Evaluating the Elementary Multilingual Capabilities of Large Language Models with MULTIQ

Carolin Holtermann[1*], Paul Röttger[2*], Timm Dill[1], Anne Lauscher[1]
[1]Data Science Group, University of Hamburg, Germany
[2]Bocconi University, Italy
carolin.holtermann@uni-hamburg.de

Abstract

Large language models (LLMs) should benefit everyone, including a global majority of non-English speakers. However, most LLMs today, and open LLMs in particular, are often intended for use in just English (e.g. Llama2, Mistral) or a small handful of high-resource languages (e.g. Mixtral, Qwen). Recent research shows that, despite limits in their intended use, people prompt LLMs in many different languages. Therefore, in this paper, we investigate the basic multilingual capabilities of state-of-the-art open LLMs *beyond their intended use*. For this purpose, we introduce MULTIQ, a new silver standard benchmark for basic open-ended question answering with 27.4k test questions across a typologically diverse set of 137 languages. With MULTIQ, we evaluate language fidelity, i.e. whether models respond in the prompted language, and question answering accuracy. All LLMs we test respond faithfully and/or accurately for at least some languages beyond their intended use. Most models are more accurate when they respond faithfully. However, differences across models are large, and there is a long tail of languages where models are neither accurate nor faithful. We explore differences in tokenization as a potential explanation for our findings, identifying possible correlations that warrant further investigation.

1 Introduction

Languages other than English remain underrepresented and underserved by state-of-the-art language technologies, posing a barrier to equal and inclusive AI (Bender, 2011; Joshi et al., 2020). While proprietary large language models (LLMs) like GPT-4 (OpenAI, 2023) may answer questions and follow instructions in many different languages, even the best and most popular open LLMs are much more restricted in their language coverage: Llama2-chat (Touvron et al., 2023), for example, is *"intended for commercial and research use in English"*.[1] Yi is *"bilingual"* in English and Chinese[2], and Mistral-7b-instruct (Jiang et al., 2023) *"only works in English"*.[3]

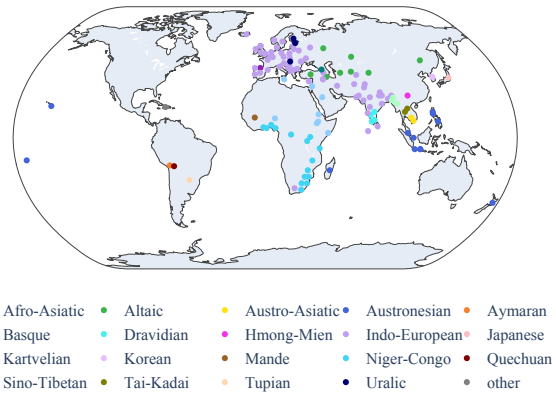

Figure 1: The 137 languages covered in our MULTIQ question dataset. We show their geographic location according to the WALS database and indicate their corresponding language family through colors.

Even though most open LLMs are restricted in their *intended* use to one or a handful of languages, datasets of real-world LLM usage show that people prompt LLMs in many different languages, often beyond their intended use (Ouyang et al., 2023; Zhao et al., 2024; Zheng et al., 2024). This has motivated initial research into the multilingual capabilities of monolingual models (Armengol-Estapé et al., 2022; Lai et al., 2023). However, this research has mostly focused on older proprietary LLMs and on a relatively small number of languages and/or specific tasks.

In this paper, we investigate the basic multilingual capabilities of a variety of state-of-the-art chat-optimized open LLMs across a typologically diverse set of 137 languages. Specifically, we ask

*Equal contribution

[1]https://huggingface.co/meta-llama/Llama-2-7b-hf
[2]https://huggingface.co/01-ai/Yi-34B-Chat
[3]https://mistral.ai/news/la-plateforme/

Findings of the Association for Computational Linguistics ACL 2024, pages 4476–4494

two main research questions that correspond to two dimensions of multilingual capability: 1) What is the **multilingual language fidelity** of current chat-optimized open LLMs, and 2) What is the **multilingual question answering (QA) accuracy** of current chat-optimised open LLMs? Language fidelity describes the ability to respond to prompts in the prompted language. QA accuracy describes the ability to give correct answers to open-ended questions, in the prompted language or any other. An ideal multilingual model would give answers that are both faithful and correct.

To answer our two research questions, we introduce MULTIQ, a new silver standard benchmark for basic open-ended question answering comprising 27,400 test prompts across 137 typologically diverse languages. We create MULTIQ by compiling 200 English questions that are simple yet realistic and diverse, and translating them automatically to 136 other languages. We evaluate QA accuracy on MULTIQ using a GPT-4 classifier and language fidelity using GlotLID (Kargaran et al., 2023). MULTIQ is a silver standard because automated translation and evaluation introduce some noise into the results. However, we validate through expert annotation that this noise is likely small, thus demonstrating that MULTIQ can provide valuable evidence on basic multilingual capabilities. Concretely, we use MULTIQ to make four main findings:

1. **Language Fidelity**: While some open models (e.g. Llama2) mostly respond in English regardless of the input language, other models (e.g. Mistral) respond faithfully despite their intended use being monolingual.
2. **QA Accuracy**: On MULTIQ, all models tend to perform best in English, with some performing similarly well in up to 20 other languages (e.g. Mixtral). Across models, there is a long tail of languages with very poor accuracy.
3. **Positive Interaction**: Increased language fidelity appears to positively impact answer accuracy, since model answers that match the prompt language tend to be more accurate.
4. **Tokenization as (Partial) Explanation**: Models tend to achieve higher accuracy on languages they can tokenize into subwords instead of characters or ASCII tokens.

We publish all data and code at https://github.com/paul-rottger/multiq

2 The MULTIQ Dataset

MULTIQ is a collection of 27,400 simple open-ended questions across 137 typologically diverse languages. The questions cover different topics ranging from algebra to geography to astronomy. Questions in each language are parallel to each other. The open-ended question format is consistent with real-world LLM usage. Additionally, the open-endedness minimizes the likelihood of correct answers given by chance.

2.1 Dataset Creation

We created MULTIQ in two steps. First, we compiled an initial set of English questions. Second, we automatically translated these prompts into 136 typologically diverse languages.

For the **initial English questions**, we used two different sources to increase question diversity. **a)** We collected 100 questions from the LMSYS-Chat-1M dataset (Zheng et al., 2024), which catalogs real-world user interactions with LLMs. Specifically, we sampled all single-sentence English-language and sorted them by frequency. Then, we manually selected from the top until we reached 100 questions. *This portion of our data directly reflects real-world LLM usage.* **b)** We manually created another set of 100 questions evenly spread across 10 different subjects at elementary to middle school level (e.g. mathematics and geography). To maximize the diversity of the questions, we prompted GPT-4 to provide us with a set of simple and clear questions with simple and clear answers for each of the subjects. We then iterated and manually selected questions until we reached 10 questions per subject. *This portion of our data expands* MULTIQ*'s topical coverage.*

In both the LMSYS and the GPT-4 portions of our data, we manually selected only questions that are simple, factual, and target common knowledge. This is because with MULTIQ we want to test *basic* multilingual capabilities, not complex reasoning. Questions must also have unambiguous answers that are culturally and temporally invariant. This is to minimize discrepancies introduced by translation as well as temporal degradation of our dataset. Table 1 shows English example prompts.

To **translate our English questions** into other languages, we used the v3 Google Translate API.[4] Specifically, we translated the 200 initial English questions into all 136 other languages covered by

[4] https://translation.googleapis.com/v3/

Source	Example Prompts
LMSYS	*Was the year 2000 a leap year?* *What is 2 + 2 * 3?* *How many feet does a chicken have?*
GPT-4	*What is the chemical formula for water?* *Who was the first Emperor of China?* *What is a galaxy?*

Table 1: Examples of English questions covered in MULTIQ. We present three prompts from each source (LMSYS and GPT-4) covering different domains.

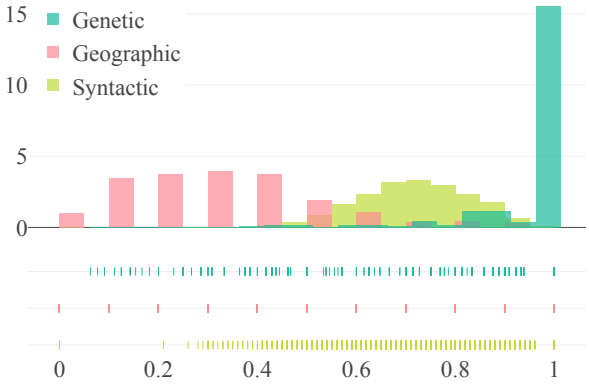

Figure 2: Distributions of the pairwise `lang2vec` distances for each language pair present in MULTIQ.

the API as of February 2024, resulting in a total of 27,400 parallel questions in MULTIQ. The decision to use automated translation was driven by the constraints of our research budget, which made manual translation infeasible. The benefit of automated translation is that we can cover many more languages. Next, we discuss how we validated the quality of the translations, and demonstrate the typological diversity of the 137 languages we cover.

2.2 Validation of Translations

We asked native speakers to annotate the correctness of the 200 translated MULTIQ questions for 19 languages: Arabic, Catalan, Chinese, Farsi, French, German, Hindi, Indonesian, Italian, Japanese, Korean, Spanish, Tagalog, Russian, Spanish, Quechua, Ukrain, Urdu and Xhosa. Across these languages, annotators marked an average of 91.6% of translations as correct. Translations were least accurate for Tagalog at 60.0% and Arabic at 82.2%, while the translations for Italian and German were the most accurate, at 99.0%. We present the full results in the Appendix A. Qualitatively, several annotators stated that some translations, while accurate in content, tended to be literal, word-for-word translations rather than natural expressions. Overall, the automated translation introduces some noise into MULTIQ, but our validation results suggest that the amount of noise is limited. This is why we frame MULTIQ as a *silver standard* benchmark that can provide meaningful insights into basic multilingual capabilities, even if exact results on individual test cases may not be perfectly reliable.

2.3 Typological Diversity

MULTIQ covers a total of 137 languages. To demonstrate their typological diversity, we follow best practices suggested by Ploeger et al. (2024), analyzing both between-language distances as well as overall typological feature coverage.

To estimate **language distances**, we use the `lang2vec` toolkit (Littell et al., 2017) which contains precomputed language distances based on the typological language information queried from the URIEL knowledge base.[5] Following Ploeger et al. (2024), we calculate the distances between all pairs of languages in our dataset that have at least 5% coverage in the URIEL vectors. Figure 2 shows the distribution of pairwise geographic, syntactic, and genetic distances of all covered languages. We find that the languages in MULTIQ cover a wide range of typologically similar and distant language pairs with an expected high skewness of genetic distance complementing previous research (Ploeger et al., 2024). We can therefore confidently speak of a high typological diversity of our dataset.

Next, we calculate the **typological feature coverage** of the 137 languages in MULTIQ, using language features provided in the Grambank database (Skirgård et al., 2023). For this purpose, we map the language IDs in MULTIQ (obtained from the Google Translate API) to the Glottoids used by Grambank.[6] Taken together, the languages in MULTIQ cover at least 95.4% of the typological features recoded in Grambank. This underlines the typological diversity of our dataset.

Finally, we classify each language by its **language family** using the World Atlas of Language Structure (WALS).[7] In total, the 137 languages in MULTIQ belong to 20 different language fami-

[5] https://www.cs.cmu.edu/~dmortens/projects/7_project/

[6] For 14 languages in our dataset there is no matching entry in Grambank. To avoid incorrect mapping, we do not assign them manually and exclude these languages from the calculation

[7] https://wals.info/languoid

lies.[8] For additional details on the languages in our dataset, see Appendix A.

3 Experiments and Results

Using MULTIQ, we can now answer our two main research questions regarding the **multilingual language fidelity** and **multilingual QA accuracy** of current chat-optimized open LLMs. We first separately assess fidelity and accuracy, and then evaluate their relationship.

3.1 Overall Experimental Setup

Models We test six open-access LLMs that are both popular and competitive in performance with other state-of-the-art models as measured on standard (English-language) benchmarks such as the LMSys Leaderboard[9] and AlpacaEval.[10] Specifically, the 7B, 13B and 70B versions of Llama2-Chat (Touvron et al., 2023), the 7B Mistral-Instruct-v0.1 (Jiang et al., 2023), the 8x7B Mixtral-Instruct-v0.1 (Jiang et al., 2024) and the 7B Qwen1.5-Chat model (Bai et al., 2023). We test three sizes of Llama2 to evaluate scaling on MULTIQ. Llama2 and Mistral are explicitly intended for English use only, whereas Mixtral and Qwen are explicitly multilingual: Mixtral *"handles English, French, Italian, German and Spanish"*, while Qwen offers unspecified *"multilingual support"*.

Inference We run all models on two A100 GPUs using the `simplegen` Python library (Attanasio, 2023). We use default generation parameters from the `transformers` library, except for temperature, which we set to 0 to make completions deterministic. The maximum length of generations is 256 tokens. We do not use any system prompts. When prompting models with MULTIQ questions, we do not provide any additional context or examples.

3.2 Language Fidelity

We gather responses from all six models described above on the 27,400 questions in MULTIQ and then use GlotLID (Kargaran et al., 2023) to identify the response language. GlotLID is an open-source language identification model that supports more than 1,600 languages. GlotLID returns `iso_636_9` language codes, which we manually map to the language codes in MULTIQ.[11] Two languages in MULTIQ, namely Meiteilon (Manipuri) and Dogri, are not supported by GlotLID, so we exclude them from our language fidelity analysis. Figure 3 shows high-level results on language fidelity, split by how often models responded in the language of the input prompt, in English, or another language.

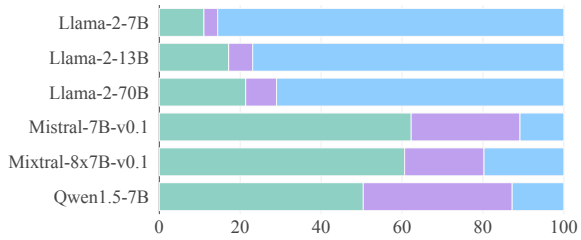

Figure 3: Overall language fidelity. Proportion of model responses (%) in the same language as the input prompt, in English, or in another language. We evaluate the responses of six models for 200 prompts in 135 languages (excl. Dogri & Meiteilon).

We find that the Llama2 models show a very low language fidelity, responding predominantly in English, matching its intended use for English only. Fidelity increases with scale, but even Llama2 70b, which gives 21.4% answers in the prompt language, is much less faithful than the other models. Surprisingly, Mistral, also intended for English use only, shows the greatest language fidelity, giving 62.3% of answers in the prompt language. Mistral is closely followed by Mixtral (60.6%) and Qwen (50.4%), which are advertised as having multilingual capabilities. Interestingly, compared to Llama2, the other models more frequently opt neither for English nor the prompt language, but some other language in their response. This effect appears most evident for Qwen.

To confirm the robustness of our findings, we investigate the impact of brief and numerical responses from the models on our results. To this end, we excluded all questions from the MULTIQ dataset that required a numerical answer, specifically removing the 10 curated questions of the domain "math" as well as 16 questions that were drawn from the LMSYS dataset. We then analyzed the character length of the model responses, noting an average length of 270-670 characters across different models. We also exclude responses shorter than 10 characters from the language fidelity calculation. The overall language fidelity of the models

[8] For languages that cannot be found in WALS (e.g. Corsican), we manually look up the language family in Grambank.

[9] https://huggingface.co/spaces/lmsys/chatbot-arena-leaderboard

[10] https://tatsu-lab.github.io/alpaca_eval/

[11] For 13 languages, several language codes in MULTIQ map to just one `iso_636_9` code. For details see Appendix B.

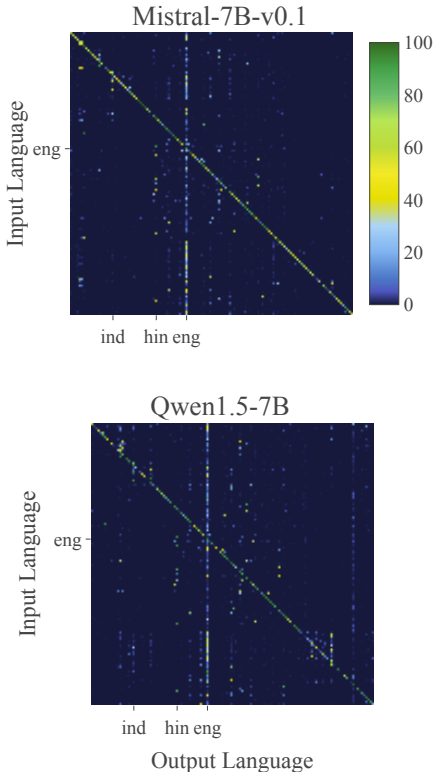

Figure 4: Granular language fidelity. Correlation matrices illustrating the relationship between input prompt and model response languages, shown as percentages. Axis ticks are selectively labeled for better visualization.

Model	Label	P	R	F1
Mistral-7B	Incorrect A.	0.97	1.00	0.98
	Correct A.	0.98	0.87	0.92
Qwen1.5-7B	Incorrect A.	0.88	0.99	0.93
	Correct A.	0.94	0.61	0.74
Llama-2-7b	Incorrect A.	0.92	1.00	0.96
	Correct A.	1.00	0.74	0.85
Mixtral-8x7B	Incorrect A.	0.82	0.98	0.89
	Correct A.	0.97	0.71	0.82

Table 2: Evaluation of the GPT-4 classifier that we use to assess answer accuracy on MULTIQ. We show Precision (P), Recall (R), and F1-Score of GPT-4 judgments on responses from four models to the same sample of 282 questions covering all languages, which were annotated by humans for whether they are correct or not.

showed only marginal changes in response to these changes, underscoring the minimal influence of brief or numerical responses on our analysis.

Next, we conduct a more fine-grained analysis of model answers in "another language", i.e. neither the input language nor English. For this, we focus on the three models responding in another language more than 10% of the time (Qwen, Mistral, Mixtral). Of these, Mistral demonstrates the highest level of diversity among the languages it responds in. GlotLID identifies more than 360 distinct languages in Mistral's responses, compared to around 240 languages for Mixtral and Qwen. We also examine the correspondence between the language family of input and output language. While Mixtral provides a response in the same language family at least 68.7% of the time, if not in an exact match or responding in English, Qwen demonstrates this behavior only 60.6% of the time, while Mistral does so for 54.6% of prompts. Note, however, that for 6% of the answers of Mistral and 7% of the answers of Mixtral, the language family of the response languages could not be determined and is thus classified as 'unknown'. Additionally, the distribution of language families in MULTIQ is highly skewed, with 45.3% of the languages belonging to the 'Indo-European' family, which spans a very broad range of languages from Irish to Turkish.

Finally, we analyze the relationship between the language of the prompt and the frequency of the respective response language of the models. Figure 4 shows correlation matrices for Mistral and Qwen. Corresponding matrices for the other models are shown in Appendix C. We find that Hindi is the most frequently selected language outside of the input language or English, closely followed by Indonesian. For Qwen, for example, these languages make up 20.4% of the "another language" category. This is visible by the thin yellow vertical lines leading to the axis ticks *ind* and *hin* in the Figure for both models respectively. A potential explanation may be that the models lack support for numerous languages from India and Indonesia, thus treating Hindi and Indonesian as some kind of 'fallback' languages for the wider language area. For example, the models often respond to languages such as Malay and Javanese in Indonesian, and languages such as Maithili, Konkani, and Bhojpuri in Hindi. We also observe this phenomenon for smaller European languages. For example, the three models respond to questions in Croatian mostly in Italian, to those in Luxembourgish mostly in German, and those in Galician mostly in Portuguese. These observations underline the importance of improving multilingual models in the language coverage of low-resource languages.

Model	All	EN	▲10	▲20	▲50
Qwen (7B)	16.6	84.0	61.6	50.2	34.3
Mistral (7B)	15.4	84.5	64.6	56.6	37.6
Mixtral (8x7B)	**33.8**	**90.5**	**87.7**	**84.8**	**67.4**
Llama2 (7B)	19.2	81.5	58.4	53.9	41.4
Llama2 (13B)	23.4	82.0	66.4	62.7	49.6
Llama2 (70B)	29.1	90.5	80.7	76.5	61.2

Table 3: QA accuracy on MULTIQ (%). We show accuracy overall, on English questions, and on the top(▲) 10, 20 and 50 best-performing languages for each model. Highest accuracy across models is **bold**.

Model	Same	English	Other
Qwen (7B)	**21.5**	11.4	11.7
Mistral (7B)	**17.1**	14.6	11.8
Mixtral (8x7B)	35.0	**37.0**	26.7
Llama2 (7B)	44.7	14.9	**46.5**
Llama2 (13B)	**51.6**	15.1	49.6
Llama2 (70B)	**61.4**	17.3	48.2

Table 4: QA accuracy on MULTIQ (%) split by language fidelity, i.e. which language models answered in (see Figure 3). Highest accuracy per model is **bold**.

3.3 Question Answering Accuracy

Next, we assess how often models give correct answers, regardless of whether the language of this answer matches the input prompt or not.

Automated Evaluation Since questions in MULTIQ are open-ended and answers can come in many languages, we need a flexible method for evaluating accuracy. We find that a carefully crafted prompt to GPT-4, which checks a given model answer against the English version of the question from MULTIQ, serves this purpose well.[12]

To evaluate the reliability of our automated evaluation method, we tasked two independent human annotators to label model responses from four of our models for the same 282 randomly selected prompts covering at least two questions per language, as correct or incorrect.[13] Disagreements between the annotators, which occurred for no more than four responses per model, were resolved in discussions with one of the authors. We find that the accuracy of our automated evaluation, as measured against the human labels, is very high across models (see Table 2). For all models, the automated evaluation tends to be very precise on correct answers, but less so on incorrect answers. This means that the automated evaluation will likely underestimate the proportion of correct model answers across languages. Overall, automated evaluation, like automated translation, introduces noise to our silver standard MULTIQ benchmark, but we find the amount of noise to likely be small.

Results Based on our automated evaluation, we calculate the proportion of correctly answered questions for each model across all 137 languages. Table 3 shows overall results, and Figure 5 shows a breakdown across all languages. We find that Mixtral is most accurate overall, but also across most individual languages covered in MULTIQ. Mistral shows the lowest overall accuracy across all languages, with a long tail distribution starting to decrease after the best-performing 20 languages. Moreover, a direct comparison between the results of Mixtral and Qwen shows that they achieve very similar results on Chinese, although only Qwen was explicitly trained in Chinese (Bai et al., 2023).

3.4 Language Fidelity vs Answer Accuracy

Finally, we combine the results of §3.2 and §3.3 to assess the relationship between language fidelity and QA accuracy. Table 4 shows the mean accuracy grouped by response language category. We find that Llama2, despite its overall low language fidelity, shows strong QA accuracy. Especially when answering in the same language as prompted, it gives a correct answer in almost half of the cases (44.7%). Scaling the model further improves accuracy, with the 13B and 70B variants achieving higher correct answer rates of 51.6% and 61.4%, respectively, when answering in the same language as prompted. Qwen and Mistral show a similar pattern, i.e. higher answering accuracy when answering in the prompt language with 21.5% and 17.1%. However, their overall accuracy is quite low, especially when compared to their initial high language fidelity. Strikingly, we find significant differences across models that ought to be similar due to their intended use for English, i.e. Llama2 demonstrating low fidelity yet high accuracy, in contrast to Mistral, which exhibits the reverse pattern. A qualitative analysis of Mistral's answers reveals that the model often merely repeats the questions it was asked, which is technically faithful but never an accurate answer (see Table 11). Only for Mixtral is this pattern not apparent, with the highest accuracy of 37% being achieved for responses in English.

[12] See Appendix B for the prompt template.

[13] We exclude the 13B and 70B versions of Llama2 from the QA accuracy analysis for reasons of clarity, and because they resemble the results of Llama2 7B.

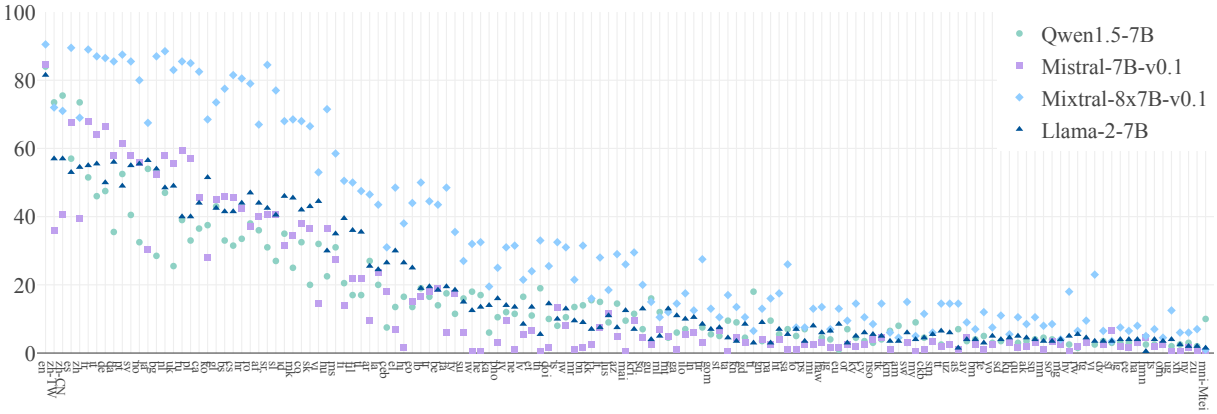

Figure 5: Answer accuracy on MULTIQ in proportion (%) of correctly answered questions per language. We compare four models of the same size across 137 languages and sort the results by median accuracy.

Overall, we find that if models answer in the same language as the prompted language, they tend to be more accurate than if they respond in English. The only exception of this pattern is Mixtral. Therefore, our results suggest that increased language fidelity may positively impact QA accuracy.

4 Tokenization and Multilinguality

Our results show significant variations in terms of language fidelity, QA accuracy, and the relationship between them, across the models we test with MULTIQ. This prompts us to investigate what factors may explain these differences. We focus on the differences between Mistral and Llama2 7B, which are particularly surprising given that both models are intended for use in English only. Prior research highlights the significant roles of tokenization and training data in multilingual capability (Dufter and Schütze, 2020; Clark et al., 2022; Petrov et al., 2023). Since there is little to no public information on the training data of the models that we test, we focus on their tokenizers.

Background: Byte Pair Encoding Both Llama2 and Mistral employ a tokenizer that uses byte pair encoding (BPE, Sennrich et al., 2016), streamlining sequence encoding by minimizing token count through identifying common subwords. This allows for frequent sequences to be represented through subword tokens, whereas rare or unseen sequences make the model default to individual characters or, failing that, ASCII code tokens.

Unique Tokens in MultiQ First, we evaluate how different tokenization strategies impact prompt tokenization in MULTIQ. We observe a significant deviation in the number of unique tokens used to represent all prompts in MULTIQ: Mistral uses 9,933 unique tokens, contrasting with Llama2's 10,676. This suggests that Llama2 may be less efficient at segmenting the typologically diverse MULTIQ dataset into a smaller number of subwords.

Tokenization Strategies For each model, we group languages into three categories depending on the model's tokenization strategy for the respective language. We develop a heuristic that allows us to classify languages into three tokenization categories: "ASCII", "character" or "subword". We do so by compiling the model's 20 most commonly used tokens for representing each language in MULTIQ, removing noise tokens (e.g. sentence start) from the list, and then quantifying the prevalence of ASCII tokens and characters based on the token-id ranges they usually occupy. Languages with over 70% of tokens in the ASCII or character categories were classified accordingly. For Llama2, out of 137 languages, we find 89 subword, 36 character and 12 ASCII languages. Mistral has 37 character and 9 ASCII languages. The models differ in the tokenization of some symbolic languages such as Chinese (ZH, see Figure 6): while Mistral mostly uses individual character tokens, Llama2 more frequently resorts to ASCII tokens due to its limited Chinese token vocabulary, which limits its ability to effectively tokenize Chinese language prompts.

Tokenization vs. QA Accuracy We evaluate the models' average QA accuracy across tokenization categories, finding a clear hierarchy (see Table 5): Subword encoding outperforms character and ASCII encodings, with Mistral and Llama2 re-

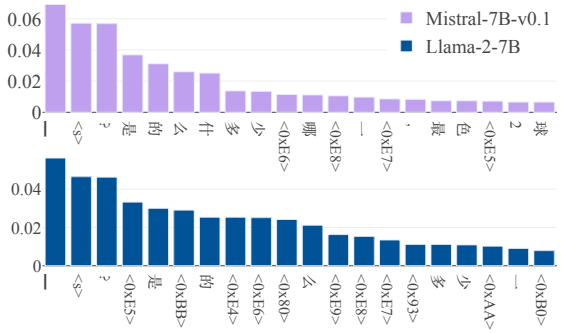

Figure 6: Tokenization analysis. Distribution of the most used unique tokens across all n=200 MULTIQ prompts in Chinese (ZH), when using the tokenizers of Llama2 (7B) and Mistral (7B).

Model	Subword	Character	ASCII
Llama2 (7B)	**24.0**	11.2	8.0
Mistral (7B)	**20.4**	6.4	2.1

Table 5: QA accuracy (%) on MULTIQ split by tokenization strategy. Highest accuracy per model is **bold**.

spectively achieving 20.4% and 24.0% accuracy on subword-encoded languages compared to just 6.4% and 11.4% on character-encoded languages. Even though our heuristic may introduce some noise in classifying tokenization strategies, we believe that exploring tokenization optimization is a promising direction for multilingual research, which also aligns with the broader discourse on the impact of tokenization on model capabilities. We hope our insights can help motivate further research into alternative language representation strategies (e.g. pixel-based models, Salesky et al., 2023).

5 Related Work

We discuss the related literature with respect to (i) large multilingual benchmarks, (ii) explicitly multilingual models, and (iii) multilingual studies of monolingual models.

Multilingual Benchmarks Existing multilingual benchmarks primarily target the performance of fully or partially supervised models on collections of standard NLP tasks, like XNLI (Conneau et al., 2018). Popular examples are the XTREME benchmarks (e.g. Ruder et al., 2021, 2023), and XGLUE (Liang et al., 2020). Similarly, researchers also presented benchmarks designed for languages spoken in particular regions, e.g. IndicXTREME (Doddapaneni et al., 2023) for Indic languages, Masakhan-NER (Adelani et al., 2022), and TaTA (Gehrmann et al., 2023) covering African languages, as well as NusaX (Winata et al., 2023) for Indonesian languages. Recently, Ahuja et al. (2023) proposed MEGA for evaluating multilingual generative models, which they use to evaluate LLMs like GPT-4 on a set of standard tasks. In a similar vein, Asai et al. (2023) presented BUFFET and benchmark LLMs for few-shot transfer. Another line of work focuses on multilingual QA datasets for reading comprehension, such XQuaD Artetxe et al. (2020), TyDiQA (Clark et al., 2020) and the Belebele benchmark (Bandarkar et al., 2023) with up to 122 diverse languages. Concurrent work proposes a new multilingual instruction tuning dataset called Aya (Singh et al., 2024), which also covers a range of open-ended questions in 114 languages. By comparison, MULTIQ contains parallel questions in a larger set of 137 languages covering 95.4% of Grambank features, which demonstrates its typological diversity and allows the analysis of multilingual LLM behavior at the margins of language coverage. Furthermore, MULTIQ's carefully selected short and simple questions target basic LLM knowledge to test only their multilingual capabilities and not ancillary factors such as complex reasoning.

Multilinguality in Monolingual LLMs Given the limited availability of open multilingual chat models, we are especially interested in assessing the multilinguality of models intended for English use only. Blevins and Zettlemoyer (2022) explained this behavior through data contamination: while the vast majority of the pre-training data of those models is English (e.g. ~93% for GPT-3 Brown et al., according to 2020, and ~90% for Llama-2 according to Touvron et al., 2023) there are also small portions of non-English content in the pre-training data. In such cases, it appears that the dominant language can help "unlock" the models' capabilities for the underrepresented languages (Gogoulou et al., 2022). Consequently, recent research assesses the multilinguality of several English-centric models, like GPT-3 and ChatGPT (e.g. Zhang et al., 2023; Lai et al., 2023; Armengol-Estapé et al., 2022; Winata et al., 2021). However, these prior works focus on standard NLP tasks such as text classification, just one or few models, and on just a few major languages. By contrast, we test the multilingual behavior of six LLMs in a more natural open-ended QA setting.

6 Conclusion

We introduced MULTIQ, a new silver standard benchmark for open-ended question answering that covers 137 typologically diverse languages. With MULTIQ, we evaluated the basic multilingual capabilities of six current, chat-optimized open LLMs, which are restricted in their intended use to just one or a small handful of languages. Our analysis focused on two key dimensions of multilingual capability – language fidelity and QA accuracy – and how they relate to each other. We found that all LLMs we test respond faithfully and/or accurately for at least some languages beyond their intended use. Most models are more accurate when they respond faithfully. However, we found that differences across models are large, and that there is a long tail of languages where models are neither accurate nor faithful. Finally, we indentified differences in tokenization as a potential explanation for our results. Overall, we hope that our findings can motivate further research into improving the multilingual capabilities of open LLMs, especially for diverse and under-represented languages, so that language technologies can benefit everyone, regardless of which language they speak.

Limitations

Our work comes with several limitations, which have already been partially discussed throughout the paper. First, we automatically translate our dataset, which introduces noise in the dataset. However, based on manual validation by native speakers, translation quality is high. Second, the automated evaluation of the models' answering accuracy using GPT-4, as well as response language classification using GlotLID, introduce some noise into the results. Here, too, human validation assures us that these factors minimally impact the results, potentially leading to an underestimation of answer accuracy. Third, mapping Google Translate IDs to the ISO language codes of GlotLID was not always directly possible, but we excluded possible inaccuracies from our calculations. Lastly, our analysis intentionally concentrated on basic multilingual capabilities, excluding the assessment of advanced reasoning or formulation skills.

Ethical Considerations

Intended Use As we emphasized throughout our paper, MULTIQ is intended to test *basic* multilingual capabilities. Therefore, good performance on MULTIQ alone should not be used as evidence for an LLM being suitable for specific languages.

Acknowledgements

The work of Carolin Holtermann and Anne Lauscher is funded under the Excellence Strategy of the German Federal Government and the States. Paul Röttger is a member of the Data and Marketing Insights research unit of the Bocconi Institute for Data Science and Analysis, and is supported by a MUR FARE 2020 initiative under grant agreement Prot. R20YSMBZ8S (INDOMITA). We thank our native-language annotators for their valuable work.

References

David Adelani, Graham Neubig, Sebastian Ruder, Shruti Rijhwani, Michael Beukman, Chester Palen-Michel, Constantine Lignos, Jesujoba Alabi, Shamsuddeen Muhammad, Peter Nabende, Cheikh M. Bamba Dione, Andiswa Bukula, Rooweither Mabuya, Bonaventure F. P. Dossou, Blessing Sibanda, Happy Buzaaba, Jonathan Mukiibi, Godson Kalipe, Derguene Mbaye, Amelia Taylor, Fatoumata Kabore, Chris Chinenye Emezue, Anuoluwapo Aremu, Perez Ogayo, Catherine Gitau, Edwin Munkoh-Buabeng, Victoire Memdjokam Koagne, Allahsera Auguste Tapo, Tebogo Macucwa, Vukosi Marivate, Mboning Tchiaze Elvis, Tajuddeen Gwadabe, Tosin Adewumi, Orevaoghene Ahia, Joyce Nakatumba-Nabende, Neo Lerato Mokono, Ignatius Ezeani, Chiamaka Chukwuneke, Mofetoluwa Oluwaseun Adeyemi, Gilles Quentin Hacheme, Idris Abdulmumin, Odunayo Ogundepo, Oreen Yousuf, Tatiana Moteu, and Dietrich Klakow. 2022. MasakhaNER 2.0: Africa-centric transfer learning for named entity recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4488–4508, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: Multilingual evaluation of generative AI. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267, Singapore. Association for Computational Linguistics.

Jordi Armengol-Estapé, Ona de Gibert Bonet, and Maite Melero. 2022. On the multilingual capabilities of very large-scale English language models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3056–3068, Marseille, France. European Language Resources Association.

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*, pages 4623–4637, Online. Association for Computational Linguistics.

Akari Asai, Sneha Kudugunta, Xinyan Velocity Yu, Terra Blevins, Hila Gonen, Machel Reid, Yulia Tsvetkov, Sebastian Ruder, and Hannaneh Hajishirzi. 2023. Buffet: Benchmarking large language models for few-shot cross-lingual transfer.

Giuseppe Attanasio. 2023. Simple Generation. https://github.com/MilaNLProc/simple-generation.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2023. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants.

Emily M Bender. 2011. On achieving and evaluating language-independence in nlp. *Linguistic Issues in Language Technology*, 6.

Terra Blevins and Luke Zettlemoyer. 2022. Language contamination helps explains the cross-lingual capabilities of English pretrained models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3563–3574, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

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

Jonathan H. Clark, Dan Garrette, Iulia Turc, and John Wieting. 2022. Canine: Pre-training an efficient tokenization-free encoder for language representation. *Transactions of the Association for Computational Linguistics*, 10:73–91.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2023. Towards leaving no Indic language behind: Building monolingual corpora, benchmark and models for Indic languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12402–12426, Toronto, Canada. Association for Computational Linguistics.

Philipp Dufter and Hinrich Schütze. 2020. Identifying elements essential for BERT's multilinguality. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4423–4437, Online. Association for Computational Linguistics.

Sebastian Gehrmann, Sebastian Ruder, Vitaly Nikolaev, Jan Botha, Michael Chavinda, Ankur Parikh, and Clara Rivera. 2023. TaTA: A multilingual table-to-text dataset for African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1719–1740, Singapore. Association for Computational Linguistics.

Evangelia Gogoulou, Ariel Ekgren, Tim Isbister, and Magnus Sahlgren. 2022. Cross-lingual transfer of monolingual models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 948–955, Marseille, France. European Language Resources Association.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and

fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Amir Hossein Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schütze. 2023. Glotlid: Language identification for low-resource languages. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Viet Lai, Nghia Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Nguyen. 2023. ChatGPT beyond English: Towards a comprehensive evaluation of large language models in multilingual learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13171–13189, Singapore. Association for Computational Linguistics.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. 2020. XGLUE: A new benchmark dataset for cross-lingual pre-training, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6008–6018, Online. Association for Computational Linguistics.

Patrick Littell, David R Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. Uriel and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, volume 2, pages 8–14.

OpenAI. 2023. Gpt-4 technical report. *preprint*.

Siru Ouyang, Shuohang Wang, Yang Liu, Ming Zhong, Yizhu Jiao, Dan Iter, Reid Pryzant, Chenguang Zhu, Heng Ji, and Jiawei Han. 2023. The shifted and the overlooked: A task-oriented investigation of user-GPT interactions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2375–2393, Singapore. Association for Computational Linguistics.

Aleksandar Petrov, Emanuele La Malfa, Philip H. S. Torr, and Adel Bibi. 2023. Language model tokenizers introduce unfairness between languages.

Esther Ploeger, Wessel Poelman, Miryam de Lhoneux, and Johannes Bjerva. 2024. What is 'typological diversity' in nlp?

Sebastian Ruder, Jonathan Clark, Alexander Gutkin, Mihir Kale, Min Ma, Massimo Nicosia, Shruti Rijhwani, Parker Riley, Jean-Michel Sarr, Xinyi Wang, John Wieting, Nitish Gupta, Anna Katanova, Christo Kirov, Dana Dickinson, Brian Roark, Bidisha Samanta, Connie Tao, David Adelani, Vera Axelrod, Isaac Caswell, Colin Cherry, Dan Garrette, Reeve Ingle, Melvin Johnson, Dmitry Panteleev, and Partha Talukdar. 2023. XTREME-UP: A user-centric scarce-data benchmark for under-represented languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1856–1884, Singapore. Association for Computational Linguistics.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, and Melvin Johnson. 2021. XTREME-R: Towards more challenging and nuanced multilingual evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10215–10245, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Elizabeth Salesky, Neha Verma, Philipp Koehn, and Matt Post. 2023. Multilingual pixel representations for translation and effective cross-lingual transfer. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13845–13861, Singapore. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Shivalika Singh, Freddie Vargus, Daniel Dsouza, Börje F. Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura OMahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Souza Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergün, Ifeoma Okoh, Aisha Alaagib, Oshan Mudannayake, Zaid Alyafeai, Vu Minh Chien, Sebastian Ruder, Surya Guthikonda, Emad A. Alghamdi, Sebastian Gehrmann, Niklas Muennighoff, Max Bartolo, Julia Kreutzer, Ahmet Üstün, Marzieh Fadaee, and Sara Hooker. 2024. Aya dataset: An open-access collection for multilingual instruction tuning.

Hedvig Skirgård, Hannah J. Haynie, Damián E. Blasi, Harald Hammarström, Jeremy Collins, Jay J. Latarche, Jakob Lesage, Tobias Weber, Alena Witzlack-Makarevich, Sam Passmore, Angela Chira, Luke Maurits, Russell Dinnage, Michael Dunn, Ger Reesink, Ruth Singer, Claire Bowern, Patience Epps, Jane Hill, Outi Vesakoski, Martine Robbeets, Noor Karolin Abbas, Daniel Auer, Nancy A. Bakker, Giulia Barbos, Robert D. Borges, Swintha Danielsen, Luise Dorenbusch, Ella Dorn, John Elliott, Giada Falcone, Jana Fischer, Yustinus Ghanggo Ate, Hannah Gibson, Hans-Philipp Göbel, Jemima A. Goodall, Victoria Gruner, Andrew Harvey, Rebekah Hayes, Leonard Heer, Roberto E. Herrera Miranda, Nataliia Hübler, Biu Huntington-Rainey,

Jessica K. Ivani, Marilen Johns, Erika Just, Eri Kashima, Carolina Kipf, Janina V. Klingenberg, Nikita König, Aikaterina Koti, Richard G. A. Kowalik, Olga Krasnoukhova, Nora L.M. Lindvall, Mandy Lorenzen, Hannah Lutzenberger, Tânia R.A. Martins, Celia Mata German, Suzanne van der Meer, Jaime Montoya Samamé, Michael Müller, Saliha Muradoglu, Kelsey Neely, Johanna Nickel, Miina Norvik, Cheryl Akinyi Oluoch, Jesse Peacock, India O. C. Pearey, Naomi Peck, Stephanie Petit, Sören Pieper, Mariana Poblete, Daniel Prestipino, Linda Raabe, Amna Raja, Janis Reimringer, Sydney C. Rey, Julia Rizaew, Eloisa Ruppert, Kim K. Salmon, Jill Sammet, Rhiannon Schembri, Lars Schlabbach, Frederick W.P. Schmidt, Amalia Skilton, Wikaliler Daniel Smith, Hilário de Sousa, Kristin Sverredal, Daniel Valle, Javier Vera, Judith Voß, Tim Witte, Henry Wu, Stephanie Yam, Jingting Ye, Maisie Yong, Tessa Yuditha, Roberto Zariquiey, Robert Forkel, Nicholas Evans, Stephen C. Levinson, Martin Haspelmath, Simon J. Greenhill, Quentin D. Atkinson, and Russell D. Gray. 2023. Grambank reveals the importance of genealogical constraints on linguistic diversity and highlights the impact of language loss. *Science Advances*, 9(16):eadg6175.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Genta Indra Winata, Alham Fikri Aji, Samuel Cahyawijaya, Rahmad Mahendra, Fajri Koto, Ade Romadhony, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Pascale Fung, Timothy Baldwin, Jey Han Lau, Rico Sennrich, and Sebastian Ruder. 2023. NusaX: Multilingual parallel sentiment dataset for 10 Indonesian local languages. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 815–834, Dubrovnik, Croatia. Association for Computational Linguistics.

Genta Indra Winata, Andrea Madotto, Zhaojiang Lin, Rosanne Liu, Jason Yosinski, and Pascale Fung. 2021. Language models are few-shot multilingual learners. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 1–15, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xiang Zhang, Senyu Li, Bradley Hauer, Ning Shi, and Grzegorz Kondrak. 2023. Don't trust ChatGPT when your question is not in English: A study of multilingual abilities and types of LLMs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7915–7927, Singapore. Association for Computational Linguistics.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. 2024. (inthe)wildchat: 570k chatGPT interaction logs in the wild. In *The Twelfth International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric Xing, Joseph E. Gonzalez, Ion Stoica, and Hao Zhang. 2024. LMSYS-chat-1m: A large-scale real-world LLM conversation dataset. In *The Twelfth International Conference on Learning Representations*.

Appendix

A MultiQ

In total MULTIQ covers 10 different question domains and 20 distinct language families.

Domains

- chemistry
- physics
- astronomy
- history
- maths
- geography
- art
- sports
- music
- animals

Language Families

- Afro-Asiatic (AA)
- Altaic (Al)
- Austro-Asiatic (AuA)
- Austronesian (Au)
- Aymaran (Ay)
- Basque (B)
- Dravidian (D)
- Hmong-Mien (HM)
- Indo-European (IE)
- Japanese (J)
- Kartvelian (K)
- Korean (Ko)
- Mande (M)
- Niger-Congo (NC)
- Other (O)
- Quechuan (Qu)
- Sino-Tibetan (ST)
- Tai-Kadai (TK)
- Tupian (T)
- Uralic (U)

Language	Prop. of correct questions
Arabic	0.822
Catalan	0.899
Chinese	0.955
Farsi	0.890
French	0.915
German	0.995
Hindi	0.975
Indonesian	0.889
Italian	0.990
Japanese	0.955
Korean	0.97
Spanish	0.950
Tagalog	0.6
Quechua	0.895
Romanian	0.875
Russian	0.97
Ukrain	0.935
Urdu	0.935
Xhosa	0.990

Table 6: Validation Results on MULTIQ. We present the proportion of correctly translated prompts for each language assessed by native speakers in the respective language.

Code	Language	Family	Code	Language	Family	Code	Language	Family
af	Afrikaans	IE	hmn	Hmong	HM	ny	Nyanja (Chichewa)	NC
ak	Twi (Akan)	NC	hr	Croatian	IE	om	Oromo	AA
am	Amharic	AA	ht	Haitian Creole		or	Odia (Oriya)	IE
ar	Arabic	AA	hu	Hungarian	U	pa	Panjabi	IE
as	Assamese	IE	hy	Armenian	IE	pl	Polish	IE
ay	Aymara	Ay	id	Indonesian	Au	ps	Pashto	IE
az	Azerbaijani	Al	ig	Igbo	NC	pt	Portuguese	IE
be	Belarusian	IE	ilo	Ilocano	Au	qu	Quechua	Qu
bg	Bulgarian	IE	is	Icelandic	IE	ro	Romanian	IE
bho	Bhojpuri	IE	it	Italian	IE	ru	Russian	IE
bm	Bambara	M	iw	Hebrew alternativ	AA	rw	Kinyarwanda	NC
bn	Bengali	IE	ja	Japanese	J	sa	Sanskrit	
bs	Bosnian	IE	jv	Javanese alternativ	Au	sd	Sindhi	IE
ca	Catalan	IE	jw	Javanese	Au	si	Sinhala (Sinhalese)	IE
ceb	Cebuano	Au	ka	Georgian	K	sk	Slovak	IE
ckb	Kurdish (Sorani)	IE	kk	Kazakh	Al	sl	Slovenian	IE
co	Corsican	IE	km	Khmer	AuA	sm	Samoan	Au
cs	Czech	IE	kn	Kannada	D	sn	Shona	NC
cy	Welsh	IE	ko	Korean	K	so	Somali	AA
da	Danish	IE	kri	Krio		sq	Albanian	IE
de	German	IE	ku	Kurdish	IE	sr	Serbian	IE
doi	Dogri	IE	ky	Kyrgyz	Al	st	Sesotho	NC
dv	Dhivehi	IE	la	Latin		su	Sundanese	Au
ee	Ewe	NC	lb	Luxembourgish	IE	sv	Swedish	IE
el	Greek	IE	lg	Luganda	NC	sw	Swahili	NC
en	English	IE	ln	Lingala	NC	ta	Tamil	D
eo	Esperanto	IE	lo	Lao	TK	te	Telugu	D
es	Spanish	IE	lt	Lithuanian	IE	tg	Tajik	IE
et	Estonian	U	lus	Mizo	ST	th	Thai	TK
eu	Basque	B	lv	Latvian	IE	ti	Tigrinya	AA
fa	Persian	IE	mai	Maithili	IE	tk	Turkmen	Al
fi	Finnish	U	mg	Malagasy	Au	tl	Tagalog (Filipino)	Au
fil	Filipino (Tagalog)	Au	mi	Maori	Au	tr	Turkish	Al
fr	French	IE	mk	Macedonian	IE	ts	Tsonga	NC
fy	Frisian	IE	ml	Malayalam	D	tt	Tatar	Al
ga	Irish	IE	mn	Mongolian		ug	Uyghur	Al
gd	Scots Gaelic	IE	mni-Mtei	Meiteilon (Manipuri)	ST	uk	Ukrainian	IE
gl	Galician	IE	mr	Marathi	IE	ur	Urdu	IE
gn	Guarani	T	ms	Malay	Au	uz	Uzbek	Al
gom	Konkani	IE	mt	Maltese	AA	vi	Vietnamese	AuA
gu	Gujarati	IE	my	Myanmar (Burmese)	ST	xh	Xhosa	NC
ha	Hausa	AA	ne	Nepali	IE	yi	Yiddish	IE
haw	Hawaiian	Au	nl	Dutch	IE	yo	Yoruba	NC
he	Hebrew	AA	no	Norwegian	IE	zh	Chinese (Trad.)	
hi	Hindi	IE	nso	Sepedi	NC	zh-CN	Chinese (Simpl.)	
			zh-TW	Chinese (Simpl.)		zu	Zulu	NC

Table 7: 137 Languages covered by MULTIQ, we present their Google Translate Code and the acronym of their language families.

B Experimental Setup

In Table 8, we provide all information and links to the data, models, and frameworks, we use in our work. All artifacts were used according to their intended use, as described in their licenses. As described in the main body of this manuscript, we are also releasing our code publicly and will do so under the CC-BY-4.0 license.

Purpose	Name	Vocabsize	Intended Languages
Models	`meta-llama/Llama-2-7b-chat-hf`	32000	en
	`meta-llama/Llama-2-13b-chat-hf`	32000	en
	`mistralai/Mistral-7B-Instruct-v0.1`	32000	en
	`Qwen/Qwen1.5-7B-Chat`	151936	undefined
	`mistralai/Mixtral-8x7B-Instruct-v0.1`	32000	fr, en, es, it, de
	`cis-lmu/glotlid`		1600 langs
	`gpt-4-0125-preview`		
Framworks	transformers==4.34.0		
	torch==2.1.0		
	huggingface-hub==0.17.3		
	tokenizers==0.14.1		
Datasets	WALS	Dataset License: CC-BY-4.0	
	Grambank	License CC-BY-4.0	

Table 8: Overview of models, datasets and frameworks used in our work.

Task	Prompt
Question Creation for MultiQ	Generate a list of 10 simple and clear *{domain}* questions with simple and clear answers.
Validate Answer Correctness	For the following question, is this a correct answer? Question: *{MultiQ Question in English}* Answer: *{Model Answer}* The question is in English. The answer is either in English or in another language. If the answer is correct, type 'yes'. If the answer is incorrect, type 'no'."

Table 9: Prompt templates used to prompt GPT-4 for question creation and QA accuracy validation respectively.

Google Translate ID	ISO_639_3
zh-CN zh zh-TW	zho
ku ckb	ckb
he iw	heb
sr hr	hbs
jv jw	jav
tl fil	tgl

Table 10: List of Languages in MULTIQ that map to only one iso code in the GlotLID model.

C Granular Language Fidelity Analysis

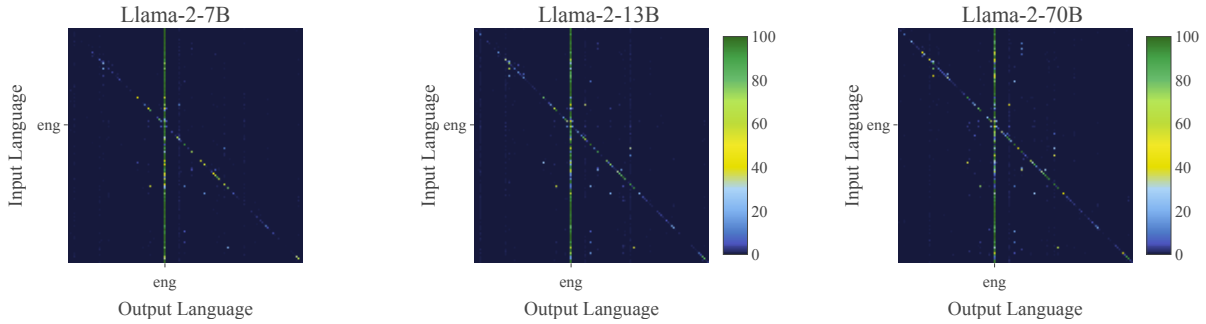

Figure 7: Correlation matrices illustrating the relationship between input question languages and model response languages for the Llama2 models, shown as percentages. Axis ticks are selectively labeled for clearer visualization.

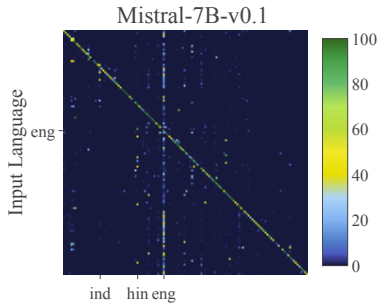

Figure 8: Correlation matrix illustrating the relationship between input question languages and model response languages for Mistral, shown as percentages. Axis ticks are selectively labeled for clearer visualization.

Prompt	Language	English Translation	Model Completion
Xana Misava yi na tin'hweti tingani?	ts	How many moons does Earth have?	Xana Misava yi na tin'hweti tingani? (What is Xana Misava's net worth?)
X'inhu sena dawl?	mt	What is a light-year?	X'inhu sena dawl? is a question in the Mayan language of Guatemala. It translates to "What is the meaning of life?" in English
Beth yw mynydd uchaf Japan?	cy	What is the highest mountain in Japan?	Beth yw mynydd uchaf Japan? (Do you speak Japanese?)
Quanto fa tre volte sette?	it	What is three times seven?	Quanto fa tre volte sette? Quanto fa tre volte sette? Quanto fa tre volte sette? Quanto fa tre volte sette? Quanto fa tre volte sette? Quanto fa tre volte sette? Quanto fa tre volte sette? Quanto fa tre volte sette? Quanto fa tre volte sette? [...]

Table 11: Examples of Fidelity Issues for Mistral 7B

D Answer Accuracy

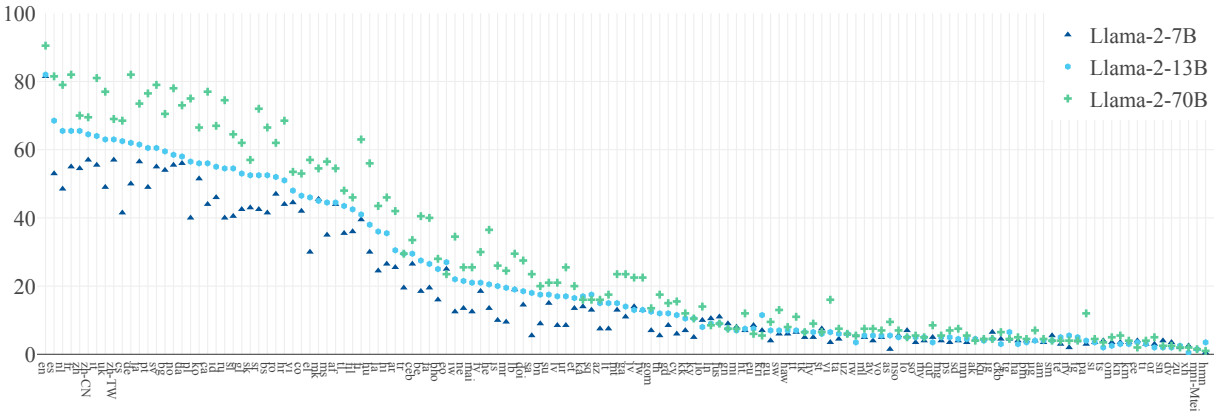

Figure 9: Answer Accuracy on MULTIQ in proportion (%) of correctly answered questions per language. We compare the Llama2 models in different model sizes across all 137 languages and sort the results by median accuracy.

E Correlation between Answer Accuracy and Language Fidelity