng Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. In *Conference on Empirical Methods in Natural Language Processing*.

Li Zhou, Taelin Karidi, Nicolas Garneau, Yong Cao, Wanlong Liu, Wenyu Chen, and Daniel Hershcovich. 2024. Does mapo tofu contain coffee? probing llms for food-related cultural knowledge. *ArXiv*, abs/2404.06833.

Ethical Considerations

The protocols for data collection and human evaluation described in this paper were reviewed and deemed *exempt* by the Institutional Review Board. All annotators provided informed consent for the use and publication of their annotations and collected questions. They were compensated fairly for their work, with their preferred rates respected for both the question collection and evaluation tasks.

A Data Collection

This appendix provides extra details about the data collection process for CALMQA. §A.1 describes the identification of websites used for data collection. §A.2 outlines the data collection methods for high- and mid-resource languages, and §A.3 details the data collection process for low-resource languages. Table 9 contains example entries from the dataset. Table 4 and Table 10 provide more details on the number of questions and languages included in the dataset.

A.1 Website Survey

We conducted a survey to find websites with non-English cultural questions. The instructions outlined the survey's goal, defined a good website, and specified what constitutes a culturally specific question. Our criteria for a good website included:

- At least 500 answered "good" questions (as defined below). Websites could contain other questions as we could filter them out.
- Most questions and answers should be in a non-English language.
- Questions should cover a diverse range of topics, not just one or two broad areas (e.g., fashion, technology).
- The website should contain culturally specific questions not found on English websites or in English QA datasets.
- The website should have a large community of contributors with many questions answered.

The survey evolved through an iterative process of piloting and refining based on the results.

Survey participants were English-proficient crowdworkers on the Prolific platform (`https://www.prolific.com`), whose native language was not English. The survey took about 10 minutes to complete, and we paid \$10 for valid responses, totaling \$510. We considered a response valid if it showed a good-faith effort, even if the website was of insufficient quality or duplicated in another response. From 51 responses, we obtained 4 websites used for question collection. Some websites were rejected despite having good questions because the proportion of good to bad questions was too low for feasible collection. Remaining websites were identified by the authors. See Table 3 for the full list of websites employed.

A.2 High- and Mid-Resource Culturally Specific Questions

Culturally specific questions in high-resource languages were collected by workers on the Prolific[17] platform from the websites in Table 3. All crowdworkers were English-proficient with their native language matching the language of their allocated websites. Each collector was required to read guidelines, pass a guidelines understanding test and complete a test pilot of 5 questions in order to qualify

[17] `https://www.prolific.com/`

Website	Language	ISO	# Q	URL
Ejaba	Arabic	ar	29	https://www.ejaba.com/
Ujeeb	Arabic	ar	56	https://ujeeb.com/
Zhihu	Chinese	zh	75	https://www.zhihu.com/
Reddit ELI5	English	en	78	https://www.reddit.com/r/explainlikeimfive/
Gutefrage	German	de	96	https://www.gutefrage.net/
Quora	Hebrew	he	96	https://he.quora.com
Let's Diskuss	Hindi	hi	91	https://hi.letsdiskuss.com/
Gyakori kérdések	Hungarian	hu	75	https://www.gyakorikerdesek.hu/
Yahoo Japan	Japanese	ja	17	https://chiebukuro.yahoo.co.jp/
OKWave	Japanese	ja	58	https://okwave.jp/
Naver	Korean	ko	75	https://kin.naver.com/qna/
Yandex	Russian	ru	75	https://yandex.ru/q/
Todoexpertos	Spanish	es	102	https://www.todoexpertos.com/

Table 3: Websites from which cultural questions were obtained, with the number of questions retrieved by website. Multiple websites were used for a given language if workers were struggling with a given website.

for the main task. This protocol was reviewed by the Institutional Review Board. Overall, our process yielded 923 questions across 11 languages with answers at a cost of $1427 USD.

Guidelines We provided a guidelines slideshow detailing the rules for selecting questions. The main rules for questions where:

1. The question should require long answer.
2. The question should be culture specific.
3. A native speaker would ask this [question].
4. The question should be objective.
5. Questions should not need pictures/links.

Guidelines Understanding Test Our guidelines understanding test consisted of a form consisting of 11 multiple-selection multiple-choice graded questions. The first question assessed question was "Which of these are listed as **important rules for questions** in the guidelines? (you should select all correct answers)", which required showing understanding of long-form culturally specific information-seeking questions. The remaining 10 questions were curated examples of questions that each may or may not have had issues. Test takers were required to select all the reasons why a question was not suitable according to the guidelines, or select that the question was suitable. We reviewed test results manually, and accordingly chose which workers to pass. We provided passing workers with the test answers, so that they could learn from their mistakes. We paid workers $3.33 USD for completing the test.

Main Collection Task We asked workers to provide examples of culturally specific questions and real users' answers from the identified websites. We manually reviewed all provided examples, using Google Translate to get English translations of website content. In cases where we deemed that an example did not meet our guidelines, we provided feedback and the worker either clarified how their example met the guidelines or replaced the example. For the final dataset, we used GPT-4-TURBO with the prompt in Table 7 instead of Google Translate to obtain the English translations of questions. We translated answers using GPT-4O, which was released after we had conducted our human eval-

LANGUAGE	CULTURALLY SPECIFIC			CULTURALLY AGNOSTIC		
	# Q	Q. BYTES (AVG/STD)	A. BYTES (AVG/STD)	# Q	Q. BYTES (AVG/STD)	A. BYTES (AVG/STD)
HIGH- & MID-RESOURCE LANGUAGES						
English	78	275.7 / 189.0	674.1 / 475.9	51	67.1 / 31.7	632.3 / 636.9
Arabic	85	74.3 / 61.3	2105.0 / 2378.6	51	108.7 / 56.4	N/A
Chinese	75	193.4 / 329.5	588.8 / 939.7	51	68.1 / 31.4	N/A
German	96	304.6 / 227.4	1169.0 / 744.7	51	82.2 / 39.8	N/A
Hebrew	96	142.5 / 84.2	2043.6 / 1934.9	51	93.0 / 42.9	N/A
Hindi	91	122.4 / 52.8	3618.8 / 1867.1	51	184.2 / 90.3	N/A
Hungarian	75	301.1 / 279.8	379.3 / 333.2	51	82.3 / 38.2	N/A
Japanese	75	512.0 / 359.3	920.6 / 637.1	51	104.3 / 50.6	N/A
Korean	75	126.3 / 138.7	1008.6 / 936.3	51	93.0 / 43.3	N/A
Russian	75	310.3 / 438.3	4546.7 / 5067.9	51	134.6 / 70.8	N/A
Spanish	102	429.9 / 271.1	852.0 / 817.9	51	83.6 / 36.1	N/A
LOW-RESOURCE LANGUAGES						
Afar	25	43.7 / 16.5	N/A	51	81.1 / 39.8	N/A
Balochi	65	122.7 / 52.4	N/A	51	96.1 / 48.5	N/A
Faroese	30	47.8 / 16.6	N/A	51	75.1 / 34.5	N/A
Fijian	75	75.0 / 36.9	N/A	51	92.5 / 40.6	N/A
Hiligaynon	65	93.4 / 39.1	N/A	51	83.6 / 39.7	N/A
Kirundi	53	64.6 / 21.2	557.2 / 160.9	51	88.2 / 43.1	N/A
Papiamento	10	66.8 / 28.5	N/A	51	74.1 / 35.3	N/A
Pashto	75	64.8 / 26.9	N/A	51	118.1 / 55.6	N/A
Samoan	25	51.2 / 19.3	N/A	51	80.5 / 37.6	N/A
Tongan	10	81.2 / 19.2	N/A	51	102.4 / 47.9	N/A
Tswana	65	87.2 / 43.4	N/A	51	88.8 / 43.4	N/A
Wolof	50	45.3 / 18.9	N/A	51	78.2 / 44.1	N/A

Table 4: Combined data statistics for culturally specific and culturally agnostic questions. For each language, we report the number of questions (# Q), average and standard deviation of question bytes (Q. Bytes) and answer bytes (A. Bytes) in UTF-8 encoding. Answer bytes for culturally agnostic questions are not available, and are marked as N/A.

uation, with the prompt in Table 8. We paid the workers $0.65-1.33 USD per question.

A.3 Low-Resource Culturally Specific Questions

Questions for low-resource languages were collected by hiring native speakers proficient in English through Upwork. They were paid $0.65 to $1.00 USD per submitted question with its English translation. Annotators were required to read the guidelines and complete a short comprehension task, for which they were paid $7 USD. Additionally, answers to all Kirundi questions were paid $2 USD per answer. This protocol was reviewed by the Institutional Review Board.

Annotators were instructed to write up to 25 questions in their native language along with English translations, ensuring the questions met the following criteria:

- The question requires a long-form answer (at least 3-4 sentences).
- The question is culturally specific, meaning it is more likely to be asked in the region where the language is spoken.
- The question is something a native speaker of the language might ask.
- The question has an objective answer (i.e., not based on opinions).

B Question Categorization

In this section we describe the process of categorizing all questions into predefined set of categories.

Method We selected 25 random culturally specific questions from the dataset. We manually created a list of broad categories with descriptions and examples, and then 2 authors independently applied the categorization on the 25 questions. We reviewed disagreements and accordingly refined the categories. Then we used GPT-4O to categorize using the prompts in Table 11, with temperature set to

You are to evaluate whether a given question is *culturally specific* to *[language]*. A question is considered *culturally specific* to a language if it meets both of the following conditions:
1. The question is written in *[language]*.
2. **At least one** of the following applies:
- The question refers to topics, concepts, objects, entities, or events that are unique to one or a few cultures associated with speakers of *[language]*.
- The answer to the question may vary depending on the cultural or regional context—meaning that the answer would be specific to the *[language]* culture and may differ when considered in the context of other cultures.

Culturally specific topics may include, but are not limited to, areas such as heritage, current events, history, religion, literature, society, travel, cuisine, fauna, economy, education, law, and the arts. **Note:** A question being written in *[language]* does not automatically make it culturally specific.

Example 1 (Culturally Specific to English):
- *Question:* "Why does Eminem give so much praise to Dr Dre? He talks about him like a father in his music."
- *Explanation:* This question refers to Eminem and Dr Dre, who are prominent figures in the American music industry—a distinct aspect of American (and related Western) culture. The question's cultural context makes it culturally specific to English.

Example 2 (Culturally Specific to English):
- *Question:* "How do you obtain a gun license?"
- *Explanation:* The process for obtaining a gun license differs significantly between countries. In the USA, the process is well-established and culturally significant, while it may differ or be non-existent in other countries. Thus, the question is culturally specific to English when referring to American cultural context.

Example 3 (Not Culturally Specific to English):
- *Question:* "Do ships cause the ocean to be higher than it normally would be?"
- *Explanation:* This question is general and universal; its answer does not depend on any cultural or regional context. Therefore, it is not culturally specific.

—

Task:
Assess whether the following question is culturally specific to *[language]* or not. Provide your explanation in English, wrapping it in `<explanation></explanation>` tags. Then, output `<result>PASS</result>` if the question is culturally specific, or `<result>FAIL</result>` if it is not. Your response should contain **only** these two tags and nothing else.

Here is the question to assess:
<question>
[question]
</question>

Table 5: Prompt used with GPT-4O-MINI to filter questions collected from community QA websites for culturally specific questions. Strings in the form *[form]* are placeholders that are replaced at runtime.

You are evaluating questions in a dataset for quality. Your task is to determine whether a given question meets the following quality criteria:

1. **Language:** The question is written in *[language]*.
2. **Long-form Answer:** The question cannot be answered with just a short phrase or entity; it requires a few sentences to answer.
3. **Not Asking For Answerer's Opinions:** The question does not ask for opinions, personal experiences, perspectives or recommendations of the **answerer**. A question that exbihits bias or implies a particular view point of the **asker** does **not** violate this criterion.
4. **Public Information:** Answering the question does not require access to non-public information.
5. **Privacy:** The question does not contain any personally identifiable information (e.g., name, username, phone number, or home address).

Example 1 (Satisfies Criteria):
- *Question:* "Why is Norton hated so much? What makes an antivirus/antimalware program good or bad anyway?"
- *Explanation:* 1. The question is in English. 2. The question requires a explanation comprising of multiple sentences. 3. The question does not ask for an opinion, even though it indicates a negative viewpoint towards Norton. 4. Answering the question does not require access to non-public information. 5. The question does not contain any personally identifiable information. The question meets all the criteria and so is satisfactory.

Example 2 (Does Not Satisfy Criteria):
- *Question:* "How would you suggest I revise mathematics before my first economics class?"
- *Explanation:* The question is explicitly asking for a recommendation and so does not meet the quality critera.

—

Task: Assess whether the following question satisfies all of the quality criteria listed above. Provide a detailed explanation of your assessment in English, wrapped in `<explanation></explanation>` tags. Then, output `<result>PASS</result>` if the question satisfies the quality criteria, or `<result>FAIL</result>` if it does not. Do not output anything outside of the `<explanation></explanation>` and `<result></result>` tags.

Here is the question to assess:
<question>
[question]
</question>

Table 6: Prompt used with GPT-4O-MINI to filter questions collected from community QA websites based on general quality criteria. Strings in the form *[form]* are placeholders that are replaced at runtime.

Your task is to translate a question from *[language]* into English. You will be given the *[language]* answer as the context.

Here is the *[language]* answer. Use it as the context to make the translation sound natural in the English:
[answer]

Translate the following question from *[language]* into English. Make it sound as natural as possible:
[question]

Table 7: Prompt used with GPT-4-TURBO to translate non-English questions into English. Strings in the form *[form]* are placeholders that are replaced at runtime.

Your task is to translate the answer of a *[language]* question from *[language]* into English. You will be given the *[language]* question as the context.

Here is the *[language]* question. Use it as the context to make the translation sound natural in the English: *[question]*

Translate the following answer from *[language]* into English. Make it sound as natural as possible: *[answer]*

Table 8: Prompt used with GPT-4o to translate non-English answers into English. Strings in the form *[form]* are placeholders that are replaced at runtime.

0.0. After minor clarifications to category descriptions, we found that GPT-4o produced adequate categories for all 25 questions. We consequently used the model to categorize all of CaLMQA. Our final categories, with descriptions and examples, can be found in Table 12.

Results Figure 5 shows the number of questions by category and language. We observe that one of Religion, Beliefs, Customs, and Traditions, Governance and Society and History is the top category of almost every language (the exceptions being English and Korean). Furthermore, Religion, Beliefs, Customs, and Traditions is the predominantly the top category for low-resource languages (10/12). This difference is likely due to the question collection process for low-resource languages.

To compare the distribution of categories between languages, we compute pairwise Bhattacharyya coefficients between the data from the languages (Figure 6). The Bhattacharyya coefficient ranges from 0 to 1 with a higher number meaning similar distributions. We see generally high coefficients, indicating that the category distributions are similar between languages.

C Automatic Evaluation

In this section of the appendix we present the details of automatic evaluation. All evaluated models are listed in Table 15. Examples of the model tendencies detected by automatic evaluation are in Table 16.

C.1 Method Details

Language accuracy Figure 13 displays the percentage of responses each model generated in the correct language, independent of correctness or fluency of the answer. We used `polyglot` (https://pypi.org/project/polyglot/) and `langid` (https://pypi.org/project/py3langid/) for language identification, choosing them based on their performance for specific languages. This identification was also applied to the questions to estimate its performance across languages. Our pipeline accurately recognized 100% of instances in 14 languages. For other languages, accuracy typically remained above 90%, with Fijian at 98.67%, Russian at 97.33%, Tongan at 96.92%, Samoan at 92.00%, and Wolof at 90% (see Table 13). However, identification accuracy for Kirundi was notably lower at 35.85%, as the libraries frequently misclassified it as the closely related Kinyarwanda. The automatic identification process failed entirely for Balochi, Hiligaynon, and Papiamento, which is reflected in seemingly low performance for these languages across all the models.

Repetitions Figure 14 illustrates the percentage of responses affected by repetitions, analyzed by language across different models. To identify these repetitions, we employed `tiktoken` (https://github.com/openai/tiktoken) with the `o200_base` encoding. We specifically identified instances where at least 20 consecutive tokens were repeated at least four times within an answer.

Claim extraction and verification pipeline We first translated the answers into English with GPT-4o. Then we extract claims using a finetuned Mistral 7B model and use them to query Serper API for evidence. Then we prompt a finetuned Mistral 7B model for verification. Both models were introduced in Song et al. (2024). The pipeline is visualized in Figure 7.

We report the mean claim count by model, language of the question, and question type in Figure 8. We exclude all answers with surface-level issues as

FIELD	CONTENT
Language	English
Question (Original)	Why does the President of the United States need to be born in the United States to be eligible to run? It seems like the country that a person was born in has little to do with their abilities to lead.
Question (English)	Why does the President of the United States need to be born in the United States to be eligible to run? It seems like the country that a person was born in has little to do with their abilities to lead.
Culturally Specific	True
Answer	It's not a matter of their leadership ability, it's intended to guarantee loyalty. The idea is that most people are loyal to their birth country, so you'd want someone born in the US to be the one acting as commander in chief of the US armed forces, among other things. May not be perfect logic, but it's tradition at this point. Nobody has made a big enough deal of it accompanied by a strong enough argument to get it changed.
Language	Hungarian
Question (Original)	Hogyan lehet kikeverni a Horthy-kori sisakok színét?
Question (English)	How can you mix the color of helmets from the Horthy era? (translated by GPT-4-TURBO)
Culturally Specific	True
Answer	M35 tábori sisak. Neten rákeresve találsz róla képeket. Nem kell megijedni a sok árnyalattól, annak idején sem volt tökéletesen egységes. Ez egy zöldesbarna szín. Talán a RAL6025 áll a legközelebb hozzá. Festékboltban kikeverik géppel. Nem drágább, mint külön megvenni hozzá egy egy dobozzal a festéket és kevergetni. Ecsettel festették az eredetit.
Language	Fijian
Question (Original)	Na cava na vuna era vinakata kina na Nasi ni veiyanuyanu mera sa lesi i Viti Levu?
Question (English)	What is the reason the nurses from the outer islands want to be assigned to Viti Levu?
Culturally Specific	True
Answer	N/A
Language	Kirundi
Question (Original)	Ni kubera iki twama dukeneye gushira ama aprikasiyo ku gihe? Hoba iki iyo tutabikoze?
Question (English)	Why do we need to constantly do software updates? What happens if I don't?
Culturally Specific	False
Answer	N/A

Table 9: Examples of entries in CALMQA. Metadata like questions source (specific website or annotator) are omitted here for simplicity.

LANGUAGE	ISO	FAMILY	Branch	MORPHOLOGY	ORDER	SCRIPT	Region	SPEAKERS
				High- & Mid-Resource				
Arabic	ar	Afro-Asiatic	Semitic	fusional	SVO	Arabic alphabet	Arab world	720M
Chinese	zh	Sino-Tibetan	Sinitic	analytic	SVO	Hanzi	Mainland China, Taiwan, Singapore	1.38B
English	en	Indo-European	Germanic	analytic	SVO	Latin	World-wide	1.5B
German	de	Indo-European	Germanic	fusional	SVO	Latin	Germany, Austria, Switzerland, etc.	133M
Hebrew	he	Afro-Asiatic	Semitic	fusional	SVO	Hebrew script	Israel	9.3M
Hindi	hi	Indo-European	Indo-Iranian	fusional	SOV	Devanagari	India	610M
Hungarian	hu	Uralic	Finno-Ugric	agglutinative	SVO	Latin	Hungary	13M
Japanese	ja	Japonic	Japanese	agglutinative	SOV	Kanji, Kana	Japan	123M
Korean	ko	Koreanic	Korean	agglutinative	SOV	Hangul	Korea	82M
Russian	ru	Indo-European	Balto-Slavic	fusional	SVO	Cyrillic	Russia, Russian-speaking world	255M
Spanish	es	Indo-European	Italic	fusional	SVO	Latin	Spain, Central and South Americas, the US	559M
				Low-Resource				
Afar	aa	Afro-Asiatic	Cushitic	agglutinative	SOV	Latin	Ethiopia, Djibouti, Eritrea	2.6M
Balochi	bal	Indo-European	Indo-Iranian	agglutinative	SOV	Balochi Standard Alphabet	Pakistan, Iran, Afghanistan	8.8M
Faroese	fo	Indo-European	Germanic	fusional	SVO	Latin	Faroe Islands, Denmark	69K
Fijian	fj	Austronesian	Malayo-Polynesian	agglutinative	VOS	Latin	Fiji	640K
Hiligaynon	hil	Austronesian	Malayo-Polynesian	analytic	VSO	Latin	Philippines	9.1M
Kirundi	rn	Niger-Kongo	Atlantic–Congo	agglutinative	SVO	Latin	Burundi	12-13M
Papiamento	pap	Portuguese-based creole	Afro-Portuguese	analytic	SVO	Latin	Aruba, Curaçao, Bonaire	300K
Pashto	ps	Indo-European	Indo-Iranian	fusional	SOV	Pashto alphabet	Afghanistan, Pakistan and Iran	58.8M
Samoan	sm	Austronesian	Malayo-Polynesian	analytic	VSO	Latin	Samoa	510K
Tongan	to	Austronesian	Polynesian	agglutinative	VSO	Latin	Tonga	187K
Tswana	tn	Niger-Kongo	Atlantic–Congo	agglutinative	SVO	Latin	Botswana, South Africa, Zimbabwe	13.9M
Wolof	wo	Niger-Kongo	Atlantic–Congo	agglutinative	SVO	Latin primarily	Senegal	12.3M

Table 10: Linguistic and usage information of the languages in the CALMQA dataset

LANGUAGE	PROMPT
English	You are categorizing questions about different cultures into specific categories. Your task is to assign one category to each question. Here are the available categories: education and career – *[Education and Career Description]*. Example: *[Education and Career Example]* *[categories]...* health and wellness – *[Health and Wellness Description]*. Example: *[Health and Wellness Example]* other – *[Other Description]* Here is the question to categorize: <original_question>*[question]*</original_question> Categorize this question into one of the categories. Output your choice in the following format: <category>category name</category> Your choice:
All except English	You are categorizing questions about different cultures into specific categories. Your task is to assign one category to each question. Here are the available categories: education and career – *[Education and Career Description]*. Example: *[Education and Career Example]* *[categories]...* health and wellness – *[Health and Wellness Description]*. Example: *[Health and Wellness Example]* other – *[Other Description]* Here is the question to categorize: <original_question>*[question]*</original_question> <translation>*[translation]*</translation> Categorize this question into one of the categories. Output your choice in the following format: <category>category name</category> Your choice:

Table 11: Prompts used with GPT-4o to categorize questions. Strings in the form *[form]* are placeholders that are replaced at runtime. The categories used are in Table 12.

CATEGORY	DESCRIPTION	EXAMPLE
EDUCATION AND CAREER	Questions related to school, education system, jobs and career paths. Includes developing new skills for new jobs.	*Why do young children drop out of school?*
GOVERNANCE AND SOCIETY	Questions about laws, governance and policies, as well as politics and social issues.	*What are the reasons why Japan cannot have casinos?*
GEOGRAPHY, TOURISM, AND CLIMATE	Questions concerning the geography, climatic conditions, environmental factors of a region, tourism and travelling.	*What is the significance of Gorée Island?*
TECHNOLOGY	Questions about the technology, technological advancements, uses of technology and digital innovation.	*Are stores that accept VISA debit cards marked VISA? Or is it a store with a VISA PLUS mark?*
ECONOMY AND INDUSTRY	Questions regarding modern-day economic practices, key industries, trade, and economic development.	*Why is our country not developing like others?*
MEDIA AND ENTERTAINMENT	Questions about the media and entertainment specific to the region.	*Why are trademarks obscured on broadcast?*
FOOD AND DRINKS	Questions related to culinary traditions, typical foods and beverages, preparation methods, culinary practices, and cultural significance of meals.	*Why is bread with sausage called a sandwich in Russia?*
HISTORY	Questions about historical events, significant figures, and important periods that have shaped a culture.	*How did the Bujumbura market burn?*
LANGUAGE, ART AND LITERATURE	Questions about the language, dialects, as well as art forms and literary works/traditions.	*Why is the Balochi Language categorized into its three main dialects?*
RELIGION, BELIEFS, CUSTOMS, AND TRADITIONS	Questions regarding religious practices, beliefs, rituals, customs, traditions, and holiday.	*In the Islamic religious teaching, what's the meaning of the seven tens?*
HEALTH AND WELLNESS	Questions related to traditional and modern health practices, public health issues, and well-being.	*Why methadone? What makes it "better" than other opioids for maintenance therapy or tapering off another drug?*
OTHER	Questions that do not fit neatly into the above categories.	

Table 12: Categories of questions in CALMQA.

Figure 5: Number of human-collected questions by category and language.

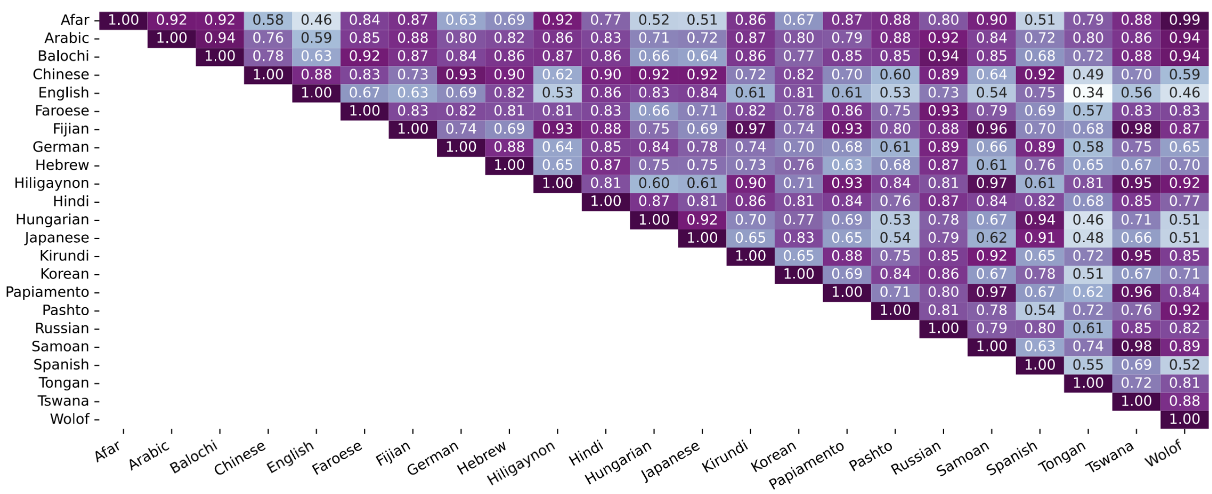

Figure 6: Bhattacharyya coefficients of the category distributions, pairwise between languages. The Bhattacharyya coefficient ranges from 0 to 1, with a higher number meaning more similar distributions.

Language	Lang ID	Accuracy (%)
Afar	aa	100.00
Arabic	ar	100.00
Balochi	bal	0.00
Chinese	zh	100.00
English	en	100.00
Faroese	fo	100.00
Fijian	fj	98.67
German	de	100.00
Hebrew	he	100.00
Hiligaynon	hil	0.00
Hindi	hi	100.00
Hungarian	hu	100.00
Japanese	ja	100.00
Kirundi	rn	35.85
Korean	ko	100.00
Papiamento	pap	0.00
Pashto	ps	100.00
Russian	ru	97.33
Samoan	sm	92.00
Spanish	es	100.00
Tongan	to	100.00
Tswana	tn	96.92
Wolof	wo	90.00

Table 13: Accuracy of the language detection pipeline on the test set made from questions in the given language. Note that the language detection libraries are often more accurate on longer texts (i.e., texts longer than the length of a single question).

well as languages for which the model produced less than 50% of valid answers (i.e., answers without identified surface level issues).

C.2 Further Analysis

Model surface-level errors We further analyzed the responses for specific textual indicators. Detected patterns in model responses are presented with examples in Table 16.

Our textual analysis demonstrates issues in Mixtral-8x22B responses for low-resource languages. 31.47% of Mixtral-8x22B responses to questions in low-resource languages contain phrases like "sorry", "apologize" or "understand" (e.g., *"I'm sorry for any confusion, but it seems you're using a language that I'm not currently able to understand or translate."*). Mixtral-8x22B responses to questions in high-resource languages do not contain these apology-related keywords, revealing an inability to answer the question specifically in low-resource languages. The apologetic textual markers were seen in less than 1% of other models' responses except for Llama-3-70B's, where they were present in 14.74% of low-resource and 10.48% of high-resource language answers.

Textual indicators also uncover deficiencies in Llama-3-70B responses. Notably, 37.87% of responses from Llama-3-70B explicitly mention the English name of the language (e.g., *"I see you're speaking in Balochi!"*), indicating that although the system recognizes the language of the question, it nonetheless responds in English. This is in contrast to Mixtral-8x22B, which does so in 7.21% of responses, GPT-4-Turbo at 1.84%, and less than 1% for other models. Additionally, approximately 19.71% of Llama-3-70B responses include terms like "translate" or "translation" (e.g., *"I apologize, but I'm having trouble understanding your question. Could you please rephrase or translate your question into a language I can understand, such as English?"*), where the system either declines to answer (with or without apology), requests an English translation, or provides a translation itself. In comparison, 8.43% of Mixtral-8x22B responses

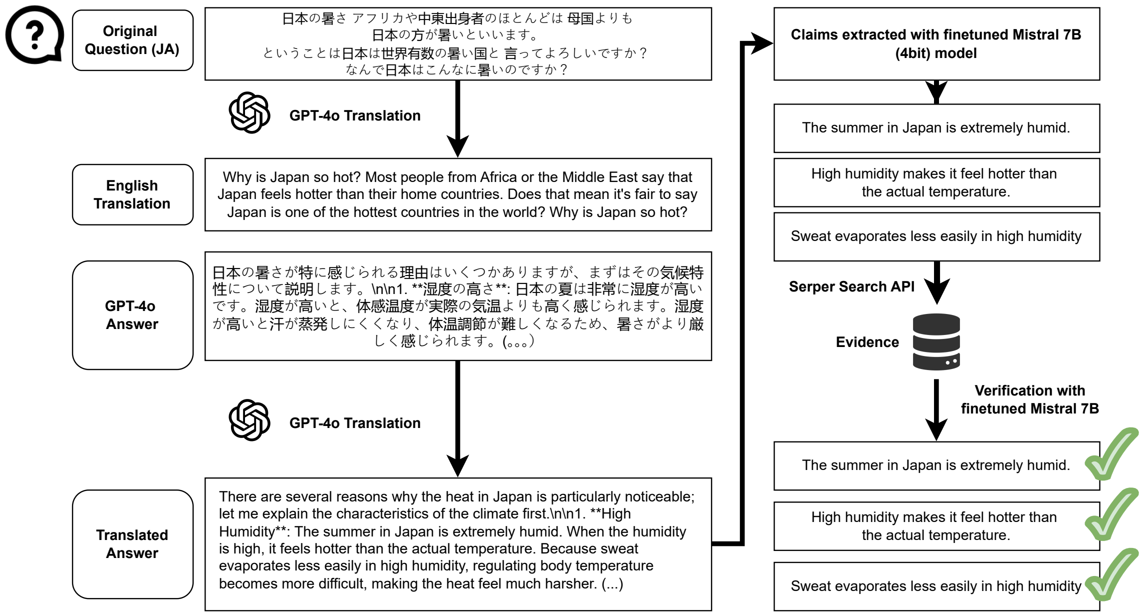

Figure 7: Claim extraction and verification pipeline. Example showing extraction and verification of claims for a question and answer in Japanese. English translations were obtained with GPT-4o. Only part of the answer is provided for readability.

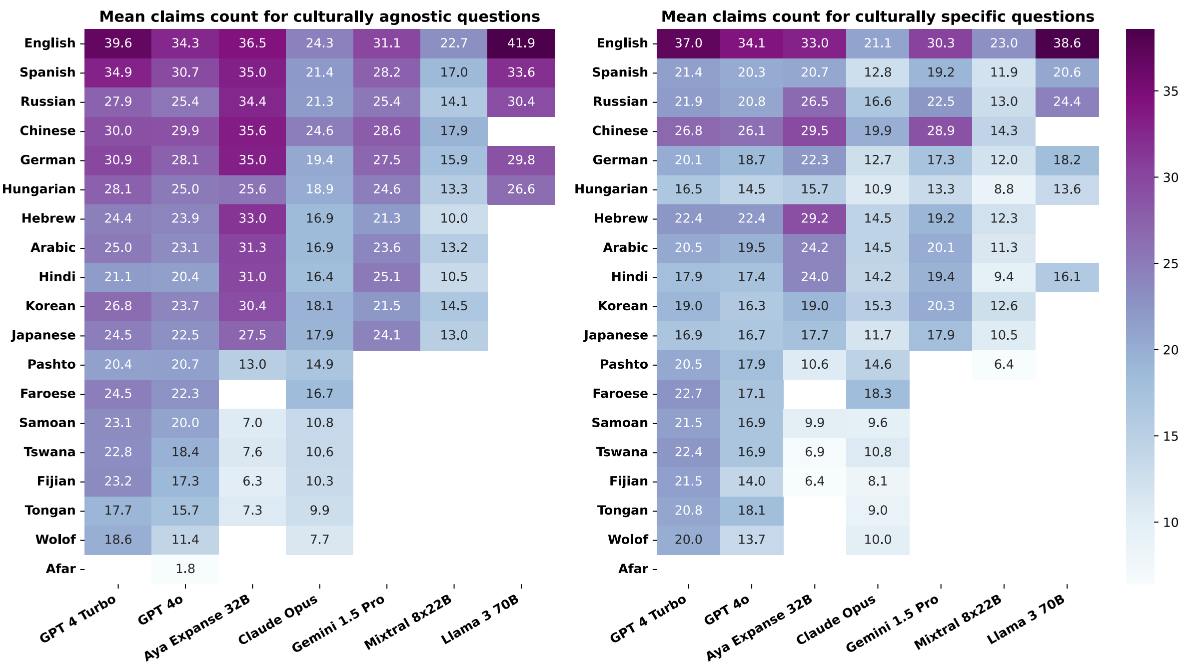

Figure 8: Mean claim count for answers without surface-level issues. The left heatmap shows the results for culturally agnostic questions while the right heatmap shows the results for culturally specific questions. Only languages where at least 10 answers were free from surface-level issues are included.

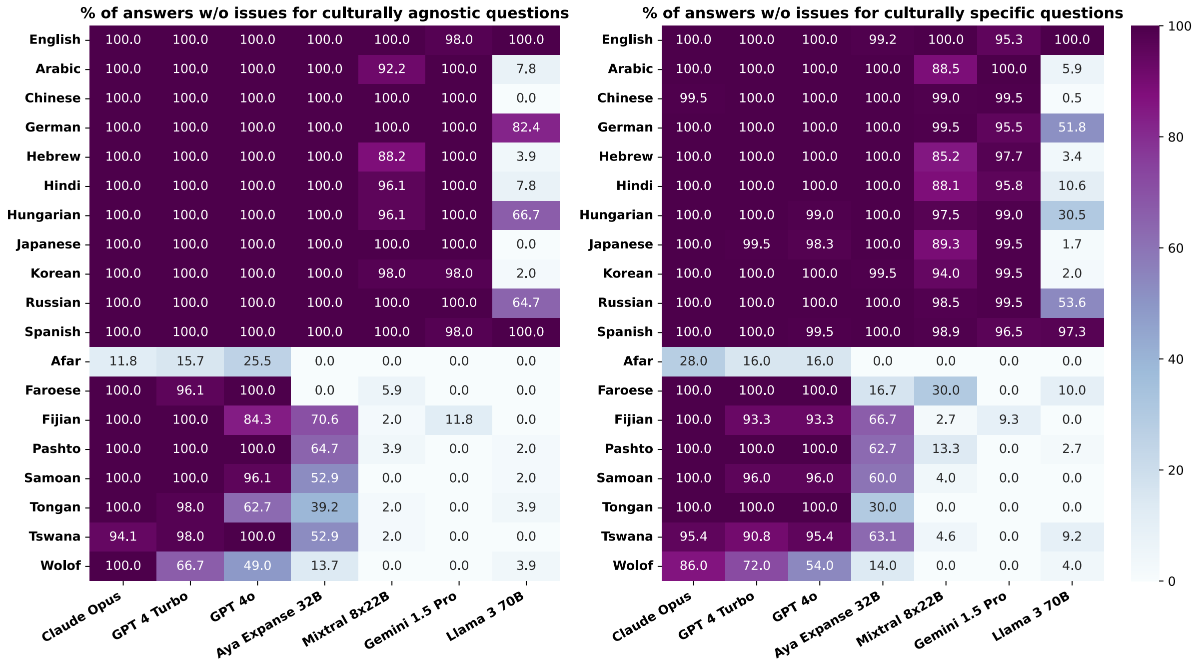

Figure 9: Percentage of model answers without surface issues per language. The left heatmap shows the results for culturally agnostic questions while the right heatmap shows the results for culturally specific questions.

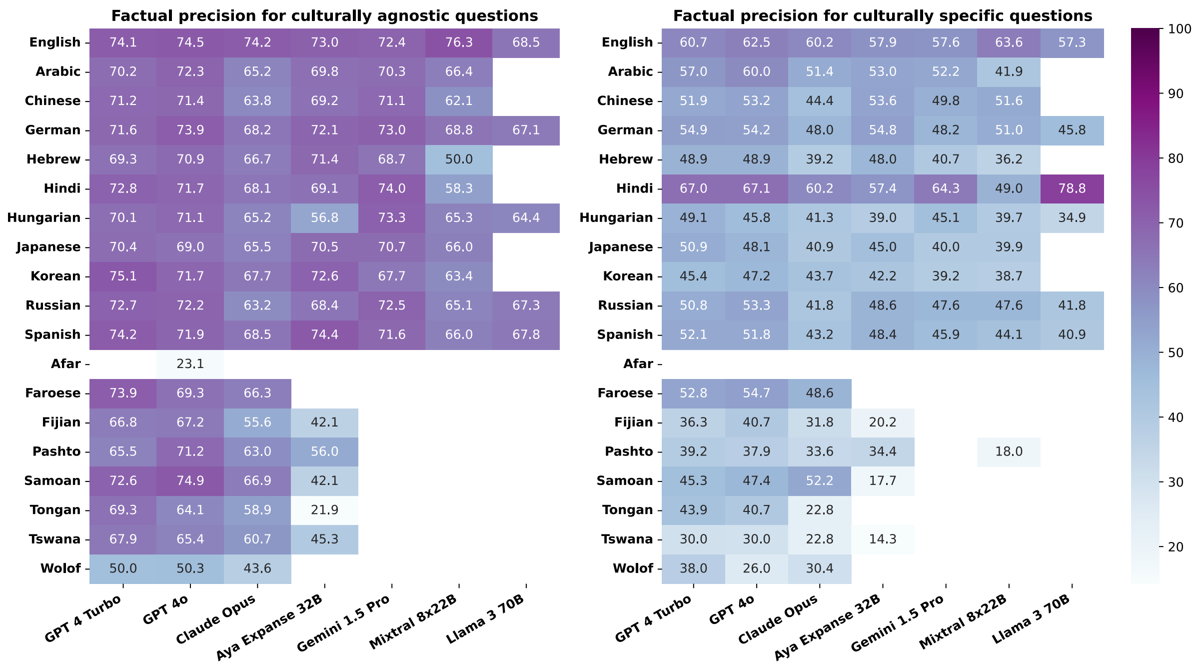

Figure 10: Factual precision for answers without surface-level issues. The left heatmap shows the results for culturally agnostic questions while the right heatmap shows the results for culturally specific questions. We remove model-language combinations for which there are not at least 10 answers without surface-level issues. Factual precision degrades on culturally specific questions, especially for low-resource languages.

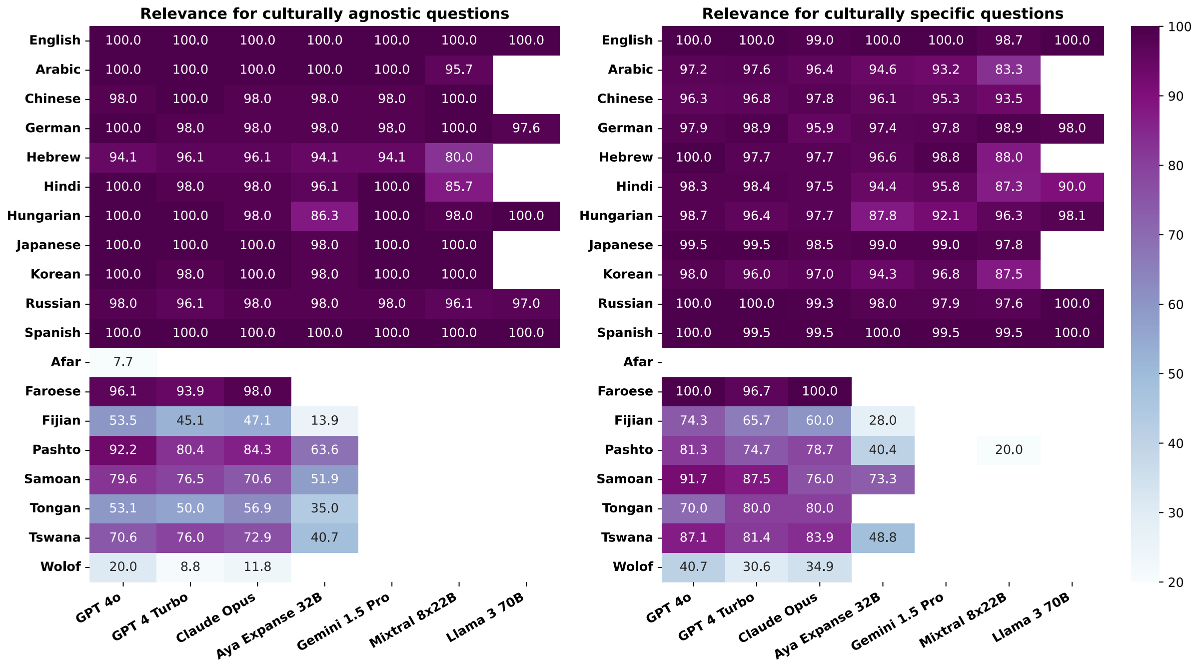

Figure 11: Relevance for answers without surface-level issues. The left heatmap shows the results for culturally agnostic questions while the right heatmap shows the results for culturally specific questions. We remove model-language combinations for which there are not at least 10 answers without surface-level issues. Answer relevance degrades for low-resource languages but is similar on culturally specific and culturally agnostic questions.

exhibit similar behavior, with less than 1% for other models. Lastly, we observed an unusually high proportion of emojis in responses generated by LLAMA-3-70B, with 17.54% containing at least one emoji.

Human- vs automatically-collected questions Table 14 shows model performance scores on human collected and automatically collected questions. We see comparable results between the two question sets, though some model rankings change. Specifically, GPT-4-TURBO and GEMINI-1.5-PRO move ahead of GPT-4O and CLAUDE-3-OPUS respectively in overall performance. Nevertheless, we see important trends like poor factuality scores and high MIXTRAL-8X22B repetitions on both sets of questions.

Figure 12 breaks down model overall performance on the human collected and automatically collected questions by language. We observe that model performance is higher on the automatically collected question for most languages. To determine whether the performance difference between the two question sets is significant, for each language we conduct a 2-sample Kolmogorov–Smirnov test on the overall answer scores for those languages, with the null hypothesis that the answer scores are drawn from the same distribution. The tests refute the null hypothesis with p-values below 0.01 for all languages except English and Hindi, for which p-values are 0.22 and 0.08 respectively. Although model performance is not identical on both question sets, analysis like Table 14 shows that model performance on automatically collected questions is an effective proxy for performance on human collected questions.

Answer statistics: We compute the lengths of generated answers using `tiktoken` with the `o200k_base` encoding. Table 17 presents statistics for the length of answers generated by each model. To account for variations in token count due to the language of generation and the presence of repetitions, we provide separate statistics for all answers and for those produced in the correct language without repetitions. Finally, we provide the percentage of answers produced in English for a non-English question in Figure 16.

D Human Evaluation

In this section, we present the details of human evaluation.

Evaluation Task The evaluation was conducted using `LabelStudio` (Tkachenko et al., 2020-2022). On the UI, annotators were presented with a question, a gold answer (if applicable), and three competitive answers in random order. The annotation

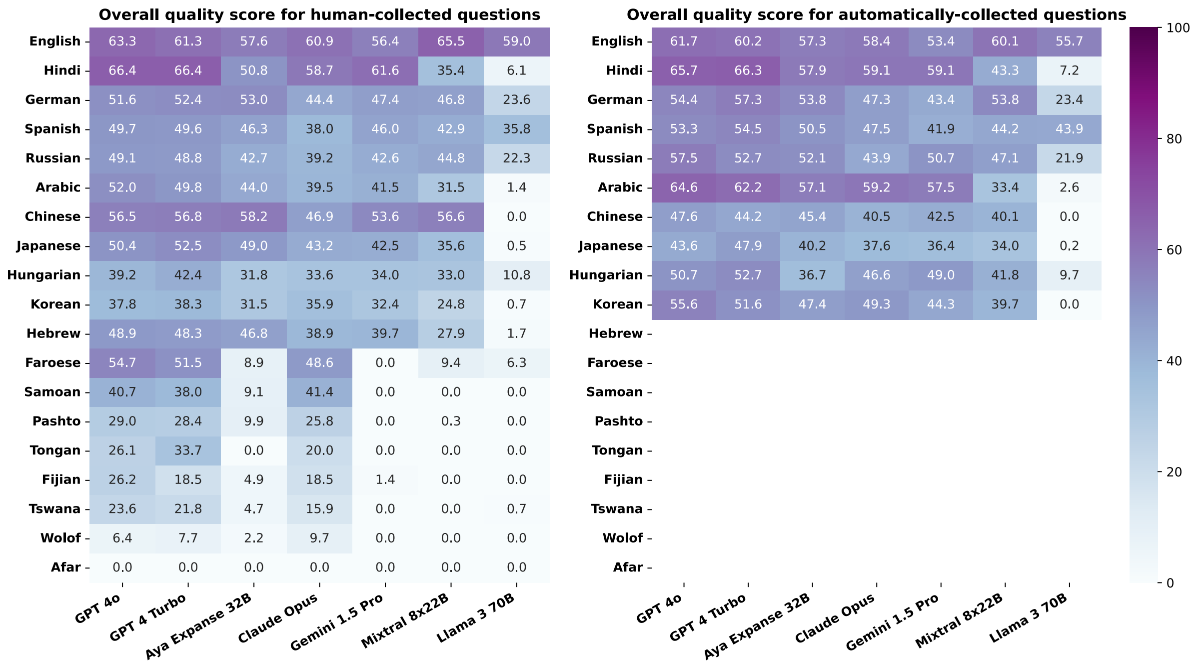

Figure 12: Answer scores S based on our quality criteria: surface issues, factuality and relevance. The left heatmap shows the results for human collected questions while the right heatmap shows the results for a subset of automatically collected questions (100 per language). Model performance differs notably between human collected and automatically collected questions in non-English languages.

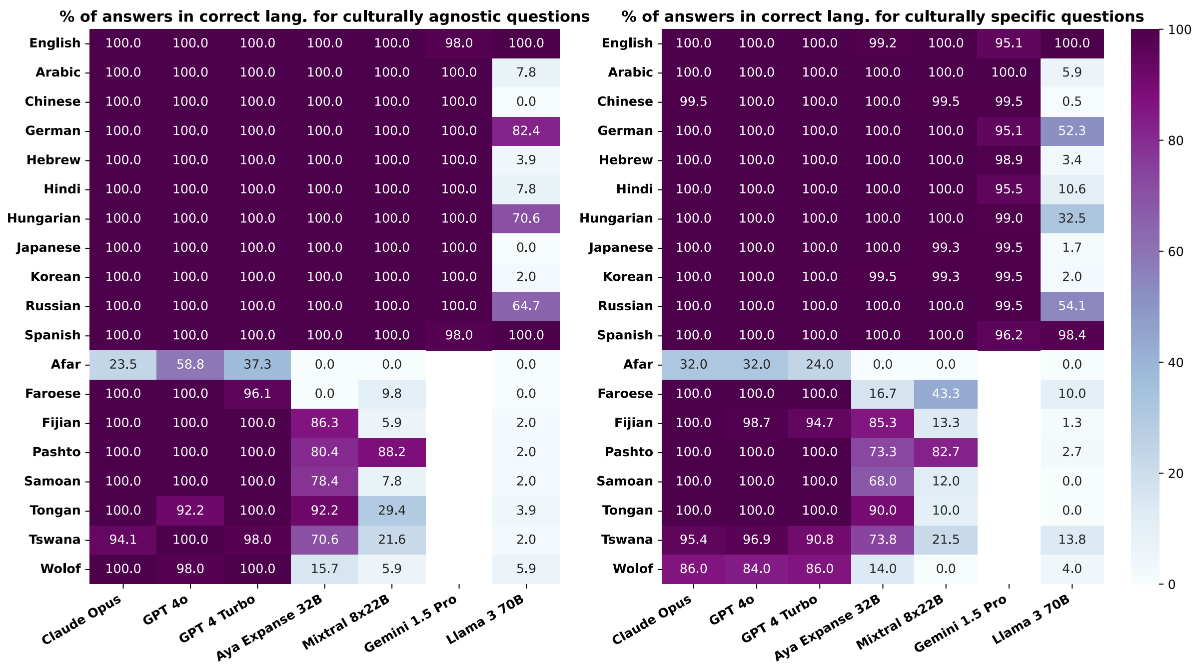

Figure 13: Percentage of responses generated in the correct language by model for culture specific and culturally agnostic questions. Blank cells for GEMINI-1.5-PRO indicate languages where the API returned an error message. Balochi, Hiligaynon, and Papiamento are omitted since language detection libraries performed poorly for these languages. Additionally, detection accuracy for Kirundi was compromised, with instances of Kirundi being incorrectly identified as Kinyarwanda. Please see Table 13 for details.

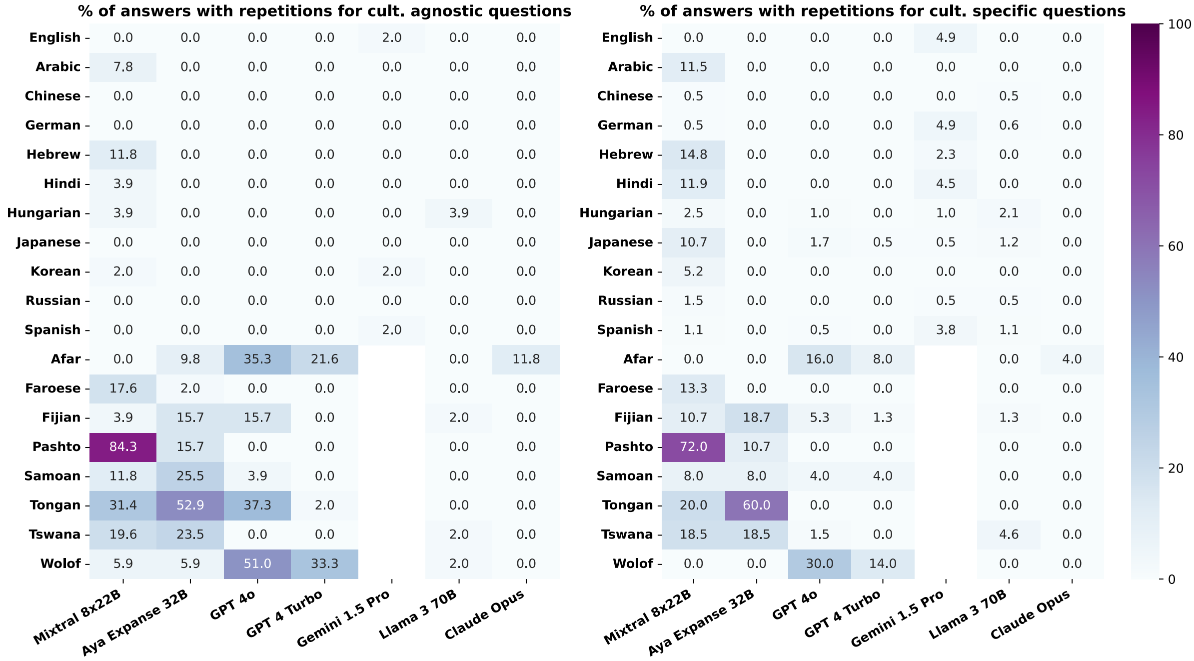

Figure 14: Percentage of answers with identified repetitions for culturally specific and culturally agnostic questions. Blank cells for GEMINI-1.5-PRO indicate languages where the API returned an error message.

You will be provided with a question and an answer, both translated from *[language]* into English. The answer was produced by an LLM.

Your task is to evaluate whether the answer is relevant to the question. **Relevance means that the answer makes a genuine attempt to address the question as stated.** The factual accuracy of the answer is not important. An answer should be judged as "irrelevant" if it:
- Does not attempt to answer the question,
- Focuses on a completely different topic,
- Is merely a refusal to answer, or
- Simply translates the content of the question without addressing it.

Instructions:

1. Read the question and the answer carefully.
2. Provide an explanation of your reasoning enclosed in `<explanation></explanation>` tags. 3. Then, give your final judgment enclosed in `<judgment></judgment>` tags with either the word `relevant` or `irrelevant`.

Content:

Question:
<question>
[question]
</question>

Answer:
<answer>
[answer]
</answer>

Figure 15: Prompt used with GPT-4O to determine the relevance of an LLM-generated answer to its question. Strings in the form *[form]* are placeholders that are replaced at runtime. The question and answer are provided in their English translations.

	Surface Level			Fine-Grained		Overall
	Wrong Lang. (%) ↓	Repetitions (%) ↓	W/o Issues (%) ↑	Factual Pr. (%) ↑	Relevance (%) ↑	Overall Score S ↑
GPT-4-TURBO	**0.0 / 0.0**	**0.0** / 0.1	**100.0 / 99.9**	**52.3** / 55.8	98.4 / 98.3	**52.0** / 55.0
GPT-4O	**0.0 / 0.0**	0.1 / 0.5	99.9 / 99.5	**52.3 / 56.5**	**98.7 / 98.5**	51.8 / **55.5**
AYA-EXPANSE-32B	0.1 / 0.1	**0.0 / 0.0**	99.9 / **99.9**	48.3 / 51.2	96.6 / 95.8	46.8 / 49.8
GEMINI-1.5-PRO	0.8 / 3.1	0.8 / 2.9	99.2 / 96.9	47.7 / 48.6	94.8 / 95.2	46.1 / 47.8
CLAUDE-3-OPUS	**0.0 / 0.0**	**0.0 / 0.0**	**100.0 / 99.9**	44.6 / 50.4	98.5 / 97.1	44.0 / 49.0
MIXTRAL-8X22B	0.3 / 0.1	4.4 / 4.7	95.4 / 95.2	45.5 / 47.7	93.8 / 94.3	41.6 / 43.8
LLAMA-3-70B	61.2 / 66.0	0.7 / 0.5	38.2 / 33.7	44.9 / 49.8	96.4 / 96.3	16.2 / 16.5

Table 14: Model performance aggregated across 10 high resource languages on human collected and a subset of automatically collected questions (100 per language). Each cell reports values on human collected / automatically collected culturally specific question.

MODEL	CHECKPOINT	AVAIL.	# PARAM
GPT-4-TURBO	`gpt-4-0125-preview`	**closed**	?
GPT-4O	`gpt-4o-2024-05-13`	**closed**	?
CLAUDE-3-OPUS	`claude-3-opus-20240229`	**closed**	?
GEMINI-1.5-PRO	`gemini-1.5-pro-preview-0514`	**closed**	?
LLAMA-3-70B	`Meta-Llama-3-70B-Instruct` x	**open-weight**	70B
MIXTRAL-8X22B	`Mixtral-8x22B-Instruct-v0.1` x	**open-weight**	8x22B
AYA-EXPANSE-32B	`aya-expanse-32b` x	**open-weight**	32B

Table 15: General information about models we evaluate using CALMQA.

process for each answer involved: (1) marking any mistakes,[18] (2) stating whether the answer is in the correct language, (3) evaluating factual accuracy, (4) noting any content omissions, (5) commenting on the overall quality of each answer, (6) rating each answer on a 5-point scale (excellent, good, average, poor, unusable). Upon completing the ratings, annotators ranked the three answers from best to worst and provided a free-form explanation for their ranking. Figure 17 illustrates the overall flow of the evaluation task. The study was submitted for the review to Institutional Review Board and received a *non-human subject* determination.

Guidelines and Consent We provided human evaluation guidelines, describing how to use the interface (including a tutorial video) and explaining each of the steps in the annotation process. The guidelines link to the consent form.

Data Human evaluation was done for answers generated by CLAUDE-3-OPUS, GPT-4-TURBO, and MIXTRAL-8X22B for questions in English, German, Hindi, Fijian and Kirundi. For culturally specific questions, annotators chose 10 questions in their language that they felt confident they knew the answer to. For culturally agnostic questions, we sampled 10 English culturally agnostic questions, and used the original English and the translations into the 4 other languages. We provided annotators with bullet-point answers in English for the culturally agnostic questions.

Workers and Cost German and Hindi annotators were recruited via Prolific, while Fijian and Kirundi annotators were recruited via Upwork. English annotations were performed by one of the authors. All annotators were native speakers of their respective languages and had participated in the question collection. Each question took approximately 20–40 minutes to evaluate, with annotators receiving compensation of $7.50 USD per question and an additional $8.00 USD for reviewing the guidelines, totaling $158 USD per language. The overall cost of the evaluation amounted to approximately $720 USD.[19]

Results Figure 18 and Figure 19 show the results of annotation for whether the answer was generated in the same language as the question (see Table 25

[18] This step was included to help the annotators visualize any issues with the answer.

[19] We also covered Upwork charges which the platform impose on the freelancers.

ISSUE	QUESTION	GENERATION
The model answers a non-English question in English	A BKV járatain egy (külföldi) nyugdíjas miként igazolja, hogy nyugdíjas a sofőrnél? … *How can a (foreign) senior citizen prove their eligibility for free travel on Budapest's public transport (BKV) to the driver? …*	I understand your concern! As a foreign pensioner, your mother is entitled to free travel on BKV (Budapest public transportation) services, but she needs to prove her pensioner status to the driver or inspector. According to the BKV website, foreign pensioners can use their national pensioner ID or a document that proves their pensioner status, along with a valid ID (e.g., passport). The document should be in English or have an English translation. Here are some options your mother can consider: … [Hungarian/LLAMA-3-70B]
A sequence of at least 20 tokens is repeated at least 4 times	Ke eng se se neng se le pharologano ka lenyalo la ga Sir Seretse Khama le Ruth Khama mme Ruth Khama o ne a itshwara jang lefa go nale pharologanyo eo? *What was different about Sir Seretse Khama and Ruth Khama's marriage and how did Ruth Khama behave despite the difference?*	Ke a leboga gore o ne a reng ditlhong tse di ntle. Ka kakaretso, ga Sir Seretse Khama le Ruth Khama o ne a itshwara jang lefa go nale pharologano ya gagwe. Ruth Khama o ne a neng a reng mabaka a a latelang a a nang le bona, a a nang le lefatsheng la Botswana, a a nang le baagi ba bangwe ba bangata ba bangwe … [Tswana/AYA-EXPANSE-32B]
The model apologizes and is unable to answer the question	Duni daddarba leh yan iyyaan duma duni core kee akah yan iyya? *How do they know how hot the earth's core is?*	I'm sorry for any confusion, but it seems like the text you've provided is in a language that I don't currently support. I can understand and generate text in several languages, but I need to know which one you're using to provide an accurate response. If you're able to translate your question into English, I'd be more than happy to help! [Afar/MIXTRAL-8X22B]
The model recognizes the language of the text	Umugirwa wo gufuba umwami wagenda gute mu Burundi? *How was the Kings' funeral done in Burundi?*	A question in Kirundi! Umugirwa wo gufuba umwami wagenda gute mu Burundi? Translation: "What is the significance of the drum in Burundian culture?" In Burundian culture, drums (umugirwa) play a significant role in various aspects of life, including: … [Kirundi/LLAMA-3-70B]
The model translates the question into English	Mahimo bala kontrolon ukon himuon sang tawo ang tiempo? Kon mahimo, hasta diin kag paano? *Can weather be controlled or man made? If so, to what extent and how?*	What a fascinating question! In English, your question translates to: "Can we control time like we control a person? If so, up to what extent and how?" Unfortunately, the answer is no, we cannot control time like we control a person. Time is a fundamental concept that governs the universe, and it is not a physical entity that can be manipulated or controlled. … [Hiligaynon/LLAMA-3-70B]

Table 16: Examples of issues in model responses that are detected as surface-level errors.

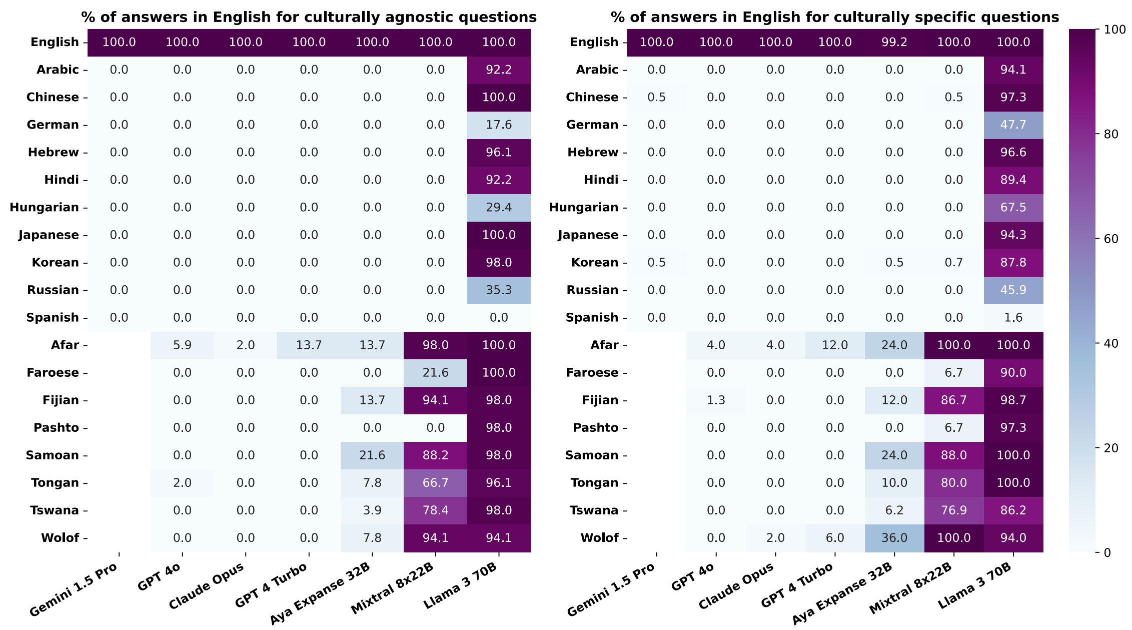

Figure 16: Percentage of answers produced in English by model which produced the answer for culturally specific and culturally agnostic questions. Blank cells for GEMINI-1.5-PRO indicate languages where the API returned an error message.

Model	All Data			Correct Lang / No Repetitions		
	Mean	Median	Std	Mean	Median	Std
Claude-3-Opus	296.4	293	88.9	302.2	297	79.2
GPT-4-Turbo	472.6	482	155.2	468.9	477	147.2
GPT-4o	446.6	425	268	434.9	430	184.8
Gemini-1.5-Pro	265.6	270	247.1	421.6	421	177.7
Aya-Expanse-32B	449.4	437	187.3	476.3	460	289.7
Llama-3-70B	395.9	410	171.4	478.7	484	138.8
Mixtral-8x22B	305.3	237	281.9	255.4	252	114

Table 17: Mean, median, and standard deviation of token counts in answers generated by different models. To account for variations in token count due to the language of generation and the presence of repetitions, we provide separate statistics for all answers and for answers produced in the correct language without repetitions. Token counts were computed using `tiktoken` with the `o200k_base` encoding.

for detailed counts). Figure 20 and Figure 21 display the annotations of the severity of factual issues in each answer (see Table 26 for detailed counts). Figure 22 and Figure 23 present the annotations of the severity of omissions in each answer (see Table 27 for detailed counts). Figure 24 and Figure 25 show the rankings of the models for both culturally specific and culturally agnostic questions. Figure 4 showsh ratings by model by question type. Finally, Figure 26 shows the distributions of scores assigned for each model by the question type and language of generation.

Statistical analysis We conducted a statistical analysis using the `clmm()` function from the `ordinal` package in R. Each model was fitted with the ordinal ratings (1–5) as the response variable and different predictors, allowing for random intercepts for annotators. Table 20 shows the results of a model with question type (either culturally specific or culturally agnostic) as the predictor. Table 21 presents the results of an analysis with model type, question type, and their interaction as predictors, complemented by Table 22, which shows the results of a post-hoc analysis. Finally, Table 23 displays the R^2 values for models with different predictors, namely model type, question type, omission ratings, factuality ratings, and language accuracy ratings.

Analysis of the annotations We conducted manual analyses of the comments provided by the annotators. For each analysis, we iteratively designed an annotation schema to analyze the submitted comments. Table 28 describes the annotation schema used for analyzing the comments on model ranking (i.e., the annotator's reason for ranking a model 1st, 2nd, or 3rd). The results of this analysis are presented in Table 29. Table 30 describes the categories used to analyze the comments on factuality. The results of this analysis are presented in Table 31. Finally, Table 32 describes the categories used to analyze the general comments left by the annotators for each answer. The results of this analysis are presented in Table 33 and Table 34.

D.1 Additional Insights

We capture here insights gained from analyzing human evaluation that we could not fit in the main body of text.

Enumerating facts makes responses seem less human-like. German and Hindi annotators remarked about the presence fact enumerations (often in the form of dot points) for some model answers. For German, the enumeration structure made responses seem artificial (e.g. *'Again very AI made structure. "here are common methods" and a following enumeration plus the asterisk titles...'*). For Hindi, listing facts makes the responses not seem human-like, though not necessarily like an AI either (e.g. *'The answer is just stating points on why is smoking harmful, so it neither sounds human-like nor artificial.'*). More broadly, the fact enumeration structure was described negatively in 5 responses, neutrally in 18 responses and positively in 2.

GPT-4-Turbo made the most grammar/spelling errors. 9 out of 12 spelling and grammar issues were noted for GPT-4-Turbo responses. 8 of these issues occurred for Fijian (e.g. 'There is a minor error, and the system might have spelled "nodra""

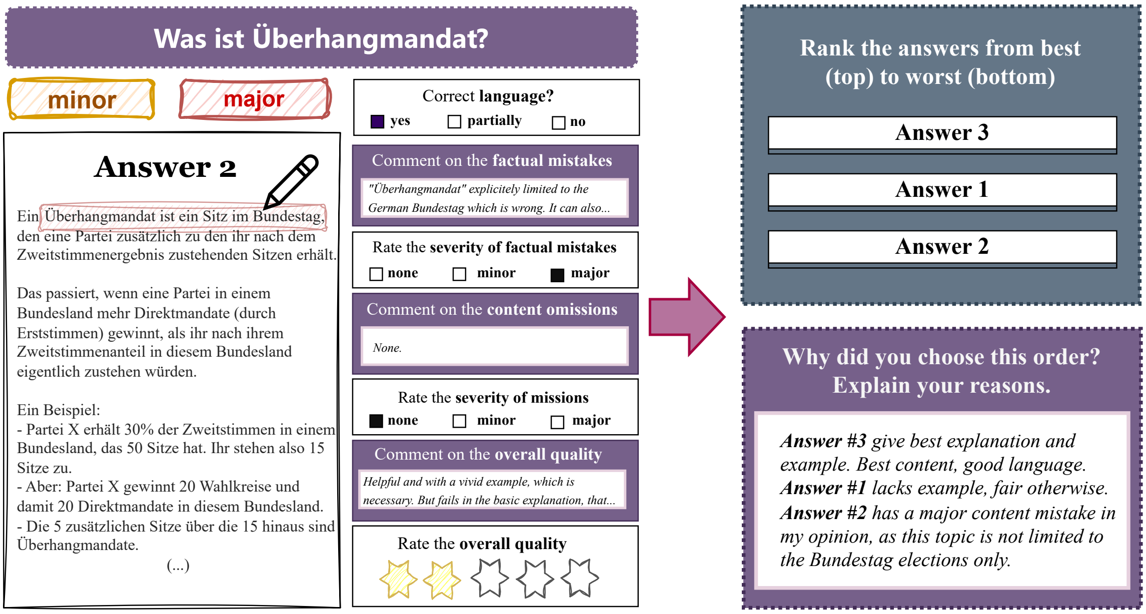

Figure 17: Our human evaluation pipeline. The annotator has to first read the answer, mark and classify all the mistake, and then comment and rate different properties of the answer. Once they have completed evaluating all three answers they are asked to rank them with respect to each other and provide a justification for the ranking. The example shows a culturally specific questions and one answer in German. The answer was produced by CLAUDE-3-OPUS.

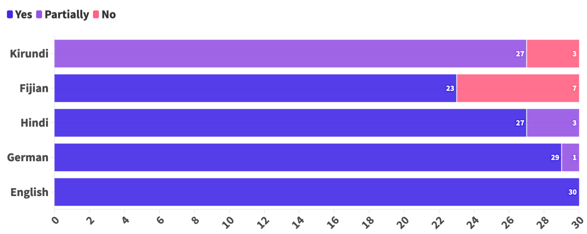

Figure 18: Annotations on Language Correctness for Culturally Specific Questions

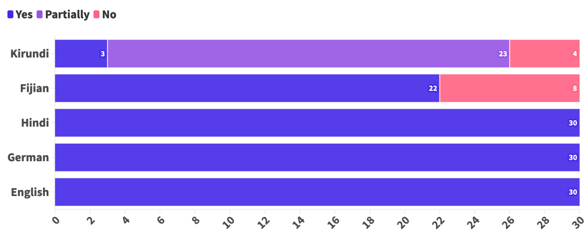

Figure 19: Annotations on Language Correctness for Agnostic-Specific Questions

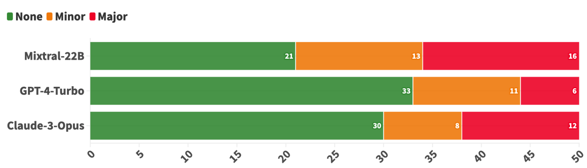

Figure 20: Factuality issues as assessed by the annotators by model for **culturally specific** questions

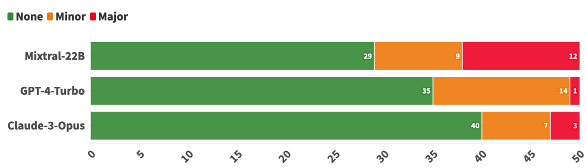

Figure 21: Factuality issues as assessed by the annotators by model for **culturally agnostic** questions

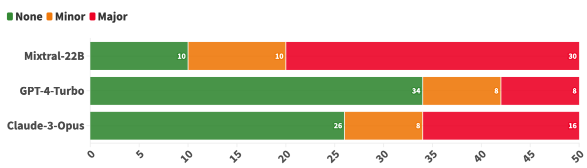

Figure 22: Omissions as assessed by the annotators by model for **culturally specific** questions

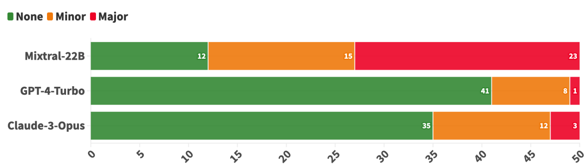

Figure 23: Omissions as assessed by the annotators by model for **culturally agnostic** questions

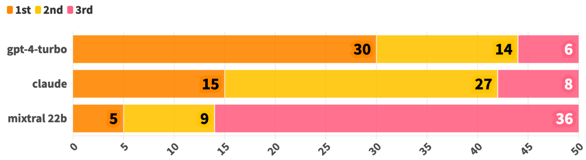

Figure 24: Number of times each model was ranked as *first*, *second*, and *last* for **culturally specific** questions.

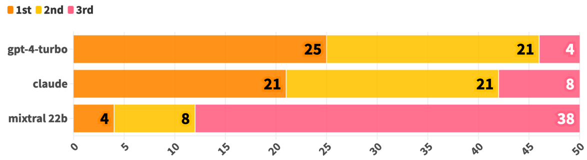

Figure 25: Number of times each model was ranked as *first*, *second*, and *last* for **culturally agnostic** questions.

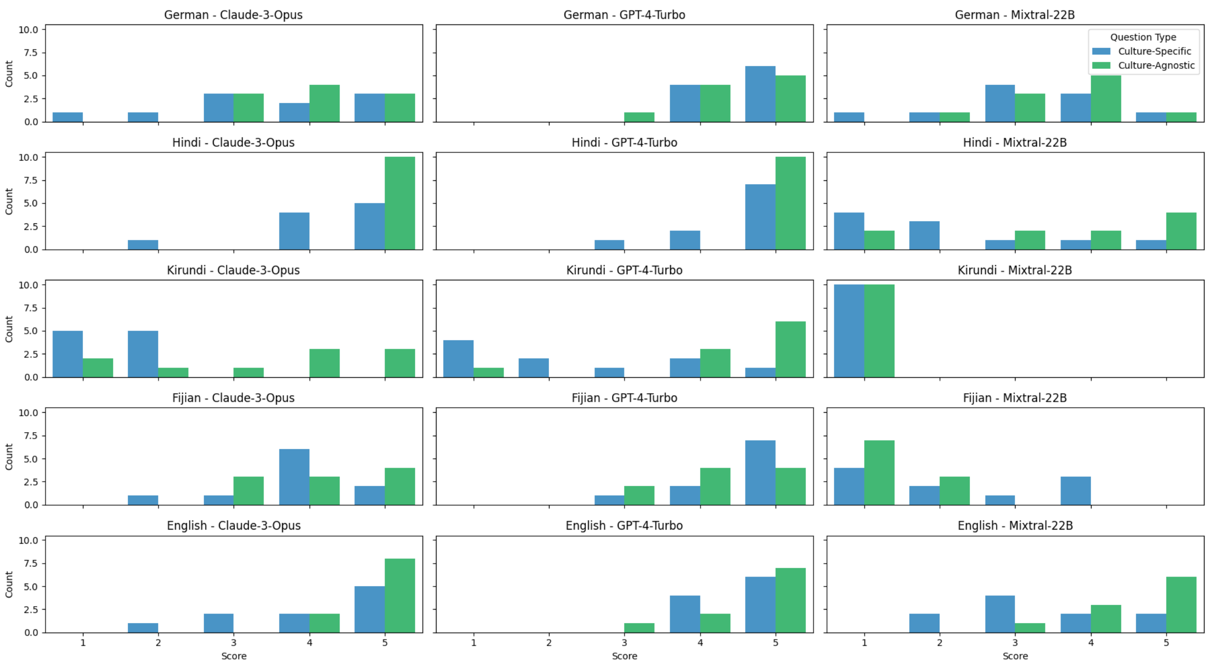

Figure 26: Scores distribution by language and model for Culturally Specific and Culturally Agnostic questions

Formula	rating ~ model + (1 | language/annotator)			
	Random Effects			
Group	Name	Variance	Std. Dev.	
language	(Intercept)	0.7175	0.847	
Number of groups:	5			
	Fixed Effects			
Coefficient	Estimate	Std. Error	z value	Pr(> $\|z\|$)
GPT-4-TURBO	0.8635	0.2885	2.993	0.00276 **
MIXTRAL-8X22B	-1.9493	0.2844	-6.854	7.18e-12 ***
Significance codes:	0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1			

Table 18: Results of cumulative link mixed model with ordinal ratings as the response variable and model as the predictor.

Contrast	**Estimate**	**SE**	**df**	***z*-ratio**	***p*-value**
CLAUDE-3-OPUS – GPT-4-TURBO	-0.863	0.288	Inf	-2.993	**0.0078**
CLAUDE-3-OPUS – MIXTRAL-8X22B	1.949	0.284	Inf	6.854	**<.0001**
GPT-4-TURBO – MIXTRAL-8X22B	2.813	0.315	Inf	8.936	**<.0001**
P value adjustment: Tukey method for comparing a family of 3 estimates					

Table 19: Post-hoc analysis for the model in Table 18. Tests performed using the emmeans library in R.

incorrectly. However, the language content is relevant so the rating is 4 out of 5, and it sounds like a human.') and the last was in German ('Defninetly helpful, complete and clear. Also fluent. One spelling mistake found: Zusammengefasend is no German word should be "zusammengefasst" or similar. But that could be a human-alike typo.'). This mistakes were present in otherwise mostly positive responses, suggesting that the issues were not due to lack of language understanding. We suspect that this phenomenon may be the result of a tokenizer issue.

Formula	rating $\sim$ type + (1 | language/annotator)			
	Random Effects			
Group	Name	Variance	Std. Dev.	
annotator	(Intercept)	0.9418	0.9705	
Number of Groups: annotator		5		
	Fixed Effects			
Coefficient	Estimate	Std. Error	z value	Pr(> $\|z\|$)
Culturally Agnostic	0.7259	0.2192	3.312	0.000926***
Significance codes:	0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1			

Table 20: Results of cumulative link mixed model with ordinal ratings as the response variable and question type (culturally specific vs culturally agnostic) as the predictor.

Formula	rating $\sim$ model * type + (1 | language/annotator)				
Random effects:					
Groups	**Name**	**Variance**	**Std.Dev.**		
Language	(Intercept)	0.7566	0.8698		
Annotator:Language	(Intercept)	0.7566	0.8698		
Fixed effects:					
Coefficient	**Estimate**	**Std. Error**	***z*-value**	**Pr(>|z|)**	
GPT-4-TURBO	1.1679	0.3868	3.020	0.002531	**
MIXTRAL-8X22B	-1.7013	0.3853	-4.415	1.01e-05	***
Culturally Agnostic	1.3561	0.3956	3.428	0.000607	***
GPT-4-TURBO:Culturally Agnostic	-0.6182	0.5849	-1.057	0.290489	
MIXTRAL-8X22B:Culturally Agnostic	-0.7163	0.5467	-1.310	0.190109	

Table 21: Cumulative link mixed model fitted with the Laplace approximation fitted with `clmm()` in R. The response variable is the ratings (an ordinal variable on a 5-point scale), with predictors being model (CLAUDE-3-OPUS, GPT-4-TURBO, or MIXTRAL-8X22B) and question type (culturally specific and culturally agnostic). Annotator nested within language is included as a random effect. The baseline model is CLAUDE-3-OPUS and the baseline question type is culturally specific. Model's conditional R^2 is 0.497 (including random effects) and marginal R^2 is 0.266 (only fixed effects). Please refer to Table 22 for post-hoc analysis.

Contrast	**Estimate**	**SE**	**df**	**z-ratio**	**p-value**
Spec. CLAUDE-3-OPUS- Agn. CLAUDE-3-OPUS	-1.356	0.396	Inf	-3.428	**0.0091**
Spec. CLAUDE-3-OPUS- Spec. GPT-4-TURBO	-1.168	0.387	Inf	-3.020	**0.0380**
Spec. CLAUDE-3-OPUS- Agn. GPT-4-TURBO	-1.906	0.424	Inf	-4.492	**0.0001**
Spec. CLAUDE-3-OPUS- Spec. MIXTRAL-8X22B	1.701	0.385	Inf	4.415	**0.0002**
Spec. CLAUDE-3-OPUS- Agn. MIXTRAL-8X22B	1.061	0.374	Inf	2.835	0.0687
Agn. CLAUDE-3-OPUS- Spec. GPT-4-TURBO	0.188	0.412	Inf	0.457	1.0000
Agn. CLAUDE-3-OPUS- Agn. GPT-4-TURBO	-0.550	0.442	Inf	-1.242	1.0000
Agn. CLAUDE-3-OPUS- Spec. MIXTRAL-8X22B	3.057	0.429	Inf	7.123	**<.0001**
Agn. CLAUDE-3-OPUS- Agn. MIXTRAL-8X22B	2.418	0.414	Inf	5.842	**<.0001**
Spec. GPT-4-TURBO- Agn. GPT-4-TURBO	-0.738	0.436	Inf	-1.694	1.0000
Spec. GPT-4-TURBO- Spec. MIXTRAL-8X22B	2.869	0.420	Inf	6.836	**<.0001**
Spec. GPT-4-TURBO- Agn. MIXTRAL-8X22B	2.229	0.404	Inf	5.514	**<.0001**
Agn. GPT-4-TURBO- Spec. MIXTRAL-8X22B	3.607	0.462	Inf	7.800	**<.0001**
Agn. GPT-4-TURBO- Agn. MIXTRAL-8X22B	2.967	0.445	Inf	6.669	**<.0001**
Spec. MIXTRAL-8X22B- Agn. MIXTRAL-8X22B	-0.640	0.383	Inf	-1.669	1.0000

Table 22: Post-hoc analysis for the model in Table 21 with Bonferroni adjustment. Spec. refers to culturally specific questions while Agn. refers to culturally agnostic questions. Tests performed using the `emmeans` library in R.

Predictor	**Conditional R^2**	**Marginal R^2**
Model	0.214	0.189
Omission	0.752	0.740
Factuality	0.614	0.560
Language Acc.	0.339	0.327
Q-Type	0.093	0.061
Model * Q-Type	0.497	0.266

Table 23: Conditional and Marginal R^2 values for different predictors. We fit cumulative link mixed models (`clmm()` in R) with *ratings* as the response variable and different predictors. All models included random intercepts for annotators. Omission, Factuality, and Language Accuracy were treated as ordinal variables (no issues > minor issues > major issues), whereas Q-Type and Model are categorical variables with two and three levels respectively. The last model was fitted with the interaction between the Model and the Q-Type. The Conditional R^2 refers to the variance explained by both fixed effects (predictors) and random effects (annotators), while Marginal R^2 refers to the variance explained by fixed effects only.

Model	Win Rate	Reason	# (*spec./agn.*)	Comment
GPT-4-Turbo	55%	Content	27 / 24	*Answer 1 (GPT-4-Turbo) is the perfect answer and and explains all the points needed to understand how to play the game 'Teen Patti'.*
		Language/ Format	11 / 7	*Answer 3 (GPT-4-Turbo) is very well structured and easy to follow. It covers all the information as well.*
		Factuality/ Relevance	10 / 3	*A3 (GPT-4-Turbo) is more factual than A1 (Claude-3-Opus) and A2 (Mixtral-8x22B).*
Claude-3-Opus	36%	Content	6 / 16	*Answer 3 (Claude-3-Opus) covers the topic in its entirety and hence is ranked 1st.*
		Language/ Format	5 / 13	*Answer 2 (Claude-3-Opus) is more readable because the information is listed as points.*
		Factuality/ Relevance	6 / 6	*A3 (Claude-3-Opus) is more detailed and factual than A1 (GPT-4-Turbo) and A2 (Mixtral-8x22B).*
Mixtral-8x22B	9%	Content	2 / 1	*A1 (Mixtral-8x22B) is better explained than A2 (Claude-3-Opus) and A3 (GPT-4-Turbo).*
		Language/ Format	3 / 2	*All answers have equal quality content, so they are distinguished by their structure/verbosity. Answer 3 (Mixtral-8x22B) has a very natural structure. Answer 2 (GPT-4-Turbo) and answer 1 (Claude-3-Opus) have redundancies and answer 1's are slightly worse.*
		Factuality/ Relevance	1 / 0	*Answer 2 (Mixtral-8x22B) was slightly more specific to Western Europe than Answer 1 (GPT-4-Turbo), but both were roughly equal in quality.*

Table 24: Win rates of the three models in human-evaluated 3-way comparisons of answers for 100 questions. Reasons behind the annotators' decisions are provided, with separate reason counts for *culturally specific* and *culturally agnostic* questions. A breakdown of reasons into finer-grained categories is provided in Table 28.

Language	Model	Culturally Specific			Culturally Agnostic		
		Yes	Partially	No	Yes	Partially	No
German	Claude-3-Opus	10	0	0	10	0	0
	GPT-4-Turbo	10	0	0	10	0	0
	Mixtral-8x22B	9	1	0	10	0	0
Hindi	Claude-3-Opus	9	1	0	10	0	0
	GPT-4-Turbo	10	0	0	10	0	0
	Mixtral-8x22B	8	2	0	10	0	0
Kirundi	Claude-3-Opus	0	10	0	1	9	0
	GPT-4-Turbo	0	10	0	1	9	0
	Mixtral-8x22B	0	7	3	1	5	4
Fijian	Claude-3-Opus	10	0	0	10	0	0
	GPT-4-Turbo	10	0	0	10	0	0
	Mixtral-8x22B	3	0	7	2	0	8
English	Claude-3-Opus	10	0	0	10	0	0
	GPT-4-Turbo	10	0	0	10	0	0
	Mixtral-8x22B	10	0	0	10	0	0

Table 25: Count of instances generated in the language of the question by model and question-type, and the language being evaluated

Language	Model	Culturally Specific			Culturally Agnostic		
		None	Minor	Major	None	Minor	Major
German	Claude-3-Opus	8	1	1	10	0	0
	GPT-4-Turbo	8	2	0	10	0	0
	Mixtral-8x22B	8	2	0	9	1	0
Hindi	Claude-3-Opus	7	2	1	10	0	0
	GPT-4-Turbo	7	3	0	10	0	0
	Mixtral-8x22B	1	4	5	6	2	2
Kirundi	Claude-3-Opus	0	1	9	3	4	3
	GPT-4-Turbo	0	4	6	3	6	1
	Mixtral-8x22B	0	0	10	0	0	10
Fijian	Claude-3-Opus	7	3	0	8	2	0
	GPT-4-Turbo	8	2	0	3	7	0
	Mixtral-8x22B	5	5	0	5	5	0
English	Claude-3-Opus	8	1	1	9	1	0
	GPT-4-Turbo	10	0	0	9	1	0
	Mixtral-8x22B	7	2	1	9	1	0

Table 26: Factuality issues in model generation by model, question type and language of the question

LANGUAGE	MODEL	CULTURALLY SPECIFIC			CULTURALLY AGNOSTIC		
		None	Minor	Major	None	Minor	Major
German	CLAUDE-3-OPUS	6	1	3	7	3	0
	GPT-4-TURBO	6	4	0	8	2	0
	MIXTRAL-8X22B	1	7	2	3	6	1
Hindi	CLAUDE-3-OPUS	8	1	1	10	0	0
	GPT-4-TURBO	9	0	1	10	0	0
	MIXTRAL-8X22B	3	1	6	4	4	2
Kirundi	CLAUDE-3-OPUS	0	0	10	4	3	3
	GPT-4-TURBO	2	2	6	6	3	1
	MIXTRAL-8X22B	0	0	10	0	0	10
Fijian	CLAUDE-3-OPUS	6	3	1	6	4	0
	GPT-4-TURBO	8	1	1	7	3	0
	MIXTRAL-8X22B	3	0	7	0	0	10
English	CLAUDE-3-OPUS	6	3	1	8	2	0
	GPT-4-TURBO	9	1	0	10	0	0
	MIXTRAL-8X22B	3	2	5	5	5	0

Table 27: Count of omission issues by severity type, model, and language for culturally specific and culturally agnostic questions

TYPE	DESCRIPTION	EXAMPLE
	Content: Completeness/Explanation	
COMPLETENESS	The answer was perceived as complete.	*Answer 3 (Claude-3-Opus) covers the topic in its entirety and hence is ranked 1st.* [Hindi]
EXPLANATION/EXAMPLES	The answer included useful explanation and/or examples.	*A3 (Gpt-4-Turbo) is better explained than A1 and A2.* [Fijian]
DETAILS/BACKGROUND	The answer included details and/or necessary background.	*Answer 1 (GPT-4-Turbo) and 2 (Claude-3-Opus) are similar but answer 1 has detailed information about the methods to measure body mass compared to 2.* [Hindi]
GENERAL	The answer was general, which was appropriate for the given question.	*Answer 3 (GPT-4-Turbo) is perfect. The writing style of the answer is the best compared to the other answers. For instance, it mentions the timeline in general rather than pointing out exact years of the event (In my opinion, different sources and online transcripts have a little variation in years in terms of history so it is the best to keep it general).* [Hindi]
	Language/Presentation	
LANGUAGE	The answer was fluent/used better language or was less AI-like.	*Both answer 1 (GPT-4-Turbo) and 3 (Claude-3-Opus) are good. Answer 1 sounds more human-like which is why it is ranked 1st.* [Hindi]
STRUCTURE	The structure of the answer was better.	*All answers have equal quality content, so they are distinguished by their structure/verbosity. Answer 3 (Mixtral-22B) has a very natural structure.* [English]
SIMPLE/CLEAR/SPECIFIC	The answer was clear and/or simple, to the point.	*Answer 3 (GPT-4-Turbo) is slightly clearer than answer 1 (Claude-3-Opus).* [English]
SUCCINCT	The answer was succinct.	*All three answers are complete by content in my view (good answer requires more, but that is more than question covers). Answer two (Mixtral-22B) I regard the best, as the density of content in a few lines is awesome - in most of the other questions, "death by long text and details" is valid, here I opt for short and good.* [German]
	Factuality/Correctness	
FACTUAL	The answer is better in terms of factuality.	*A3 (GPT-4-Turbo) is more factual than A1 (Claude-3-Opus) and A2 (Mixtral-22B).* [Fijian]
RELEVANT	The answer is the most relevant to the question. Often mentioned when other answers were irrelevant.	*Answer 3 (GPT-4-Turbo) is placed in the first position because it is relevant though not specific to Burundi. It could be used if it were specific. It has some important information (...).* [Kirundi]
NO ISSUES	There were no apparent issues in the answer.	*Answer 3 (Claude-3-Opus) had no notable issues.* [English]
PARTIAL ANSWER	The answer at least partially addresses the question (while other answers may be refusals, repetitions, or simply irrelevant/wrong).	*Answer 3 (GPT-4-Turbo) is the only one that tries to answer the question. The other 2 just point out differences between the medications.* [English]

Table 28: Categories used for analysis of reasons for specific ranking of the answers

		Completeness - Explanation				Language - Presentation				Factuality - Correctness			
Model	Type	Complete	Explanation	Details	General	Language	Structure	Simple	Succinct	Factual	Relevant	No issues	Partial ans
CLAUDE-3-OPUS	*Spec.*	4	0	2	0	1	2	1	1	1	1	2	2
GPT-4-TURBO	*Spec.*	11	5	9	2	3	2	4	2	4	1	2	3
MIXTRAL-8X22B	*Spec.*	1	1	0	0	1	0	1	1	1	0	0	0
CLAUDE-3-OPUS	*Agn.*	11	0	4	1	2	1	7	3	1	1	4	0
GPT-4-TURBO	*Agn.*	14	4	6	0	1	4	1	1	0	0	2	1
MIXTRAL-8X22B	*Agn.*	0	0	1	0	0	1	0	1	0	0	0	0

Table 29: Count of different reasons mentioned by the annotator for ranking each model's answer as the best out of three. Note that in some cases more than one reason might have been give by the annotator. *Spec.* refers to Culturally Specific questions, while *Agn.* refers to Culturally Agnostic questions.

Type	Description	Comment Example (Language/Model)
	Direct Factual Errors	
Date	Issues involving incorrect temporal references.	*It is mentioned that Nifty was launched in 1995 but it was actually launched in 1996.* [Hindi/GPT-4-Turbo]
Entity	Incorrect entity such as a person, place, or organization.	*Almost everything is incorrect because the answer states that Ntare Rugamba is the person who accepted to die in the place of the king, while Ntare Rugamba is the king who ruled before the King Mwezi Gisabo.* [Kirundi/GPT-4-Turbo]
Event	Errors in the details or occurrence of events.	*It says that Aurangzeb got the mosque built at the place, however this claim is very strong. He got the temple destroyed but it is not sure if he got the mosque built, as it was a decade after the demolition of the temple.* [Hindi/Mixtral-8x22B]
Reason	Incorrect reasons or causative explanations for events or situations.	*"Ni o lobika na ligamu, o sa vakalevutaka na kena yawa mai yalomu" means folding your elbow increase the distance from your spirit"* [Fijian/GPT-4-Turbo]
	Contextual and Logical Errors	
Scope	Errors involving the incorrect extent or range of a fact.	*"Überhangmandat" explicitely limited to the German Bundestag which is wrong. It can also apply for regional votes for a single state for instance.* [German/Claude-3-Opus]
Illogical	Statements that are logically inconsistent or defy common sense.	*Yes, there is an error in the phrases "e rawa ni vakavuna na turu ni wai ni mata ena vuku ni kena vakacacani na mata ni mata" means the cutting of onion can damage the eye of the eye sounds grammatically incorrect.* [Fijian/GPT-4-Turbo]
Negation	Incorrect use of negation or statements that present the opposite of what is true.	*Answer implies that internists are not PCPs* [English/Claude-3-Opus]
Placement	Inaccuracies resulting from the improper contextual placement of information.	*Putting "7. secondhand smoke" as a dot point incorrectly suggests that it is a reason why smoking is bad for the heart* [English/Claude-3-Opus]
	Interpretative and Relevance Errors	
Irrelevant or Hallucinated	All information in the answer is irrelevant or is fabricated.	*The entire answer is wrong. The question is to state the difference between General and OBC category of people (based on caste and society status), but the answer is about general and gynecology practitioners (doctors). It is irrelevant to the question.* [Hindi/Mixtral-8x22B]
Different Culture	The answer includes references to the wrong culture.	*The answer given is incorrect because it is not specific to Burundi. It is in Kinyarwanda and talks about Rwanda.* [Kirundi/Claude-3-Opus]
Unclear or Undocumented	Facts that are not clearly expressed or lack sufficient documentation or sourcing.	*The phrase "leweka" is not clear in the content. This issue might have a typo error.* [Fijian/Claude-3-Opus]
Not an answer	The model refuses to answer usually producing the response in English sometimes accompanied with an attempt to translate the question.	*Incorrect language. (refusal to answer produced in English)* [Fijian/Mixtral-8x22B]

Table 30: Categories used for analysis of comments on the factuality of the answers

MODEL	TYPE	UNCLEAR/ UNDOC.	DATE	ENTITY	EVENT	REASON	SCOPE
CLAUDE-3-OPUS	*Spec.*	1	1	2	1	0	2
GPT-4-TURBO	*Spec.*	0	1	1	0	0	1
MIXTRAL-8X22B	*Spec.*	1	1	2	2	1	1
CLAUDE-3-OPUS	*Agn.*	2	0	0	0	0	0
GPT-4-TURBO	*Agn.*	5	0	0	0	1	0
MIXTRAL-8X22B	*Agn.*	2	0	0	0	1	0
MODEL	TYPE	ILLOGICAL	PLACEMENT	NEGATION	IRRELEVANT OR HALLUCINATED	DIFFERENT CULTURE	NOT AN ANSWER
CLAUDE-3-OPUS	*Spec.*	4	0	1	6	4	0
GPT-4-TURBO	*Spec.*	2	1	0	0	7	0
MIXTRAL-8X22B	*Spec.*	3	0	2	7	0	8
CLAUDE-3-OPUS	*Agn.*	1	1	0	6	0	0
GPT-4-TURBO	*Agn.*	1	1	1	4	0	0
MIXTRAL-8X22B	*Agn.*	0	0	1	5	0	10

Table 31: Count of different types of factuality issues mentioned by annotators in their comments. The issues are presented by question type (*culturally specific* or *culturally agnostic*) and by model which generated the answer. The taxonomy used for this annotation can be found in Table 30.

Category	Description	Comment Example (Language/Model)
		Content Issues
Unhelpfulness	Annotator cannot discern the question's answer from the provided answer text	*This answer is not very helpful because it not specific to Burundi. Of course it contains some relevant information, but it lacks specificity.* [Kirundi/GPT-4-Turbo]
Irrelevance	Contains information not needed to answer the question	*The answer is irrelevant to the question and it isn't helpful. It explains a completely different thing (explains the importance of blood sample rather than explaining why the blood sample is taken from the elbow area).* [Hindi/Mixtral-8x22B]
Underspecificity	Ideas are not explained in sufficient detail	*Some instances of generic reasons without much detail about how it applies in Western Europe in particular.* [English/GPT-4-Turbo]
Incompleteness	Extra information is needed to understand the answer	*The points mentioned in the answer are true but they do not cover the entirety of the topic. ... So it could be said that the answer is subjective rather than complete and clear.* [Hindi/GPT-4-Turbo]
Repetitiveness	Some phrase or idea is repeated	*The phrases "Sa vakamuria na noda vakarau ni dautama o ira na iTaukei vulagi.....," is repeated in almost every line...* [Fijian/Mixtral-8x22B]
Grammar/spelling	Grammar or spelling mistakes are present	*There is a minor error, and the system might have spelled "nodra" incorrectly.* [Fijian/Claude-3-Opus]
		Content Merits
Helpfulness	Information needed to answer the question is present and understandable	*It is helpful because it tells us the reasons for updating apps: Security, Good performance and New features.* [Kirundi/GPT-4-Turbo]
Completeness	Answer does not need extra information to be understood	*It is complete, even makes an arc from outer core to inner core and a comparison to the sun surface temperature. So it covers more for what has been asked for but I like this additional information.* [German/GPT-4-Turbo]
		Presentation Issues
Artificiality	Perceived by the annotator to be written by AI	*Structure with asterix and enumeration very artificial / non human like.* [German/GPT-4-Turbo]
Non-fluency	Reading the text is awkward/difficult for the annotator	*...Also Argument 5 is "estimation" and this is again mentioned in the summary -> a bit stange and hence not fluent.* [German/Claude-3-Opus]
Inclarity	Answer ideas or terms are difficult to understand	*Fluent and complete, but not clear, as complex word i.e. Zytokinen (Cytokines) are not explained.* [German/Claude-3-Opus]
		Presentation Merits
Naturality	Perceived by the annotator to be human-written/natural	*Complete, but so different to the other answers. most human-like answer as being relatively general to a more or less vague question.* [German/Claude-3-Opus]
Fluency	Written in a smooth manner and easy to read	*There are no inaccuracies in the answer which makes it helpful. It makes use of easier vocabulary which sounds fluent.* [Hindi/Claude-3-Opus]
Clarity	Ideas are expressed in an easy-to-understand manner	*The answer is complete and helpful. It is very clear because the information is subdivided into general and OBC sections and it is easy to follow.* [Hindi/GPT-4-Turbo]

Table 32: Categories used for the analysis of annotators' general comments on the quality of answers

Model	Type	Unhelpfulness*	Incompleteness*	Artificiality*	Non-fluency*	Inclarity*
Claude-3-Opus	*Spec.*	9	13	4	4	10
GPT-4-Turbo	*Spec.*	6	7	4	2	5
Mixtral-8x22B	*Spec.*	20	17	19	10	15
Claude-3-Opus	*Agn.*	3	5	5	1	1
GPT-4-Turbo	*Agn.*	1	4	11	1	5
Mixtral-8x22B	*Agn.*	16	14	17	8	9
Model	Type	Irrelevance	Underspecifity	Repetitiveness	Grammar/Spelling	
Claude-3-Opus	*Spec.*	1	1	0	1	
GPT-4-Turbo	*Spec.*	4	1	0	2	
Mixtral-8x22B	*Spec.*	4	7	4	1	
Claude-3-Opus	*Agn.*	1	1	2	1	
GPT-4-Turbo	*Agn.*	0	0	0	7	
Mixtral-8x22B	*Agn.*	3	1	4	0	

Table 33: Counts of different types of issues noted in annotators' comments about general answer quality. The issues are presented by question type (*culturally specific* or *culturally agnostic*) and by model which generated the answer. The taxonomy used for this annotation can be found in Table 32. Our UI suggested to annotators to make comments (positive or negative) about categories marked with *.

Model	Type	Helpfulness	Completeness	Naturality	Fluency	Clarity
Claude-3-Opus	*Spec.*	24	19	24	22	22
GPT-4-Turbo	*Spec.*	24	20	27	22	19
Mixtral-8x22B	*Spec.*	13	7	18	15	12
Claude-3-Opus	*Agn.*	26	20	30	22	30
GPT-4-Turbo	*Agn.*	29	24	26	17	25
Mixtral-8x22B	*Agn.*	16	12	22	13	16

Table 34: Counts of different types of merits noted in annotators' comments about general answer quality. The merits are presented by question type (*culturally specific* or *culturally agnostic*) and by model which generated the answer. The taxonomy used for this annotation can be found in Table 32. Our UI suggested to annotators to make comments (positive or negative) about all these categories.