MEGAVERSE : Benchmarking Large Language Models Across Languages, Modalities, Models and Tasks

Sanchit Ahuja **Divyanshu Aggarwal** **Varun Gumma** **Ishaan Watts**
Ashutosh Sathe **Millicent Ochieng** **Rishav Hada** **Prachi Jain**
Mohamed Ahmed **Kalika Bali** **Sunayana Sitaram**
Microsoft Corporation
{t-sahuja,sunayana.sitaram}@microsoft.com

Abstract

There has been a surge in LLM evaluation research to understand LLM capabilities and limitations. However, much of this research has been confined to English, leaving LLM building and evaluation for non-English languages relatively unexplored. Several new LLMs have been introduced recently, necessitating their evaluation on non-English languages. This study aims to perform a thorough evaluation of the non-English capabilities of SoTA LLMs (GPT-3.5-Turbo, GPT-4, PaLM2, Gemini-Pro, Mistral, Llama2, and Gemma) by comparing them on the same set of multilingual datasets. Our benchmark comprises 22 datasets covering 83 languages, including low-resource African languages. We also include two multimodal datasets in the benchmark and compare the performance of LLaVA models, GPT-4-Vision and Gemini-Pro-Vision. Our experiments show that larger models such as GPT-4, Gemini-Pro and PaLM2 outperform smaller models on various tasks, notably on low-resource languages, with GPT-4 outperforming PaLM2 and Gemini-Pro on more datasets. We also perform a study on data contamination and find that several models are likely to be contaminated with multilingual evaluation benchmarks, necessitating approaches to detect and handle contamination while assessing the multilingual performance of LLMs.

1 Introduction

Large Language Models (LLMs) have surpassed the performance of previous generation of language models on several tasks and benchmarks, sometimes even approaching or exceeding human performance (Hubert et al., 2024). However, the root cause of the observed capabilities in these models is not always apparent, whether stemming from augmented model capabilities or other factors like contamination in test datasets and the absence of datasets that genuinely measure the capabilities of these models (Balloccu et al., 2024). Thus, evaluation of Large Language Models has become an important field of study.

Most of the work on evaluating LLMs via benchmarking (Liang et al., 2022), qualitative tests for specific capabilities (Bubeck et al., 2023) or human evaluation have focused solely on English. However, studies have shown that there is a large gap between the capabilities of LLMs in English and other languages (Choudhury et al., 2023). Evaluation of LLMs in languages other than English is challenging due to a variety of factors, including the lack of benchmarks covering a large number of languages from diverse language families and the lack of multilingual benchmarks covering tasks such as reasoning, chat, and dialogue. Therefore, it is crucial to prioritize multilingual evaluation to enhance the development of more effective multilingual models. Neglecting this critical aspect may result in a significant population being left behind and may widen the digital divide (Joshi et al., 2021).

Our prior work on evaluating multilingual capabilities of LLMs, MEGA (Ahuja et al., 2023), yielded the following observations: GPT-4 (OpenAI, 2023a) comes close to the performance of SOTA fine-tuned language models such as TULRv6 (Patra et al., 2023). GPT models perform worse on languages that are written in non-Latin scripts, and on low-resource languages. Other LLMs such as BLOOMZ (Muennighoff et al., 2023) usually perform worse than GPT-4. However, several newer models are comparable to GPT-4 in performance on English, and it is essential to study their multilingual performance as well. Moreover, there is a rising interest in Large Multimodal Models (LMMs), and the convergence of multimodal and multilingual LLMs remains an understudied area (Hu et al., 2024). Our contributions are as follows:

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 2598–2637

- We build on top of the MEGA benchmark and add 6 new datasets, thus extending coverage to 22 datasets and 83 languages including many low-resource African languages.
- We benchmark nine new SOTA text LLMs - PaLM2 (Google, 2023), Llama2 (3 variants) (Touvron et al., 2023), Mistral-v1.0 (2 variants), (Jiang et al., 2023), Gemma (2 variants) (Mesnard et al., 2024), Gemini 1.0 pro (Anil et al., 2023a) in addition to GPT-4 and GPT-3.5-Turbo.
- We benchmark the multimodal LLaVA family models (Liu et al., 2023), GPT-4-Vision (OpenAI, 2023b) and Gemini-Pro-Vision (Anil et al., 2023a) on two multilingual multimodal datasets.
- We present a thorough contamination study of both commercial and open-source set of LLMs on a subset of our datasets.
- We study the overall trends in our experiments by studying the deviation of performance across language families and tasks, and provide directions for future research.

2 Related work

Evaluation of LLMs Recently, there has been an increasing interest in evaluating LLMs on a wide range of capabilities, given the surge in their popularity and effectiveness. BIG-Bench (Srivastava et al., 2023) consists of 204 tasks to evaluate LLMs.

While BIG-Bench includes tasks in non-English languages as well, they are largely related to translation. Liang et al. (2022) proposed HELM, defining a taxonomy of scenarios and metrics that define the space of LLM evaluation, and evaluating 30 language models on 42 scenarios and 7 metrics. However, all the scenarios are focused on datasets in standard English or dialects, and they highlight coverage of languages as an important area for improvement. Bubeck et al. (2023), has pointed out the limitations of using standard NLP benchmarks to evaluate generative models, due to the pace at which these benchmarks become saturated. There are also concerns about benchmark contamination in LLM evaluation. Zhou et al. (2023) show that test dataset contamination in training and fine-tuning data leads to a significant impact on LLM performance.

Multilingual Benchmarks and Evaluation Bang et al. (2023) evaluates the multilingual capabilities of ChatGPT and shows that it fails to generalize to low-resource languages with non-Latin scripts. However, multilingual evaluation is performed only on a few tasks, and a subset of 50-100 examples are used for testing the model. Hendy et al. (2023) evaluate the translation abilities of GPT-3.5 models and find that these models perform well in translating high-resource languages, but their capabilities for low-resource languages are limited. BUFFET (Asai et al., 2023) covering 54 languages across 15 datasets and Lai et al. (2023) covering 37 languages across 7 datasets also perform multilingual benchmarking of LLMs such as ChatGPT and BLOOMZ. Yang et al. (2023) does a comprehensive study of GPT4-Vision's capabilities that include analyzing its performance on multilingual image description, scene text recognition, and translation. Our work builds on the MEGA benchmarking effort (Ahuja et al., 2023), which evaluates GPT models across 16 datasets. We extend the MEGA benchmark to more tasks including multimodal tasks, evaluate several SoTA LLMs, and perform a more comprehensive analysis of contamination.

Contamination Several techniques have been proposed to study the contamination of publicly available evaluation datasets. Ahuja et al. (2023) study contamination by prompting the models to fill dataset cards. Other methodologies encompass Golchin and Surdeanu (2023b), which does not provide quantification of contamination, and Oren et al. (2023), which requires access to log probabilities, thereby limiting their studies to open-sourced LLMs.

3 Experimental Setup

3.1 Datasets

We perform experiments on the 16 datasets that are part of the MEGA suite - XNLI (Conneau et al., 2018), IndicXNLI (Aggarwal et al., 2022), GLUECoS NLI (Khanuja et al., 2020a), PAWS-X (Yang et al., 2019), XCOPA (Ponti et al., 2020), XStoryCloze (Lin et al., 2022), GLUECoS Sentiment Analysis (En-Es-CS) (Vilares et al., 2016), TyDiQA-GoldP (Clark et al., 2020), MLQA (Lewis et al., 2020), XQUAD (Artetxe et al., 2020), IndicQA (Doddapaneni et al., 2023), PAN-X (Pan et al., 2017), UDPOS (Nivre et al., 2018), Jigsaw (Kivlichan et al., 2020), WinoMT (Stanovsky et al.,

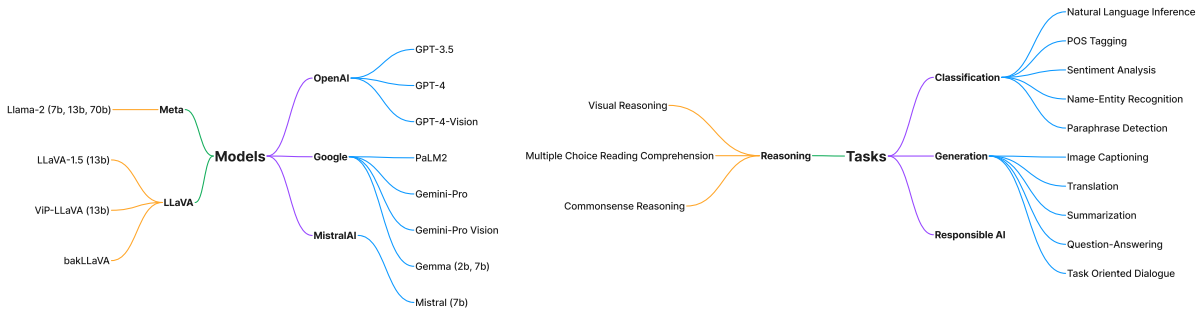

Figure 1: Hierarchy of Models and Tasks spread across MEGAVERSE

2019) and XLSum (Hasan et al., 2021). These datasets include a mix of classification, Question Answering, Sequence Labeling, and Natural Language Generation datasets, along with two datasets covering the Responsible AI tasks of toxicity detection and gender bias. The datasets we include also contain a mix of translated datasets verified by native speakers, as well as datasets created independently for each language. Figure 1 shows a hierarchy of models and tasks spread across MEGAVERSE. For a more detailed description of the datasets included in the original MEGA benchmark, we refer the readers to Ahuja et al. (2023). We describe the six datasets added to our study below.

3.1.1 AfriQA

AfriQA (Ogundepo et al., 2023) is a QA dataset that does not have a context passage. It covers 10 African languages - Bemba, Fon, Hausa, Igbo, Kinyarwanda, Swahili, Twi, Wolof, and Yorùbá. We use the few-shot size of $k = 4$ and the monolingual prompting strategy to perform experiments only on the GPT and Llama models, as the PaLM2 model only supports Swahili.

3.1.2 Belebele

Belebele (Bandarkar et al., 2023) is a multiple choice machine reading comprehension (MRC) dataset parallel across 122 languages. Each question is linked to a short passage from the FLORES-200 dataset (Costa-jussà et al., 2022). The human annotation procedure was carefully curated to create questions that discriminate between different levels of language comprehension. We evaluated Arabic, Czech, Danish, German, English, Spanish, Finnish, French, Hebrew, Hungarian, Italian, Japanese, Korean, Dutch, Norwegian, Polish, Portuguese, Russian, Swedish, Thai, Turkish, Chinese Simplified and Chinese Traditional. Results for Llama2 and GPT-3.5-Turbo are reported from the dataset paper. We perform zero-shot monolingual prompting for our experiments, as this dataset does not have a dev set.

3.1.3 IN22

IN22 (Gala et al., 2023) is a translation benchmark for all 22 scheduled Indic languages. IN22-Gen is a general-purpose multi-domain evaluation subset of IN22 which has been curated from two sources: Wikipedia and Web Sources offering diverse content spanning news, entertainment, culture, legal, and India-centric topics. IN22-Conv is the conversation domain subset of IN22. Due to resource constraints, we evaluate 14 languages: Assamese, Bengali, English, Gujarati, Hindi, Kannada, Kashmiri, Malayalam, Marathi, Nepali, Odia, Punjabi, Tamil, Telugu, and Urdu.

3.1.4 MaRVL

MaRVL (Multicultural Reasoning over Vision and Language) (Liu et al., 2021) is a dataset of images and associated captions. The concepts and images collected were entirely driven by native speakers and are representative of various cultures across the globe and span 5 languages, i.e., Indonesian, Chinese, Swahili, Tamil, and Turkish. Each instance in the dataset consists of a pair of images (left image and right image) and a statement, and the task is to determine whether the statement is consistent for the given pair of images.

3.1.5 XM-3600

CrossModal-3600 (Thapliyal et al., 2022) is a multilingual image captioning dataset consisting of 3600 geographically diverse images directly captioned in 36 different languages, avoiding any inconsistencies due to translations. We experimented on 20 out of 36 languages due to resource constraints:

Arabic, Chinese, Czech, Danish, Dutch, English, Finnish, French, German, Italian, Japanese, Korean, Norwegian, Polish, Portuguese, Russian, Spanish, Swedish, Thai, and Turkish.

3.1.6 XRiSAWOZ

XRiSAWOZ (Moradshahi et al., 2023) is a task-oriented dialogue modeling dataset. The dataset is a multilingual (English, Hindi, French, Korean) translation of the Chinese-only RiSAWOZ dataset (Quan et al., 2020). XRiSAWOZ also includes an **English-Hindi code mixed** setting. For each conversation, the agent must make use of structured knowledge from the databases to answer user queries. The task consists of 4 subtasks: "Dialogue State Tracking" (DST), "API Call Detection" (API), "Dialogue Act Generation" (DA) and "Response Generation" (RG). The metrics used for evaluation include BLEU, Slot Error Rate (SER) (factual correctness of generated response) (Wen et al., 2015), (averaged/task) success rate (Lin et al., 2021), API call accuracy, dialogue act accuracy and joint goal accuracy (Budzianowski et al., 2018). We refer the reader to Moradshahi et al. (2023) for detailed descriptions of subtasks and metrics. We perform experiments on 10% of the data i.e. about 400 dialogue turns across 3 domains due to limited compute.

3.2 Models[1]

Below is a list of all the models we evaluate:

- **GPT-3.5-Turbo** (Ouyang et al., 2022)
- **GPT-4** (OpenAI, 2023a)
- **GPT-4-Vision** (OpenAI, 2023b)
- **Llama2 (7B, 13B, 70B)** (Touvron et al., 2023)
- **PaLM2** (Anil et al., 2023b)
- **Gemini-Pro** (Anil et al., 2023a)
- **Gemini-Pro-Vision** (Anil et al., 2023a)
- **Gemma (2B, 7B)** (Mesnard et al., 2024)
- **Mistral** (Jiang et al., 2023)
- **BakLLaVA-v1** (Liu et al., 2023)
- **ViP-LLaVA (13B)** (Cai et al., 2023)
- **LLaVA-1.5 (13B)** (Liu et al., 2023)

3.3 Prompting strategies

Ahuja et al. (2023) explore three prompting variations based on the language of the few-shot and test examples, and find that monolingual prompting, featuring few-shot examples in the target language, outperforms zero-shot cross-lingual prompting in English for most datasets. Translate-test excels over monolingual for certain low-resource languages but with minimal gaps for models like GPT-4. Therefore, we default to monolingual prompting unless otherwise specified. Zero-shot cross-lingual prompting (zs-cl) is used when dev datasets are unavailable in the target language. English instructions are maintained for prompts, proven to outperform instructions in the target language (Ahuja et al., 2023). Prompt templates for our new datasets are in the Appendix A.2.

3.3.1 XRiSAWOZ

Moradshahi et al. (2023) presents results in both end-to-end and turn-by-turn evaluation settings. We perform end-to-end evaluation with regex based careful filtering of the generated responses for DST/API/DA tasks after every turn. This is required to ensure correctness of the syntax in the state descriptions for these tasks. No such postprocessing is done for the RG task. For inferring a subtask on a dialogue turn, we provide in-context examples corresponding to the same turn from other domains. If for a particular turn, sufficient in-context examples are not available, we look for the *latest previous turn* for which sufficient in-context examples are available. E.g. Assume the following turn to count distribution and $k = 4$ (number of in-context examples). Turns 1–4: more than 10 examples, Turn 5: 3 examples, and Turn 6 has 1 example.

At turns 5 and 6, we do not have sufficient examples from turn 5 or 6. Therefore, we sample in-context examples from turn 4 for both of them. Our prompts for each subtasks can be seen in Fig. 9, 10, 11, 12, 13.

4 Results

4.1 XNLI

All models perform best on English, with slightly lower performance on Greek and German, and lower performance on languages like Hindi, Thai, Urdu, and Swahili. Overall PaLM2 performs best, closely followed by GPT-4. GPT-3.5-Turbo is worse on all languages, however, we find that all three Llama models perform substantially worse, with Mistral performing the worst. Since XNLI is a popular dataset, dataset contamination cannot be

[1]We set the temperature parameter equal to 0 (or close to 0) for all our models to ensure deterministic output and reproducibility

ruled out. (Figure 18, Table 2).

4.2 IndicXNLI

We performed experiments on IndicXNLI on the GPT models, Mistral as well as Llama models, however, the Llama models gave scores of 0 for all languages, which is why we do not plot them. The Mistral model also performs poorly. We find that GPT-4 outperforms GPT-3.5-Turbo on all languages with the highest scores on Hindi, Punjabi, and Bengali. However, the overall accuracy is not very high on any language compared to the XNLI results seen earlier, and fine-tuned baselines such as MuRIL perform best. (Figure 19, Table 3).

4.3 GLUECoS NLI

All models do well on this NLI task, with GPT-4 performing best. (Figure 26, Table 14).

4.4 PAWS-X

PaLM2 outperforms the GPT models on all languages and all models perform well, which could be because this dataset contains high-resource languages. However, dataset contamination cannot be ruled out, as shown in Ahuja et al. (2023). The performance on English performs is the best, followed closely by Latin script languages, and a drop in performance for languages in other scripts. The Llama and Mistral models perform worse than the GPT models and PaLM2, although the difference in performance is not as large as in some of the other datasets. (Figure 20, Table 4).

4.5 XCOPA

The performance of GPT-4, Gemma, Gemini and PaLM2 are comparable, with GPT-4 having the best perforamnce. Notably, they are all better than GPT-3.5-Turbo, which performs substantially better than the Llama2 and Mistral models except in Quechua, for which no model performs well. However, the results on all other languages for GPT-4 and PaLM2 are extremely high, which may be due to dataset contamination. (Figure 21, Table 5).

4.6 XStoryCloze

Since the Llama models gave scores of 0 for all languages, we omit it from our analysis. We find that the gap between the GPT models and PaLM2 is very high, with both GPT models performing extremely well. For all languages except Telugu, Basque and Burmese Gemini-pro performs well. The contamination study from Ahuja et al. (2023) show a low chance of dataset contamination for GPT-4, which indicates that the GPT models can perform this task well. (Figure 22, Table 13).

4.7 Sentiment Analysis (En-Es-CS)

Surprisingly, GPT-3.5-Turbo outperforms both GPT-4 and PaLM2 on this task, with the mBERT baseline performing the best, while Gemini-pro performs the worst by a large margin. (Figure 26, Table 14).

4.8 TyDiQA GoldP

The TuLR model performs best, followed by GPT-4, PaLM2, Gemini-Pro, and BLOOMZ, while Llama models perform poorly, with Mistral being slightly better. Smaller models, in particular, demonstrate a significant performance gap between English and all other languages. However, dataset contamination cannot be ruled out, as shown in Ahuja et al. (2023). (Figure 23, Table 7).

4.9 MLQA

TULR and GPT-4 outperform all other models for this dataset except for German. English exhibits superior performance, with Spanish (es), German (de), and Vietnamese (vi) following closely. The most significant gaps are noted between English and Arabic (ar), Hindi (hi), and Chinese (zh) The Llama2-13B model performs well for some languages, such as Arabic, German, and Spanish but performs poorly on Chinese Hindi, and Vietnamese, but is still better than Mistral and Gemma. This is one of the datasets where PaLM2 struggles, particularly for Arabic and Chinese. Dataset contamination in GPT-4 cannot be ruled out, as shown in Ahuja et al. (2023). Smaller versions of the Llama model outperform the Llama 70B model across all languages. (Figure 24, Table 8).

4.10 XQUAD

TuLRv6 performs best across almost all languages in the XQuAD dataset, followed by GPT-4, PaLM 2, Gemini-Pro, and BLOOMZ. BLOOMZ's performance declines significantly in Greek and Thai as shown in Figure 2. PaLM2 and Gemini-Pro exhibit competitive performance, closely trailing GPT-4-32K and TuLRv6 – XXL across languages from high to mid-resource tiers. All three Llama models perform poorly on this dataset. Gemma and Mistral perform slightly better than Llama on all languages but lags behind the larger models and finetuned models. Dataset contamination in GPT-4 cannot be

ruled out, as shown in Ahuja et al. (2023). (Figure 2, Table 6).

4.11 IndicQA

Since the Llama models gave scores of 0 for all languages, we omit it from our analysis. We use the zero-shot cross-lingual prompting strategy due to the absence of a dev set. GPT-4 performs better than GPT-3.5-Turbo, with the best performance seen for Hindi, Marathi, and Bengali, while the smaller models like Gemma perform poorly. (Figure 25, Table 9).

4.12 PAN-X

GPT-4 and GPT-3.5-Turbo outperform PaLM2 and gemini-pro for most languages. However, all models perform poorly on Thai, Japanese, and Chinese on this sequence labeling task. Since this is an older dataset, GPT-4 data contamination cannot be ruled out as shown in Ahuja et al. (2023). (Figure 31, Table 12).

4.13 UDPOS

PaLM2 performs the best followed by GPT-4, GPT-3.5-Turbo and Gemini-pro being the worst on average. All models show similar high performance across languages, except for Arabic, Greek, Hebrew, Hindi, and Vietnamese, where PaLM2 performs best. GPT-4 data contamination cannot be ruled out as shown in Ahuja et al. (2023). (Figure 33, Table 11).

4.14 Jigsaw

We perform experiments on the Jigsaw dataset for GPT-3.5-Turbo and PaLM2 using the monolingual prompting strategy and find that both models perform very well on all languages. Since the dataset cannot be accessed without download, models are less likely to be contaminated with this dataset. (Figure 30, Table 19).

4.15 WinoMT

We perform experiments on the WinoMT dataset only for GPT-3.5-Turbo using the monolingual prompting strategy and report the results for completeness. We find that the model does not perform well on any of the languages. (Figure 29, Table 20).

4.16 XLSum

GPT-4 outperforms all other models, with some exceptions. GPT-3.5-Turbo performs best for African languages like Swahili, Somali, and Yoruba, while the Llama models perform best for Arabic, Kyrgyz, Vietnamese, and Welsh. According to the contamination analysis in Ahuja et al. (2023), it is possible, though less likely that GPT-4 is contaminated with this dataset. (Figure 34, Table 15).

4.17 Belebele

Gemini-Pro has the best performance amongst all the models for most languages, while for smaller models only Llama models come close. GPT-4 and PaLM2 outperform GPT-3.5-Turbo, Llama2, and Mistral, which performs worst. Most models do well due to the multiple-choice question-answering nature of the task, which makes parsing outputs and evaluation simpler and increases the probability of success even for weaker models. (Figure 16, Table 17).

4.18 AfriQA

GPT-4 has best performance, while the Llama2 and Mistral models perform very poorly on all languages. (Figure 15, Table 10).

4.19 IN22

We report our results on the IN22-Gen and IN22-Conv subsets (Figure 35) where we randomly select $k = 8$ translation pairs from the development set of FLORES-200 (Costa-jussà et al., 2022) as in-context examples. We also report GPT-3.5-Turbo 0-shot and IndicTrans2 scores from Gala et al. (2023) for comparison. For consistency, we use the `indic_nlp_library`[2] and the evaluation scripts[3] from Gala et al. (2023) to tokenize the predictions and references before computing chrF++ (Popović, 2017) for Indic languages. We do not evaluate PaLM2 on this dataset, as most languages in this dataset are not supported by it.

Llama2 and Mistral perform poorly on all Indic languages in the En-Indic direction, whereas the performance is better on the Indic-En direction. Gemma-7B performs significantly better than both Llama2 and Mistral in both directions and on all languages. GPT-4 performs the best among all LLM models considered. All LLMs perform better in the Indic-En direction and Conversational dataset since they are finetuned with chat or conversational style data. We compare results to IndicTrans2 Gala et al. (2023) and find that it fares

[2] https://github.com/anoopkunchukuttan/indic_nlp_library

[3] https://github.com/AI4Bharat/IndicTrans2

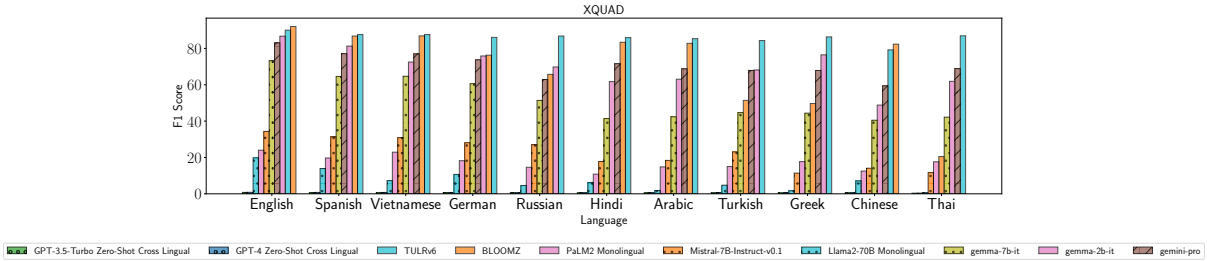

Figure 2: Results for XQUAD across all languages and models for zero-shot cross-lingual prompting

significantly better than LLMs. (Figure 35, Tables 21 - 24).

4.20 XRiSAWOZ

We compare DA accuracy of various models in Figure 17. Table 25 shows the comparison with fine-tuned models as well. We find that GPT-4's performance on DA accuracy is the closest and comparable to fine-tuned baselines for the task. Poorer scores on other models seem to correlate with the model's hallucination tendencies.

We compare results on all 6 metrics in Table 26 to better understand model behavior. We find that PaLM2,GPT-4 and Gemini-pro generate very concise responses leading to consistently higher BLEU scores as compared to other models. On all other metrics, GPT family of models significantly outperforms both PaLM/Gemini and open-source models. Notably, all the proprietary models achieve less than 10% SER on Chinese hinting contamination of RiSAWOZ (the original Chinese-only dataset). Open source models often hallucinated non-existent entities in their responses while proprietary models did not show this tendency.

In the code-mixed English-Hindi setting, the performance is worse than both English and Hindi on average across most metrics for all models. (Figure 17, Tables 25, 26). This could indicate challenges in understanding as well as generating effective code mixed text for all models.

4.21 MaRVL

We evaluate LLaVA models, GPT-4-Vision[4], and Gemini-Pro-Vision on the multimodal datasets with monolingual and translate-test prompting (Figure 27). The Azure BING translate module was utilized for translating the sentences into English. We find that accuracy scores border on random classification LLaVA models, with the lowest score on Tamil and Chinese. The translate-test strategy is comparable to monolingual. However, the performance is still the same as a random classification. GPT-4-Vision is significantly better than LLaVA, and the gains due to translate-test are only visible on Turkish. Gemini-Pro-Vision performs slightly better than random, and the translate-test is preferable except in the case of Chinese. (Figure 27, Table 16).

4.22 XM-3600

We test the LLaVA models, GPT-4-Vision[5], and Gemini-Pro-Vision models on the XM-3600 image captioning dataset and use the chrF metric (Popović, 2015) to report the performance, unlike the original paper (Thapliyal et al., 2022) that uses CIDEr. We see that the LLaVA models are poor for most languages that are not written in Latin script, especially Japanese, Korean, Russian, Thai, and Chinese. bakLLaVA-v1 performs much worse compared to LLaVA-v1.5-13B and ViP-LLaVA-13B (except English), and the latter two are comparable on all languages. Most Latin script high-resource languages such as French, German, Dutch, Spanish, and Italian outperform or come close to English performance, with lower-resource languages such as Danish, Czech Polish, and Norwegian performing worse. GPT-4-Vision significantly outperforms LLaVA models on all languages, however, the scores on Chinese, Japanese, and Thai are still very poor. French has the highest score followed by Italian, Spanish, and then English, which again shows that GPT-4-Vision is good at Latin script and European languages. Gemini-Pro-Vision is the second-best model on all languages, and the results follow the same trend as GPT-4-Vision. (Figure 28, Table 18).

[4]Given the API costs and constraints, we evaluate a random sample of 300 data instances per language.

[5]Due to API costs and constraints, we evaluate a random sample of 488 data instances per language.

4.23 The deviation of performance across language families and tasks

Given the experiments conducted, we look at how performance for a given Language Family or Task varies from the average performance (across the models covered in MEGAVERSE). In doing so we are interested in ranking how well models support different Language Families or Tasks.

The deviation for a given experiment i in the Language Family or Task (j) is defined as:

$$\Delta_{(i,j)} = p\_score_{(i,j)} - \frac{1}{N}\sum_{i}^{N} p\_score_{(i,j)}$$

Where $p\_score_{(i,j)}$ is the penalized score for the experiment i, and a high positive value indicates that a given subject (Language Family or Task) performs better than average where as a low negative value indicates that the subject performs lower than the average (across all models). $p\_score_{(i,j)}$ is calculated as:

$$p\_score_{(i,j)} = (\frac{|X_j|}{\sum_i |X_j|}) * score_i$$

Where $score_i$ is the normalized score for the experiment, penalized by the ratio of the instances in a given language family/task (j) to the total number of instances in all the language families/tasks.

Because of the sparsity in (Language, Dataset, Model) combinations (see Table 1), we apply the size penalization to limit the bias of outliers and combinations with little support. For example, there are total of 320 IE: Iranian Language family experiments in our data, with an average score of 0.31, and a penalized score of 0.05, compared to Basque which has 10 experiments with an average score of 0.54, but a penalized score of 0.003.

Figure 3 gives the distribution of the $\Delta_{(i,j)}$ scores for Language Families and Tasks. We observe that languages in IE:Germanic Family, which ranks at the top, attain a significantly higher score that the mean, while at the the opposite end, Bantu and Afro-Asiatic languages significantly underperform the mean across models and datasets. We also find that the models tested are significantly better at tasks such as MCQ Reading Comprehension and Parts of Speech Tagging (across all languages), than more open tasks such as Q&A and text Summarization.

5 Contamination Analysis

5.1 Commercial Model Contamination Study

In our work, we follow the method described by Golchin and Surdeanu (2023a) where we try to quantify contamination for commercial models such as PaLM2 and GPT-4. First, we prompt the model to generate three perturbations of the test set data points. Next, we provide these perturbations appended with the original text as four options to the model, and prompt it to pick a preferred option. We measure contamination as the chance adjusted accuracy using Cohen's Kappa (κ) and account for LLM's position bias towards a particular option by adjusting the calculation of κ, called κ_{fixed}.

We study contamination on GPT-4 and PaLM2 for 5 datasets: PAWS-X, UDPOS, TyDiQA, XNLI, and XCOPA, on 100 data points per language in each dataset. Our results show that all datasets are highly contaminated except for UDPOS, and for all datasets, contamination is higher for GPT-4, than for PaLM2. Contamination values for all datasets across different languages are reported in Appendix A.6. Contamination values differ significantly across languages for the same dataset, which could be due to bad perturbations generated by models owing to their varying performance in different languages. Another limitation of this approach is that Golchin and Surdeanu (2023a) study position bias only for GPT models and append the original text as the fourth option based on their observations. However, this could vary for different models.

5.2 Open-Source Model Contamination study

We follow the Black Box test for contamination study of open-source model described by Oren et al. (2023). This test is statistical test which provides provable guarantees that a given test set is contaminated. To achieve these guarantees, they exploit the fact that many datasets have a property known as *exchangeability*, where the order of examples in the dataset can be shuffled without affecting its join distribution. If a model has seen a benchmark dataset, it will have a preference for the canonical order (i.e. the order that examples are given in the public repositories) over randomly shuffled example orderings. If the difference between the said canonical order and the shuffled order is statistically significant, then the dataset is considered to be contaminated according to this method.

We conducted tests on the 7B instruction-tuned

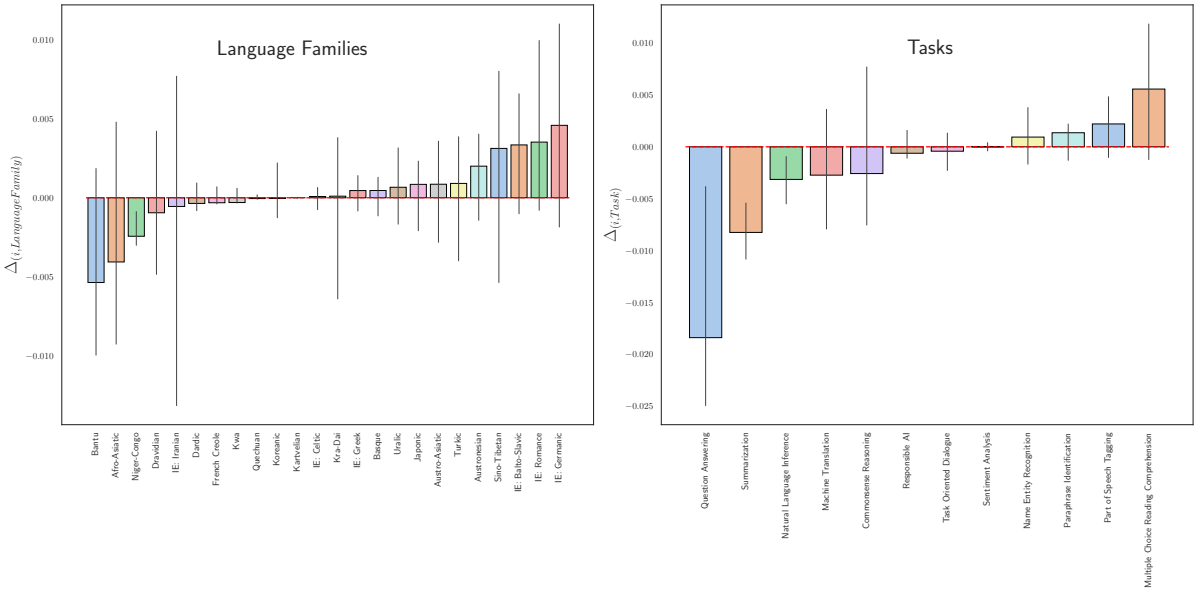

Figure 3: The positive scores of the bar-plots denote that the current LLMs are relatively good with those language families / tasks.

variants of Llama2, Mistral, and Gemma across the following evaluation datasets: PAWS-X, XCOPA, XNLI, XQUAD, XRiSAWOZ, and XstoryCloze. The significance level for our analysis was set at 0.001. We observed (Table 33) that all the models that we study, exhibited contamination. Specifically, datasets such as PAWS-X, XCOPA, XQUAD, and XRiSAWOZ were found to have their p-values less than the significant value for Gemma 7B Instruct, Llama2 7B Instruct and Mistral 7B Instruct indicating contamination.

6 Discussion

In this work, we benchmark 22 datasets covering 83 languages across several models – GPT-3.5-Turbo, GPT-4, PaLM2, Gemini-Pro, Gemma, Llama2, Mistral as well as multimodal models. We find similar trends across most datasets we study - larger commercial models such as GPT-4 and Gemini-pro outperform smaller models like Gemma, Llama and Mistral models, particularly on low-resource languages. This suggests that multilingual performance is a challenge for smaller models, and directions such as language-specific models, language family-based models and fine-tuning should be explored for better multilingual performance.

GPT-4, PaLM2 and Gemini-Pro excel on different datasets, with GPT-4 showing superior performance overall on multilingual datasets compared to both PaLM2 and Gemini-Pro. GPT-4-Vision outperforms LLaVA and Gemini-Pro-Vision on the multimodal datasets we study. Tokenizer fertility is correlated with Language Model performance (Rust et al., 2021; Ali et al., 2023). We plot the fertility analysis of all the tokenizers (Figure: 14) for the models that we studied in this work. We noticed that on average, Latin script languages such as Spanish, English had lower fertility as compared to languages that are morphologically complex languages like Telugu, Malay and Malayalam having high fertility amongst all the tokenizers.

Dataset contamination is a critical issue that affects English and non-English language benchmarking studies. Our contamination analysis on open source and commercial models shows that almost all models are contaminated with datasets included in MEGAVERSE. New multilingual evaluation datasets are difficult to create due to resource and funding constraints, hence, care should be taken to make sure that they are not included in the training data of LLMs. To achieve this, we need to enhance our ability to identify instances of contamination, as well as implement measures to avoid future contamination.

7 Limitations

Our work is subject to the following limitations:

Model comparison We have covered a wide array of Large Language Models. We realize that access to the commercial models (GPT, PaLM2, etc.) is via an API endpoint. These models might be

running various post-processing modules and classifiers resulting in an inflated performance as compared to the Open-source models (LLaVA, Llama, Mistral).

Dataset contamination We perform the dataset contamination exercise on a few set of datasets for PaLM2 and GPT-4 on a granular level. We also perform a thorough analysis of the open-source models covered in MEGAVERSE. However, there were certain limitations that we discuss in depth in Section 5. We were also limited by the compute and time, therefore we did not perform the contamination study on all our datasets and only covered the 7B variants of our open-source models.

Prompt tuning LLMs are sensitive to prompting, and we do not perform extensive prompt tuning for the new datasets. We also do not experiment with prompting variations, such as translate-test and zero-shot cross-lingual prompting, or more complex strategies such as Chain of Thought prompting due to resource constraints.

Experiments on limited data and datasets Due to resource constraints, we perform experiments on partial datasets when indicated, and do not evaluate all models on all datasets. We plan to do so in future work.

Focus on task accuracy We perform limited experiments on RAI datasets and do not perform experiments on other important dimensions such as fairness, bias, robustness, efficiency, etc., mainly due to the lack of such datasets for non-English languages. This is an important future research direction.

References

Judit Ács. 2019. Exploring BERT's Vocabulary. *Blog Post*.

Divyanshu Aggarwal, Vivek Gupta, and Anoop Kunchukuttan. 2022. IndicXNLI: Evaluating multilingual inference for Indian languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10994–11006, 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.

Mehdi Ali, Michael Fromm, Klaudia Thellmann, Richard Rutmann, Max Lübbering, Johannes Leveling, Katrin Klug, Jan Ebert, Niclas Doll, Jasper Schulze Buschhoff, Charvi Jain, Alexander Arno Weber, Lena Jurkschat, Hammam Abdelwahab, Chelsea John, Pedro Ortiz Suarez, Malte Ostendorff, Samuel Weinbach, Rafet Sifa, Stefan Kesselheim, and Nicolas Flores-Herr. 2023. Tokenizer choice for llm training: Negligible or crucial?

Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R. Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, Alexandre Frechette, Charlotte Smith, Laura Culp, Lev Proleev, Yi Luan, Xi Chen, James Lottes, Nathan Schucher, Federico Lebron, Alban Rrustemi, Natalie Clay, Phil Crone, Tomas Kocisky, Jeffrey Zhao, Bartek Perz, Dian Yu, Heidi Howard, Adam Bloniarz, Jack W. Rae, Han Lu, Laurent Sifre, Marcello Maggioni, Fred Alcober, Dan Garrette, Megan Barnes, Shantanu Thakoor, Jacob Austin, Gabriel Barth-Maron, William Wong, Rishabh Joshi, Rahma Chaabouni, Deeni Fatiha, Arun Ahuja, Ruibo Liu, Yunxuan Li, Sarah Cogan, Jeremy Chen, Chao Jia, Chenjie Gu, Qiao Zhang, Jordan Grimstad, Ale Jakse Hartman, Martin Chadwick, Gaurav Singh Tomar, Xavier Garcia, Evan Senter, Emanuel Taropa, Thanumalayan Sankaranarayana Pillai, Jacob Devlin, Michael Laskin, Diego de Las Casas, Dasha Valter, Connie Tao, Lorenzo Blanco, Adrià Puigdomènech Badia, David Reitter, Mianna Chen, Jenny Brennan, Clara Rivera, Sergey Brin, Shariq Iqbal, Gabriela Surita, Jane Labanowski, Abhi Rao, Stephanie Winkler, Emilio Parisotto, Yiming Gu, Kate Olszewska, Yujing Zhang, Ravi Addanki, Antoine Miech, Annie Louis, Laurent El Shafey, Denis Teplyashin, Geoff Brown, Elliot Catt, Nithya Attaluri, Jan Balaguer, Jackie Xiang, Pidong Wang, Zoe Ashwood, Anton Briukhov, Albert Webson, Sanjay Ganapathy, Smit Sanghavi, Ajay Kannan, Ming-Wei Chang, Axel Stjerngren, Josip Djolonga, Yuting Sun, Ankur Bapna, Matthew Aitchison, Pedram Pejman, Henryk Michalewski, Tianhe Yu, Cindy Wang, Juliette Love, Junwhan Ahn, Dawn Bloxwich, Kehang Han, Peter Humphreys, Thibault Sellam, James Bradbury, Varun Godbole,

Sina Samangooei, Bogdan Damoc, Alex Kaskasoli, Sébastien M. R. Arnold, Vijay Vasudevan, Shubham Agrawal, Jason Riesa, Dmitry Lepikhin, Richard Tanburn, Srivatsan Srinivasan, Hyeontaek Lim, Sarah Hodkinson, Pranav Shyam, Johan Ferret, Steven Hand, Ankush Garg, Tom Le Paine, Jian Li, Yujia Li, Minh Giang, Alexander Neitz, Zaheer Abbas, Sarah York, Machel Reid, Elizabeth Cole, Aakanksha Chowdhery, Dipanjan Das, Dominika Rogozińska, Vitaly Nikolaev, Pablo Sprechmann, Zachary Nado, Lukas Zilka, Flavien Prost, Luheng He, Marianne Monteiro, Gaurav Mishra, Chris Welty, Josh Newlan, Dawei Jia, Miltiadis Allamanis, Clara Huiyi Hu, Raoul de Liedekerke, Justin Gilmer, Carl Saroufim, Shruti Rijhwani, Shaobo Hou, Disha Shrivastava, Anirudh Baddepudi, Alex Goldin, Adnan Ozturel, Albin Cassirer, Yunhan Xu, Daniel Sohn, Devendra Sachan, Reinald Kim Amplayo, Craig Swanson, Dessie Petrova, Shashi Narayan, Arthur Guez, Siddhartha Brahma, Jessica Landon, Miteyan Patel, Ruizhe Zhao, Kevin Villela, Luyu Wang, Wenhao Jia, Matthew Rahtz, Mai Giménez, Legg Yeung, Hanzhao Lin, James Keeling, Petko Georgiev, Diana Mincu, Boxi Wu, Salem Haykal, Rachel Saputro, Kiran Vodrahalli, James Qin, Zeynep Cankara, Abhanshu Sharma, Nick Fernando, Will Hawkins, Behnam Neyshabur, Solomon Kim, Adrian Hutter, Priyanka Agrawal, Alex Castro-Ros, George van den Driessche, Tao Wang, Fan Yang, Shuo yiin Chang, Paul Komarek, Ross McIlroy, Mario Lučić, Guodong Zhang, Wael Farhan, Michael Sharman, Paul Natsev, Paul Michel, Yong Cheng, Yamini Bansal, Siyuan Qiao, Kris Cao, Siamak Shakeri, Christina Butterfield, Justin Chung, Paul Kishan Rubenstein, Shivani Agrawal, Arthur Mensch, Kedar Soparkar, Karel Lenc, Timothy Chung, Aedan Pope, Loren Maggiore, Jackie Kay, Priya Jhakra, Shibo Wang, Joshua Maynez, Mary Phuong, Taylor Tobin, Andrea Tacchetti, Maja Trebacz, Kevin Robinson, Yash Katariya, Sebastian Riedel, Paige Bailey, Kefan Xiao, Nimesh Ghelani, Lora Aroyo, Ambrose Slone, Neil Houlsby, Xuehan Xiong, Zhen Yang, Elena Gribovskaya, Jonas Adler, Mateo Wirth, Lisa Lee, Music Li, Thais Kagohara, Jay Pavagadhi, Sophie Bridgers, Anna Bortsova, Sanjay Ghemawat, Zafarali Ahmed, Tianqi Liu, Richard Powell, Vijay Bolina, Mariko Iinuma, Polina Zablotskaia, James Besley, Da-Woon Chung, Timothy Dozat, Ramona Comanescu, Xiance Si, Jeremy Greer, Guolong Su, Martin Polacek, Raphaël Lopez Kaufman, Simon Tokumine, Hexiang Hu, Elena Buchatskaya, Yingjie Miao, Mohamed Elhawaty, Aditya Siddhant, Nenad Tomasev, Jinwei Xing, Christina Greer, Helen Miller, Shereen Ashraf, Aurko Roy, Zizhao Zhang, Ada Ma, Angelos Filos, Milos Besta, Rory Blevins, Ted Klimenko, Chih-Kuan Yeh, Soravit Changpinyo, Jiaqi Mu, Oscar Chang, Mantas Pajarskas, Carrie Muir, Vered Cohen, Charline Le Lan, Krishna Haridasan, Amit Marathe, Steven Hansen, Sholto Douglas, Rajkumar Samuel, Mingqiu Wang, Sophia Austin, Chang Lan, Jiepu Jiang, Justin Chiu, Jaime Alonso Lorenzo, Lars Lowe Sjösund, Sébastien Cevey, Zach Gleicher, Thi Avrahami, Anudhyan Boral, Hansa Srinivasan, Vittorio Selo, Rhys May, Konstantinos Aisopos, Léonard Hussenot, Livio Baldini Soares, Kate Baumli, Michael B. Chang, Adrià Recasens, Ben Caine, Alexander Pritzel, Filip Pavetic, Fabio Pardo, Anita Gergely, Justin Frye, Vinay Ramasesh, Dan Horgan, Kartikeya Badola, Nora Kassner, Subhrajit Roy, Ethan Dyer, Víctor Campos, Alex Tomala, Yunhao Tang, Dalia El Badawy, Elspeth White, Basil Mustafa, Oran Lang, Abhishek Jindal, Sharad Vikram, Zhitao Gong, Sergi Caelles, Ross Hemsley, Gregory Thornton, Fangxiaoyu Feng, Wojciech Stokowiec, Ce Zheng, Phoebe Thacker, Çağlar Ünlü, Zhishuai Zhang, Mohammad Saleh, James Svensson, Max Bileschi, Piyush Patil, Ankesh Anand, Roman Ring, Katerina Tsihlas, Arpi Vezer, Marco Selvi, Toby Shevlane, Mikel Rodriguez, Tom Kwiatkowski, Samira Daruki, Keran Rong, Allan Dafoe, Nicholas FitzGerald, Keren Gu-Lemberg, Mina Khan, Lisa Anne Hendricks, Marie Pellat, Vladimir Feinberg, James Cobon-Kerr, Tara Sainath, Maribeth Rauh, Sayed Hadi Hashemi, Richard Ives, Yana Hasson, YaGuang Li, Eric Noland, Yuan Cao, Nathan Byrd, Le Hou, Qingze Wang, Thibault Sottiaux, Michela Paganini, Jean-Baptiste Lespiau, Alexandre Moufarek, Samer Hassan, Kaushik Shivakumar, Joost van Amersfoort, Amol Mandhane, Pratik Joshi, Anirudh Goyal, Matthew Tung, Andrew Brock, Hannah Sheahan, Vedant Misra, Cheng Li, Nemanja Rakićević, Mostafa Dehghani, Fangyu Liu, Sid Mittal, Junhyuk Oh, Seb Noury, Eren Sezener, Fantine Huot, Matthew Lamm, Nicola De Cao, Charlie Chen, Gamaleldin Elsayed, Ed Chi, Mahdis Mahdieh, Ian Tenney, Nan Hua, Ivan Petrychenko, Patrick Kane, Dylan Scandinaro, Rishub Jain, Jonathan Uesato, Romina Datta, Adam Sadovsky, Oskar Bunyan, Dominik Rabiej, Shimu Wu, John Zhang, Gautam Vasudevan, Edouard Leurent, Mahmoud Alnahlawi, Ionut Georgescu, Nan Wei, Ivy Zheng, Betty Chan, Pam G Rabinovitch, Piotr Stanczyk, Ye Zhang, David Steiner, Subhajit Naskar, Michael Azzam, Matthew Johnson, Adam Paszke, Chung-Cheng Chiu, Jaume Sanchez Elias, Afroz Mohiuddin, Faizan Muhammad, Jin Miao, Andrew Lee, Nino Vieillard, Sahitya Potluri, Jane Park, Elnaz Davoodi, Jiageng Zhang, Jeff Stanway, Drew Garmon, Abhijit Karmarkar, Zhe Dong, Jong Lee, Aviral Kumar, Luowei Zhou, Jonathan Evens, William Isaac, Zhe Chen, Johnson Jia, Anselm Levskaya, Zhenkai Zhu, Chris Gorgolewski, Peter Grabowski, Yu Mao, Alberto Magni, Kaisheng Yao, Javier Snaider, Norman Casagrande, Paul Suganthan, Evan Palmer, Geoffrey Irving, Edward Loper, Manaal Faruqui, Isha Arkatkar, Nanxin Chen, Izhak Shafran, Michael Fink, Alfonso Castaño, Irene Giannoumis, Wooyeol Kim, Mikołaj Rybiński, Ashwin Sreevatsa, Jennifer Prendki, David Soergel, Adrian Goedeckemeyer, Willi Gierke, Mohsen Jafari, Meenu Gaba, Jeremy Wiesner, Diana Gage Wright, Yawen Wei, Harsha Vashisht, Yana Kulizhskaya, Jay Hoover, Maigo Le, Lu Li, Chimezie Iwuanyanwu, Lu Liu, Kevin Ramirez, Andrey Khorlin, Albert Cui, Tian LIN, Marin Georgiev, Marcus Wu, Ricardo Aguilar, Keith Pallo, Abhishek Chakladar, Alena Repina, Xihui Wu, Tom van der Weide, Priya Ponnapalli, Caroline Kaplan, Jiri Simsa, Shuangfeng Li, Olivier

Dousse, Fan Yang, Jeff Piper, Nathan Ie, Minnie Lui, Rama Pasumarthi, Nathan Lintz, Anitha Vijayakumar, Lam Nguyen Thiet, Daniel Andor, Pedro Valenzuela, Cosmin Paduraru, Daiyi Peng, Katherine Lee, Shuyuan Zhang, Somer Greene, Duc Dung Nguyen, Paula Kurylowicz, Sarmishta Velury, Sebastian Krause, Cassidy Hardin, Lucas Dixon, Lili Janzer, Kiam Choo, Ziqiang Feng, Biao Zhang, Achintya Singhal, Tejasi Latkar, Mingyang Zhang, Quoc Le, Elena Allica Abellan, Dayou Du, Dan McKinnon, Natasha Antropova, Tolga Bolukbasi, Orgad Keller, David Reid, Daniel Finchelstein, Maria Abi Raad, Remi Crocker, Peter Hawkins, Robert Dadashi, Colin Gaffney, Sid Lall, Ken Franko, Egor Filonov, Anna Bulanova, Rémi Leblond, Vikas Yadav, Shirley Chung, Harry Askham, Luis C. Cobo, Kelvin Xu, Felix Fischer, Jun Xu, Christina Sorokin, Chris Alberti, Chu-Cheng Lin, Colin Evans, Hao Zhou, Alek Dimitriev, Hannah Forbes, Dylan Banarse, Zora Tung, Jeremiah Liu, Mark Omernick, Colton Bishop, Chintu Kumar, Rachel Sterneck, Ryan Foley, Rohan Jain, Swaroop Mishra, Jiawei Xia, Taylor Bos, Geoffrey Cideron, Ehsan Amid, Francesco Piccinno, Xingyu Wang, Praseem Banzal, Petru Gurita, Hila Noga, Premal Shah, Daniel J. Mankowitz, Alex Polozov, Nate Kushman, Victoria Krakovna, Sasha Brown, MohammadHossein Bateni, Dennis Duan, Vlad Firoiu, Meghana Thotakuri, Tom Natan, Anhad Mohananey, Matthieu Geist, Sidharth Mudgal, Sertan Girgin, Hui Li, Jiayu Ye, Ofir Roval, Reiko Tojo, Michael Kwong, James Lee-Thorp, Christopher Yew, Quan Yuan, Sumit Bagri, Danila Sinopalnikov, Sabela Ramos, John Mellor, Abhishek Sharma, Aliaksei Severyn, Jonathan Lai, Kathy Wu, Heng-Tze Cheng, David Miller, Nicolas Sonnerat, Denis Vnukov, Rory Greig, Jennifer Beattie, Emily Caveness, Libin Bai, Julian Eisenschlos, Alex Korchemniy, Tomy Tsai, Mimi Jasarevic, Weize Kong, Phuong Dao, Zeyu Zheng, Frederick Liu, Fan Yang, Rui Zhu, Mark Geller, Tian Huey Teh, Jason Sanmiya, Evgeny Gladchenko, Nejc Trdin, Andrei Sozanschi, Daniel Toyama, Evan Rosen, Sasan Tavakkol, Linting Xue, Chen Elkind, Oliver Woodman, John Carpenter, George Papamakarios, Rupert Kemp, Sushant Kafle, Tanya Grunina, Rishika Sinha, Alice Talbert, Abhimanyu Goyal, Diane Wu, Denese Owusu-Afriyie, Cosmo Du, Chloe Thornton, Jordi Pont-Tuset, Pradyumna Narayana, Jing Li, Sabaer Fatehi, John Wieting, Omar Ajmeri, Benigno Uria, Tao Zhu, Yeongil Ko, Laura Knight, Amélie Héliou, Ning Niu, Shane Gu, Chenxi Pang, Dustin Tran, Yeqing Li, Nir Levine, Ariel Stolovich, Norbert Kalb, Rebeca Santamaria-Fernandez, Sonam Goenka, Wenny Yustalim, Robin Strudel, Ali Elqursh, Balaji Lakshminarayanan, Charlie Deck, Shyam Upadhyay, Hyo Lee, Mike Dusenberry, Zonglin Li, Xuezhi Wang, Kyle Levin, Raphael Hoffmann, Dan Holtmann-Rice, Olivier Bachem, Summer Yue, Sho Arora, Eric Malmi, Daniil Mirylenka, Qijun Tan, Christy Koh, Soheil Hassas Yeganeh, Siim Põder, Steven Zheng, Francesco Pongetti, Mukarram Tariq, Yanhua Sun, Lucian Ionita, Mojtaba Seyedhosseini, Pouya Tafti, Ragha Kotikalapudi, Zhiyu Liu, Anmol Gulati, Jasmine Liu, Xinyu Ye, Bart Chrzaszcz, Lily Wang, Nikhil Sethi, Tianrun Li, Ben Brown, Shreya Singh, Wei Fan, Aaron Parisi, Joe Stanton, Chenkai Kuang, Vinod Koverkathu, Christopher A. Choquette-Choo, Yunjie Li, TJ Lu, Abe Ittycheriah, Prakash Shroff, Pei Sun, Mani Varadarajan, Sanaz Bahargam, Rob Willoughby, David Gaddy, Ishita Dasgupta, Guillaume Desjardins, Marco Cornero, Brona Robenek, Bhavishya Mittal, Ben Albrecht, Ashish Shenoy, Fedor Moiseev, Henrik Jacobsson, Alireza Ghaffarkhah, Morgane Rivière, Alanna Walton, Clément Crepy, Alicia Parrish, Yuan Liu, Zongwei Zhou, Clement Farabet, Carey Radebaugh, Praveen Srinivasan, Claudia van der Salm, Andreas Fidjeland, Salvatore Scellato, Eri Latorre-Chimoto, Hanna Klimczak-Plucińska, David Bridson, Dario de Cesare, Tom Hudson, Piermaria Mendolicchio, Lexi Walker, Alex Morris, Ivo Penchev, Matthew Mauger, Alexey Guseynov, Alison Reid, Seth Odoom, Lucia Loher, Victor Cotruta, Madhavi Yenugula, Dominik Grewe, Anastasia Petrushkina, Tom Duerig, Antonio Sanchez, Steve Yadlowsky, Amy Shen, Amir Globerson, Adam Kurzrok, Lynette Webb, Sahil Dua, Dong Li, Preethi Lahoti, Surya Bhupatiraju, Dan Hurt, Haroon Qureshi, Ananth Agarwal, Tomer Shani, Matan Eyal, Anuj Khare, Shreyas Rammohan Belle, Lei Wang, Chetan Tekur, Mihir Sanjay Kale, Jinliang Wei, Ruoxin Sang, Brennan Saeta, Tyler Liechty, Yi Sun, Yao Zhao, Stephan Lee, Pandu Nayak, Doug Fritz, Manish Reddy Vuyyuru, John Aslanides, Nidhi Vyas, Martin Wicke, Xiao Ma, Taylan Bilal, Evgenii Eltyshev, Daniel Balle, Nina Martin, Hardie Cate, James Manyika, Keyvan Amiri, Yelin Kim, Xi Xiong, Kai Kang, Florian Luisier, Nilesh Tripuraneni, David Madras, Mandy Guo, Austin Waters, Oliver Wang, Joshua Ainslie, Jason Baldridge, Han Zhang, Garima Pruthi, Jakob Bauer, Feng Yang, Riham Mansour, Jason Gelman, Yang Xu, George Polovets, Ji Liu, Honglong Cai, Warren Chen, XiangHai Sheng, Emily Xue, Sherjil Ozair, Adams Yu, Christof Angermueller, Xiaowei Li, Weiren Wang, Julia Wiesinger, Emmanouil Koukoumidis, Yuan Tian, Anand Iyer, Madhu Gurumurthy, Mark Goldenson, Parashar Shah, MK Blake, Hongkun Yu, Anthony Urbanowicz, Jennimaria Palomaki, Chrisantha Fernando, Kevin Brooks, Ken Durden, Harsh Mehta, Nikola Momchev, Elahe Rahimtoroghi, Maria Georgaki, Amit Raul, Sebastian Ruder, Morgan Redshaw, Jinhyuk Lee, Komal Jalan, Dinghua Li, Ginger Perng, Blake Hechtman, Parker Schuh, Milad Nasr, Mia Chen, Kieran Milan, Vladimir Mikulik, Trevor Strohman, Juliana Franco, Tim Green, Demis Hassabis, Koray Kavukcuoglu, Jeffrey Dean, and Oriol Vinyals. 2023a. Gemini: A family of highly capable multimodal models.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023b. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

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.

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. *arXiv cs.CL 2305.14857*.

Simone Balloccu, Patrícia Schmidtová, Mateusz Lango, and Ondřej Dušek. 2024. Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source llms. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.

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. *arXiv preprint arXiv:2308.16884*.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium. Association for Computational Linguistics.

Mu Cai, Haotian Liu, Siva Karthik Mustikovela, Gregory P. Meyer, Yuning Chai, Dennis Park, and Yong Jae Lee. 2023. Making large multimodal models understand arbitrary visual prompts. *arXiv preprint arXiv: 2312.00784*.

De Choudhury et al. 2023. Ask me in english instead: Cross-lingual evaluation of large language models for healthcare queries. *arXiv preprint arXiv:2310.13132*.

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

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of EMNLP 2018*, pages 2475–2485.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

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.

Jay Gala, Pranjal A Chitale, A K Raghavan, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar M, Janki Atul Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M Khapra, Raj Dabre, and Anoop Kunchukuttan. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *Transactions on Machine Learning Research*.

Shahriar Golchin and Mihai Surdeanu. 2023a. Data contamination quiz: A tool to detect and estimate contamination in large language models.

Shahriar Golchin and Mihai Surdeanu. 2023b. Time travel in llms: Tracing data contamination in large language models.

Google. 2023. Palm-2 technical report.

Tahmid Hasan, Abhik Bhattacharjee, Md Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M Sohel Rahman, and Rifat Shahriyar. 2021. Xl-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

Jinyi Hu, Yuan Yao, Chongyi Wang, Shan Wang, Yinxu Pan, Qianyu Chen, Tianyu Yu, Hanghao Wu, Yue Zhao, Haoye Zhang, Xu Han, Yankai Lin, Jiao Xue, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2024. Large multilingual models pivot zero-shot multimodal learning across languages.

Kent F Hubert, Kim N Awa, and Darya L Zabelina. 2024. The current state of artificial intelligence generative language models is more creative than humans on divergent thinking tasks.

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

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2021. The state and fate of linguistic diversity and inclusion in the nlp world.

Simran Khanuja, Sandipan Dandapat, Sunayana Sitaram, and Monojit Choudhury. 2020a. A new dataset for natural language inference from code-mixed conversations. In *Proceedings of the The 4th Workshop on Computational Approaches to Code Switching*, pages 9–16.

Simran Khanuja, Sandipan Dandapat, Anirudh Srinivasan, Sunayana Sitaram, and Monojit Choudhury. 2020b. Gluecos: An evaluation benchmark for code-switched nlp. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3575–3585.

Ian Kivlichan, Jeffrey Sorensen, Julia Elliott, Lucy Vasserman, Martin Görner, and Phil Culliton. 2020. Jigsaw multilingual toxic comment classification.

Viet Dac Lai, Nghia Trung Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023. Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning.

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.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zhaojiang Lin, Andrea Madotto, Genta Winata, Peng Xu, Feijun Jiang, Yuxiang Hu, Chen Shi, and Pascale N Fung. 2021. Bitod: A bilingual multi-domain dataset for task-oriented dialogue modeling. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1. Curran.

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021. Visually grounded reasoning across languages and cultures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10467–10485.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023. Improved baselines with visual instruction tuning.

Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Pier Giuseppe Sessa, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. 2024. Gemma: Open models based on gemini research and technology.

Mehrad Moradshahi, Tianhao Shen, Kalika Bali, Monojit Choudhury, Gael de Chalendar, Anmol Goel, Sungkyun Kim, Prashant Kodali, Ponnurangam Kumaraguru, Nasredine Semmar, Sina Semnani, Jiwon Seo, Vivek Seshadri, Manish Shrivastava, Michael Sun, Aditya Yadavalli, Chaobin You, Deyi Xiong, and Monica Lam. 2023. X-RiSAWOZ: High-quality end-to-end multilingual dialogue datasets and few-shot agents. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2773–2794, Toronto, Canada. Association for Computational Linguistics.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

Joakim Nivre, Mitchell Abrams, Željko Agić, Lars Ahrenberg, Lene Antonsen, Maria Jesus Aranzabe, Gashaw Arutie, Masayuki Asahara, Luma Ateyah, Mohammed Attia, et al. 2018. Universal dependencies 2.2.

Odunayo Ogundepo, Tajuddeen R Gwadabe, Clara E Rivera, Jonathan H Clark, Sebastian Ruder, David Ifeoluwa Adelani, Bonaventure FP Dossou, Abdou Aziz DIOP, Claytone Sikasote, Gilles Hacheme, et al. 2023. Afriqa: Cross-lingual open-retrieval question answering for african languages. *arXiv preprint arXiv:2305.06897*.

OpenAI. 2023a. Gpt4 technical report.

OpenAI. 2023b. Gptv system card. `https://cdn.openai.com/papers/GPTV_System_Card.pdf`. Accessed: 2023-12-13.

Yonatan Oren, Nicole Meister, Niladri Chatterji, Faisal Ladhak, and Tatsunori B Hashimoto. 2023. Proving test set contamination in black box language models. *arXiv preprint arXiv:2310.17623*.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958.

Barun Patra, Saksham Singhal, Shaohan Huang, Zewen Chi, Li Dong, Furu Wei, Vishrav Chaudhary, and Xia Song. 2023. Beyond English-centric bitexts for better multilingual language representation learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15354–15373, Toronto, Canada. Association for Computational Linguistics.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. Xcopa: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376.

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

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Jun Quan, Shian Zhang, Qian Cao, Zizhong Li, and Deyi Xiong. 2020. RiSAWOZ: A large-scale multi-domain Wizard-of-Oz dataset with rich semantic annotations for task-oriented dialogue modeling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 930–940, Online. Association for Computational Linguistics.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How good is your tokenizer? on the monolingual performance of multilingual language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3118–3135, Online. Association for Computational Linguistics.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts,

Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Bryan Orinion, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, César Ferri Ramírez, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Chris Waites, Christian Voigt, Christopher D. Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Dylan Schrader, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodola, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A. Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Wang, Gonzalo Jaimovitch-López, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Shevlin, Hinrich Schütze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B. Simon, James Koppel, James Zheng, James Zou, Jan Kocoń, Jana Thompson, Janelle Wingfield, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Batchelder, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Guerr, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh D. Dhole, Kevin Gimpel, Kevin Omondi, Kory Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros Colón, Luke Metz, Lütfi Kerem Şenel, Maarten Bosma, Maarten Sap, Maartje ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramírez Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L. Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael A. Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Swędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mitch Walker, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan A. Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nicole Martinez, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramon Risco, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan LeBras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Ruslan Salakhutdinov, Ryan Chi, Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel S. Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima, Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven T. Piantadosi, Stuart M. Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsu Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Ramasesh, Vinay Uday Prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu.

2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models.

Gabriel Stanovsky, Noah A Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684.

Ashish V Thapliyal, Jordi Pont Tuset, Xi Chen, and Radu Soricut. 2022. Crossmodal-3600: A massively multilingual multimodal evaluation dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 715–729.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

David Vilares, Miguel A Alonso, and Carlos Gómez-Rodríguez. 2016. En-es-cs: An english-spanish code-switching twitter corpus for multilingual sentiment analysis. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4149–4153.

Tsung-Hsien Wen, Milica Gašić, Nikola Mrkšić, Pei-Hao Su, David Vandyke, and Steve Young. 2015. Semantically conditioned LSTM-based natural language generation for spoken dialogue systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1711–1721, Lisbon, Portugal. Association for Computational Linguistics.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. In *Proceedings of EMNLP 2019*, pages 3685–3690.

Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2023. The dawn of lmms: Preliminary explorations with gpt-4v(ision).

Kun Zhou, Yutao Zhu, Zhipeng Chen, Wentong Chen, Wayne Xin Zhao, Xu Chen, Yankai Lin, Ji-Rong Wen, and Jiawei Han. 2023. Don't make your llm an evaluation benchmark cheater. *arXiv preprint arXiv:2311.01964*.

A Appendix

A.1 Tasks and Datasets

We benchmark 22 datasets encompassing 83 languages. A breakdown of this is described here in Table 1

Dataset	Task	Languages
XNLI	Natural Language Inference	15
Indic-XNLI	Natural Language Inference	11
GLUECoS	Natural Language Inference	2
PAWS-X	Paraphrase Identification	7
XCOPA	Commonsense Reasoning	10
XStoryCloze	Commonsense Reasoning	11
TyDiQA-GoldP	Question Answering	9
MLQA	Question Answering	6
XQuAD	Question Answering	11
IndicQA	Question Answering	10
AfriQA	Question Answering	10
MaRVL	Visual Question Answering	5
UDPOS	Part of Speech Tagging	38
PANX	Name Entity Recognition	48
XRiSAWOZ	Task Oriented Dialogue	6
WinoMT	Responsible AI	8
GLUECoS	Sentiment Analysis	2
Jigsaw	Toxicity Classification	6
XLSum	Summarization	44
IN22	Machine Translation	14
XM-3600	Image Captioning	20
BeleBele	Multiple Choice Reading Comprehension	23

Table 1: Dataset and language coverage

A.2 Prompts

Figures 4 to 13 shows the various prompts used in our benchmarking study.

A.3 Results for Fertility Analysis

Figure 14 shows fertility analysis.

A.4 Results - Figures

Figures 15 to 34 show our results on various models, languages, and datasets.

A.5 Results - Tables

Tables 2 to 26 show our results on various models, languages, and datasets.

A.6 Contamination

Tables 27 to 32 show the contamination values for the various datasets for the commercial models. For the p-values of the statistic test performed on the open-source models, please refer to Table 33.

Task Instruction $\mathcal{I}$: You are an NLP assistant trained to answer questions directly. For each question provided, respond with the most accurate and concise answer. The answer should be in the same language as the question.

Template f_{temp}:
Q: {question}
A: {answer}

Figure 4: AfriQA Prompt

Prompt: For Belebele fig: 5, we evaluated our models on zero-shot prompting using instructions proposed by Bandarkar et al. (2023) [6].
For chat-based (e.g. Llama2 chat) models and the X-RiSAWOZ prompt (fig: 9), we drop the "Learning example. . . " and "Target example. . . " and use the ChatGPT-like prompt format with task prompt in the "system" prompt, {Turn ID, Database, Context} in the "user" prompt and "Answer" in the "assistant" prompt. We use the dataset provided by Moradshahi et al. (2023) in which the context is preprocessed to include all the relevant information (e.g. previous dialogue acts or states) for a task.

```
Task Instruction 𝓘:You are an AI assistant whose purpose is to
perform reading comprehension task. Given the following passage,
query, and answer choices, output the letter corresponding to the
correct answer.

Template f_temp:
{instruction}
###
Passage:
{passage}
###
Query:
{query}
###
Choices:
(A) {A}
(B) {B}
(C) {C}
(D) {D}
###
Answer:
```

Figure 5: Belebele MRC Prompt

```
You are an AI assistant whose purpose is to perform translation.
Given the following sentence in {source}, translate it to {target}.
```

Figure 6: Translation Prompt

```
Is the below statement in {language} correct with respect to the
left and right images? Return 'TRUE' if it is true, else 'FALSE'.
CAPTION: {caption}
```

Figure 7: MaRVL Prompt

```
Generate a **brief** coco style caption for the given image in
{language}.
```

Figure 8: XM-3600 Prompt

[6] https://github.com/EleutherAI/lm-evaluation-harness/pull/885

```
⟨ TASK PROMPT. Refer to each task below. ⟩
{
Learning example #i:
Turn ID: turn_id
Database: db_id
Context: gold_context
Answer: gold_answer
} for i in range(k) # (in-context examples)
Target example #i:
Turn ID: turn_id
Database: db_id
Context: gold_context
Answer: ⟨model-completion-here⟩
```

Figure 9: General prompt structure for X-RiSAWOZ

```
You are a helpful NLP assistant solving the "Task Oriented Dialogue"
problem. In particular, you are solving the "Dialogue State
Prediction" subtask. In Dialogue State Prediction, you must describe
what is the state of the dialogue given the history using SQL-like
structure. The syntax can be understood from the examples below.
Based on the learning examples given below, complete the "Answer"
part of the target example. Do not print any additional information.
```

Figure 10: Task prompt for "DST" subtask in X-RiSAWOZ

```
You are a helpful NLP assistant solving the "Task Oriented Dialogue"
problem. In particular, you are solving the "API Call Detection"
subtask. In API call detection, your task is to identify whether
the dialogue can be continued with whatever context we already have.
"yes" here means that additional data must be queried using an API
for continuing the dialog while "no" means that API call is not
required. Based on the learning examples given below, complete the
"Answer" part of the target example. Do not print any additional
information.
```

Figure 11: Task prompt for "API" subtask in X-RiSAWOZ

```
You are a helpful NLP assistant solving the "Task Oriented Dialogue"
problem. In particular, you are solving the "Dialogue Act
Prediction" subtask. In Dialogue Act Generation, you must generate
the next dialogue action based on the given context. This will be an
SQL-like structure. The syntax can be understood from the examples
below. Based on the learning examples given below, complete the
"Answer" part of the target example. Do not print any additional
information.
```

Figure 12: Task prompt for "DA" subtask in X-RiSAWOZ

```
You are a helpful NLP assistant solving the "Task Oriented Dialogue"
problem. In particular, you are solving the "Response Generation"
subtask. In Response Generation, your task is to produce a
natural language response from the chatbot given the context of the
conversation. Based on the learning examples given below, complete
the "Answer" part of the target example. Do not print any additional
information.
```

Figure 13: Task prompt for "RG" subtask in X-RiSAWOZ

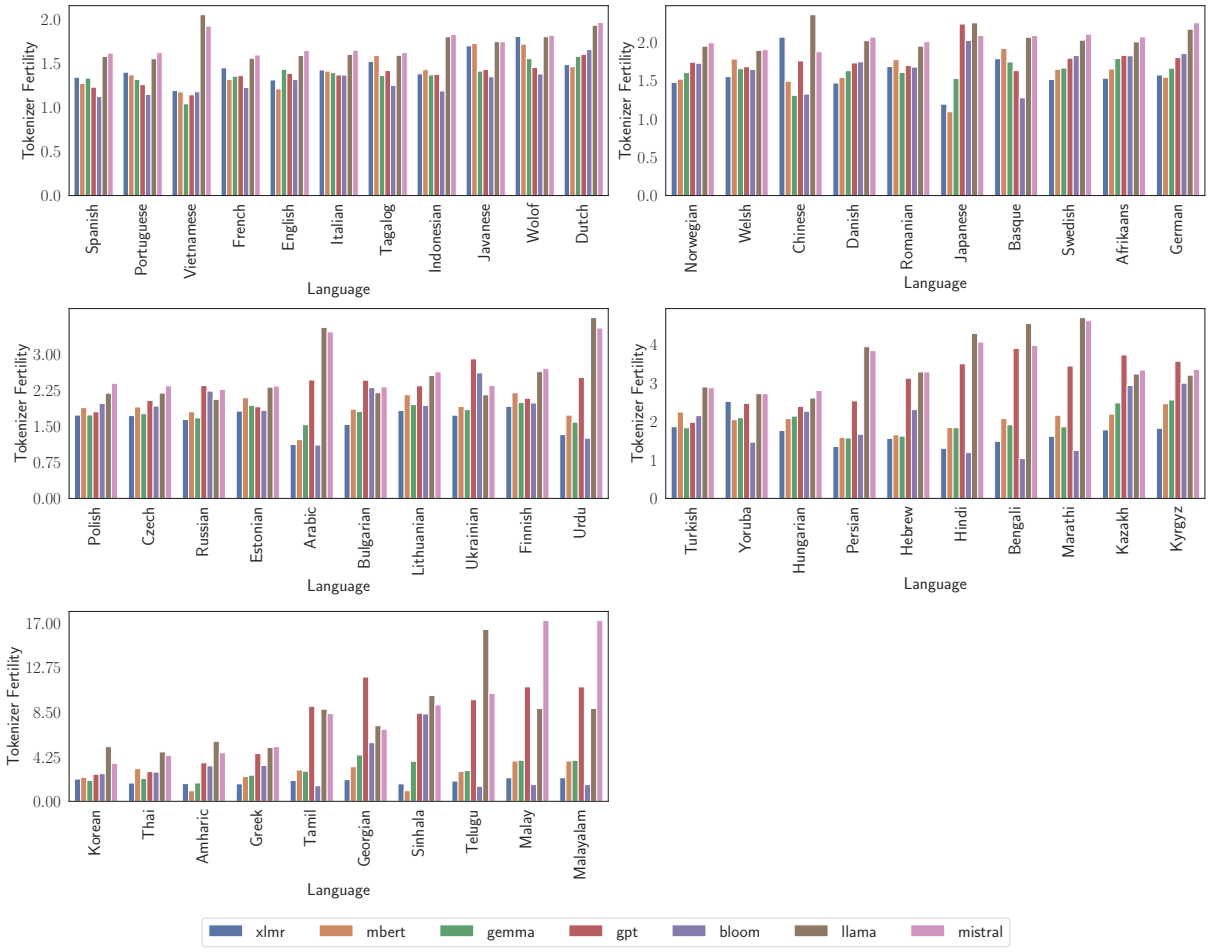

Figure 14: Fertility analysis was performed for all assessed models, with the exception of PaLM2 and Gemini, which was excluded due to a lack of available information about its tokenizer. (Ács, 2019)

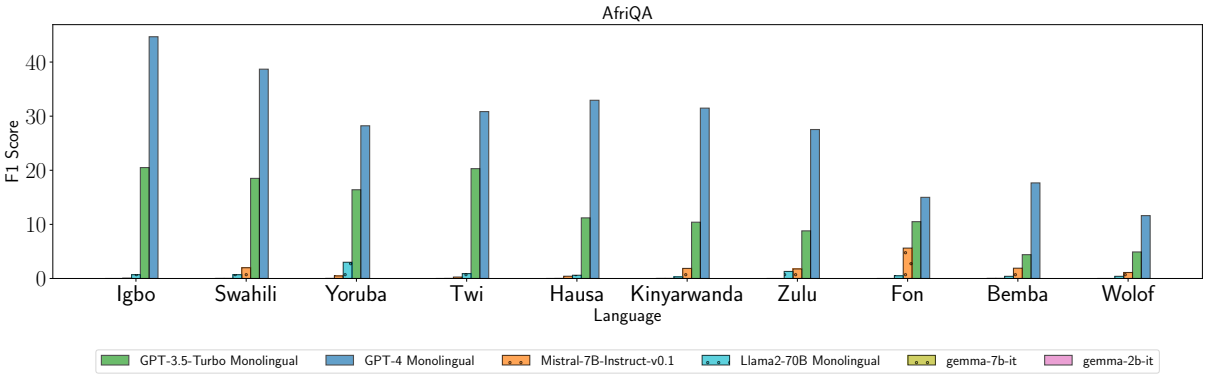

Figure 15: Results for AfriQA across all languages and models for monolingual prompting

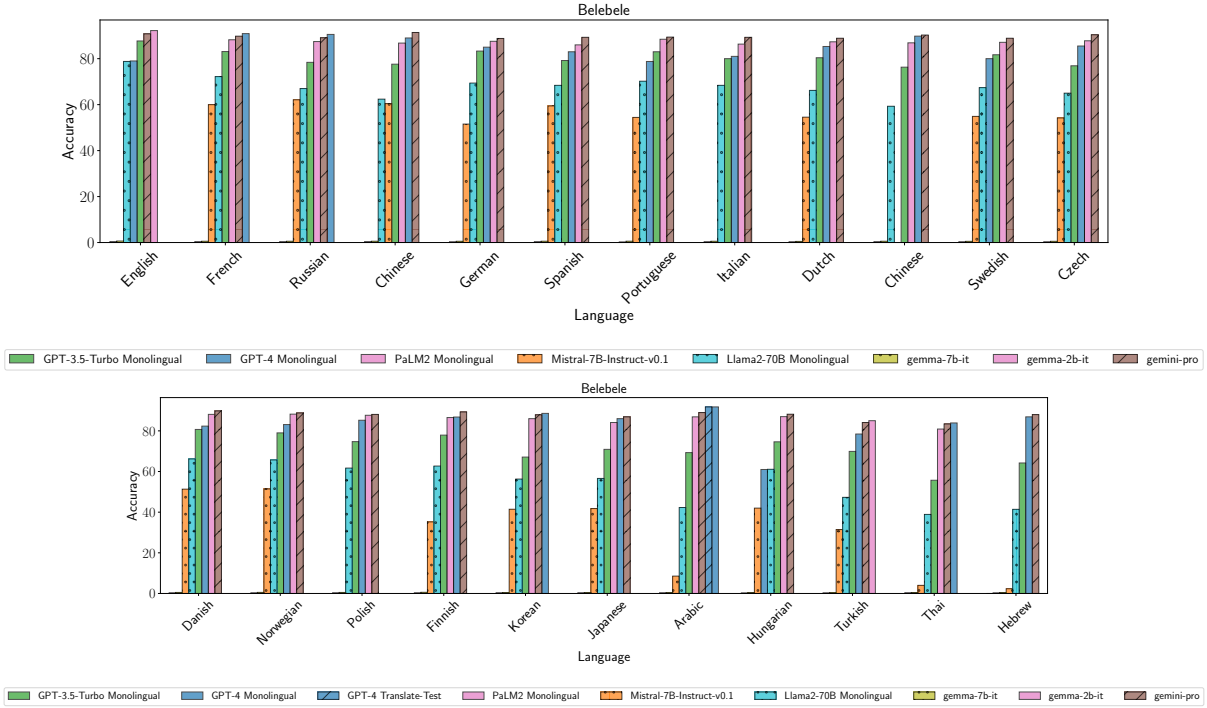

Figure 16: Results for Belebele across all languages and models for monolingual prompting

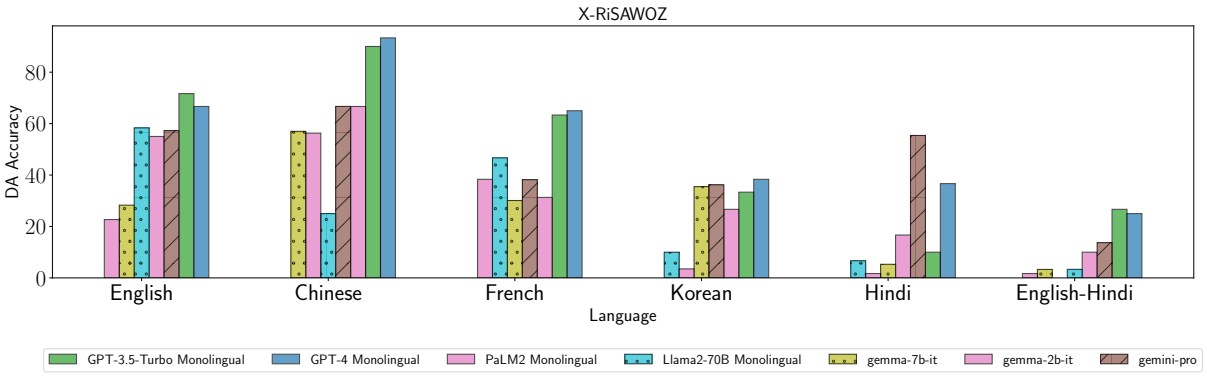

Figure 17: Results for X-RiSAWOZ across all languages and models for monolingual prompting

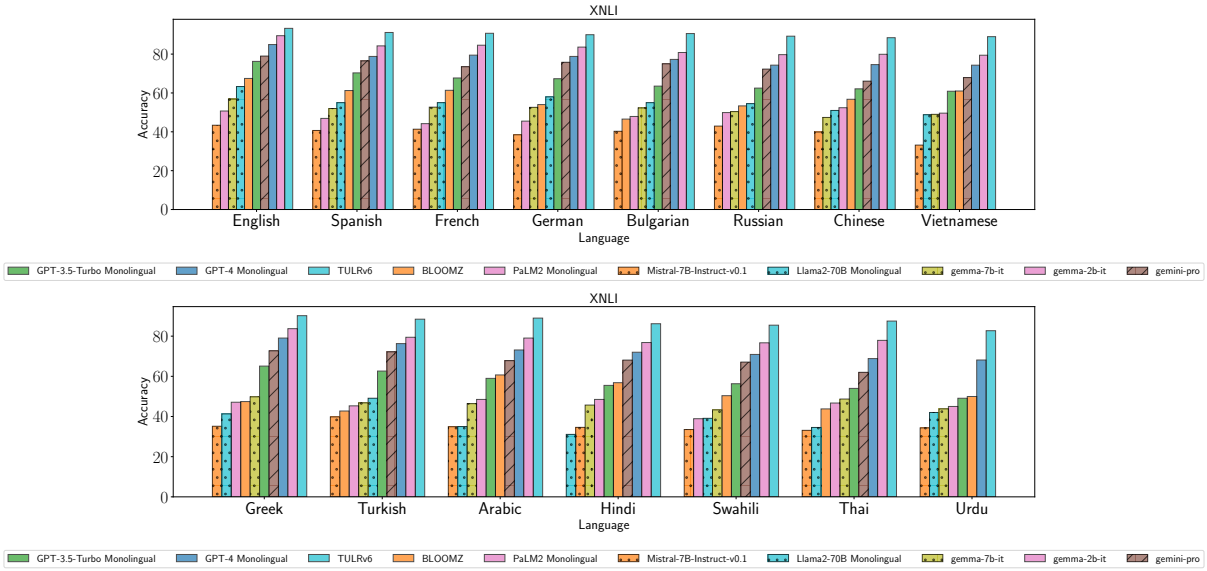

Figure 18: Results for XNLI across all languages and models for monolingual prompting

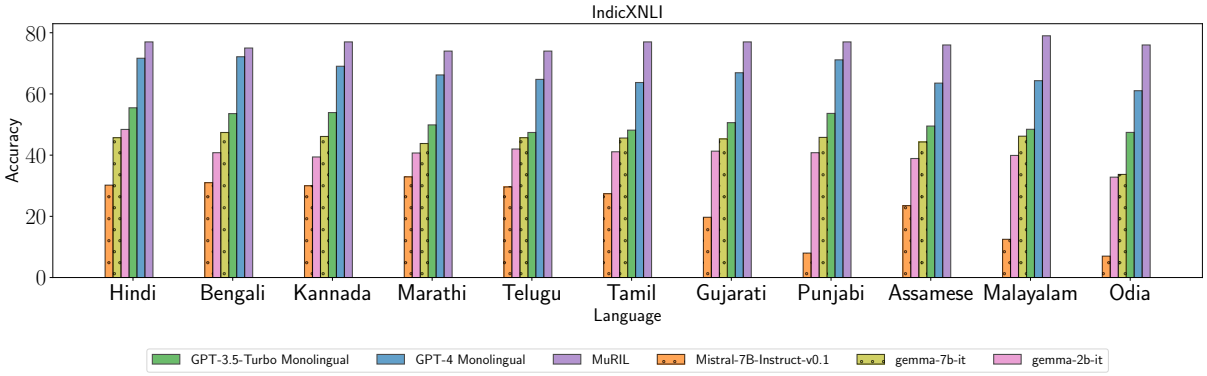

Figure 19: Results for IndicXNLI across all languages and models for monolingual prompting

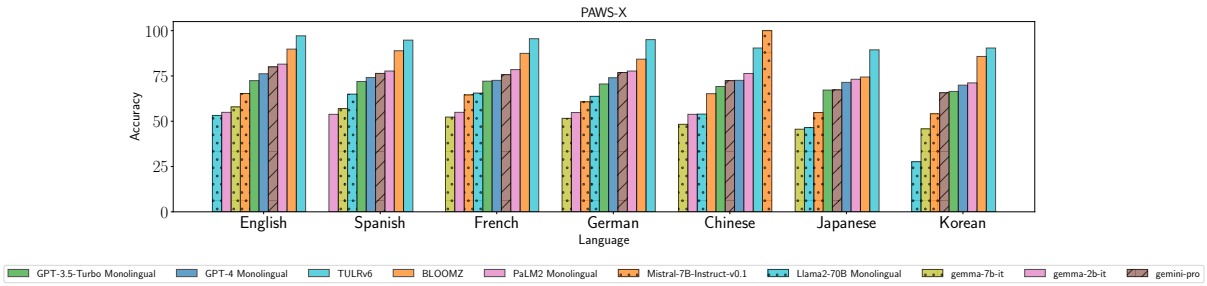

Figure 20: Results for PAWSX across all languages and models for monolingual prompting

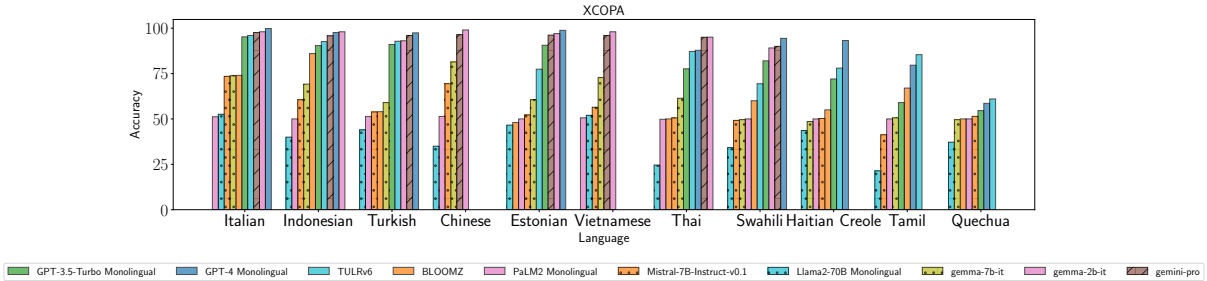

Figure 21: Results for XCOPA across all languages and models for monolingual prompting

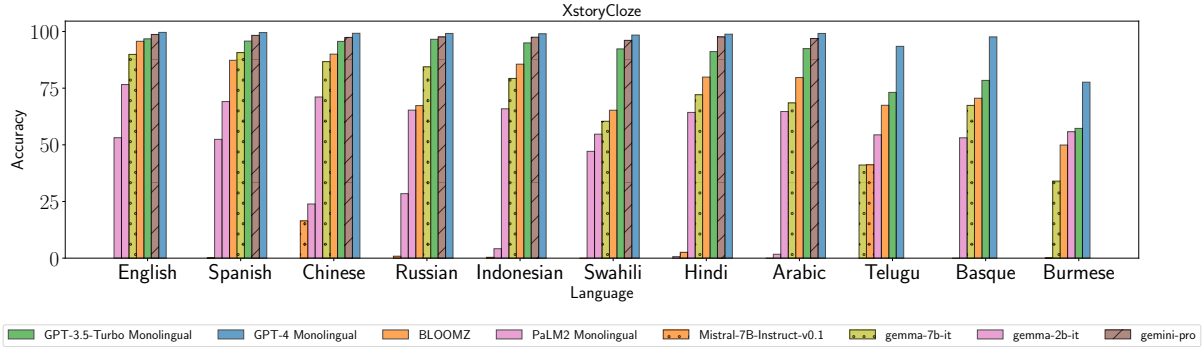

Figure 22: Results for XStoryCloze across all languages and models for monolingual prompting

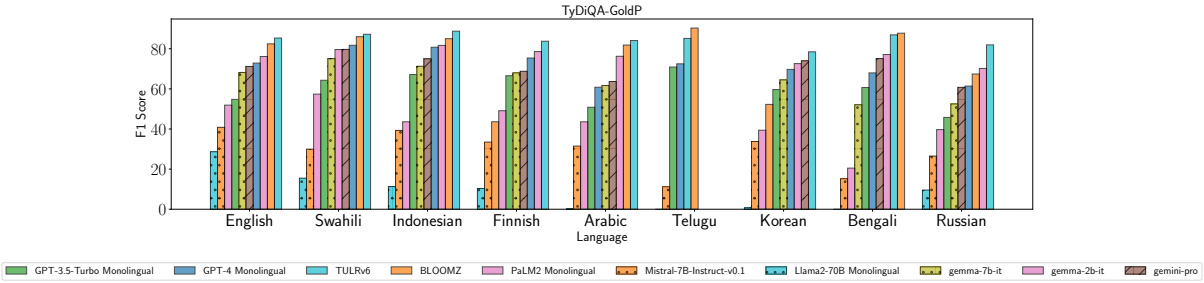

Figure 23: Results for TyDiQA across all languages and models for monolingual prompting

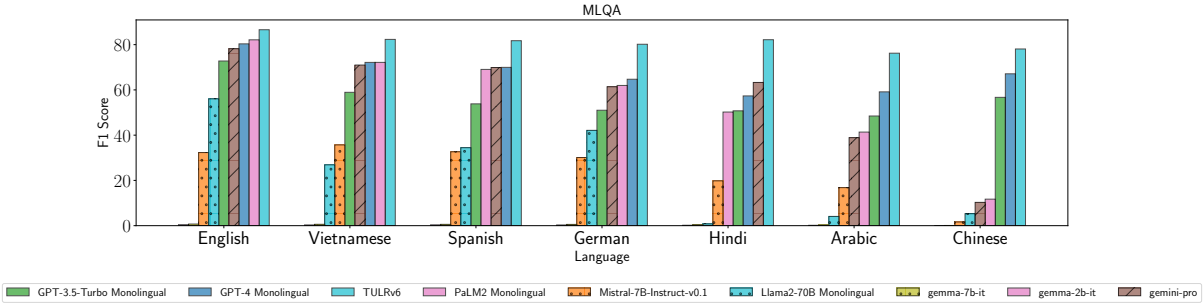

Figure 24: Results for MLQA across all languages and models for monolingual prompting

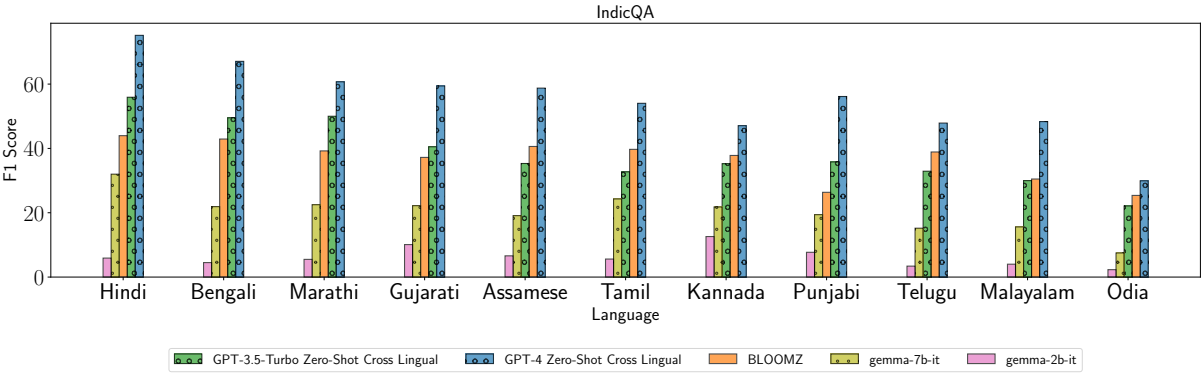

Figure 25: Results for IndicQA across all languages and models with zero-shot cross-lingual prompting

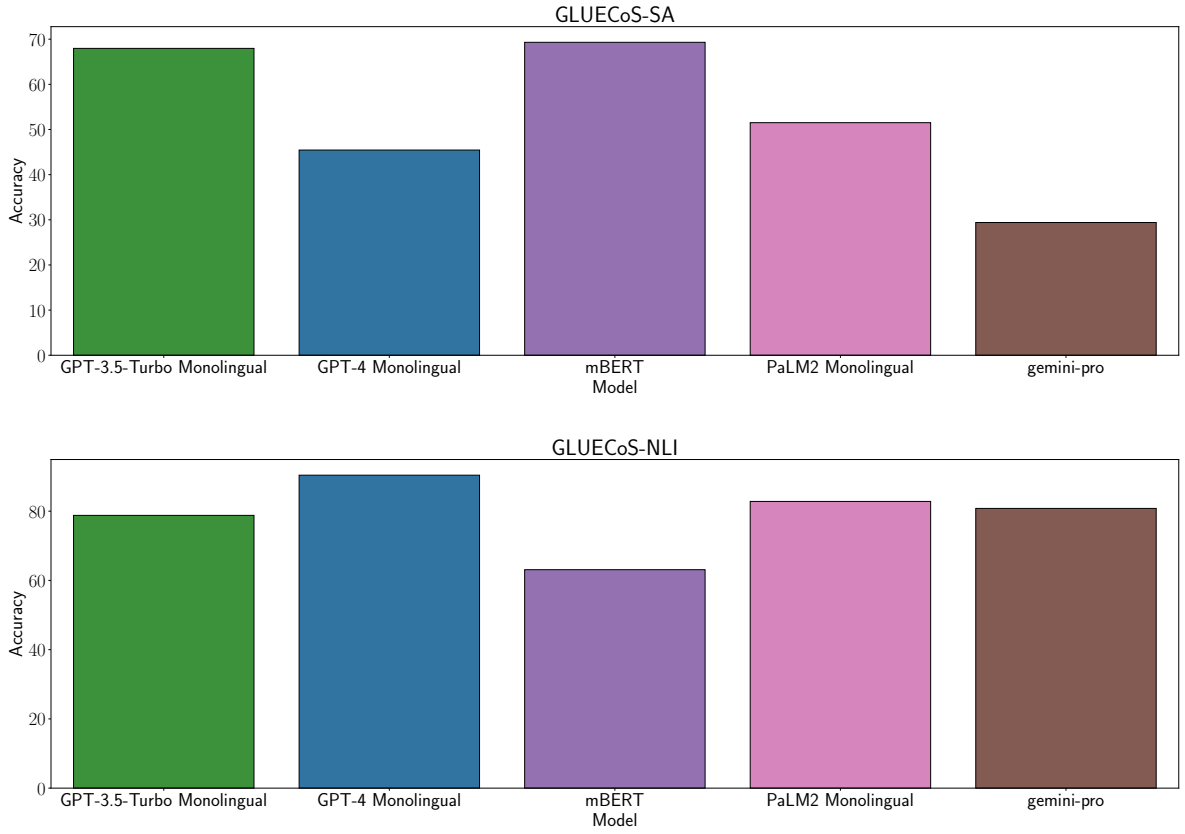

Figure 26: Results for the GLUECoS dataset on the Sentiment Classification (English-Spanish, En-Es-CS) and the NLI (English-Hindi) task

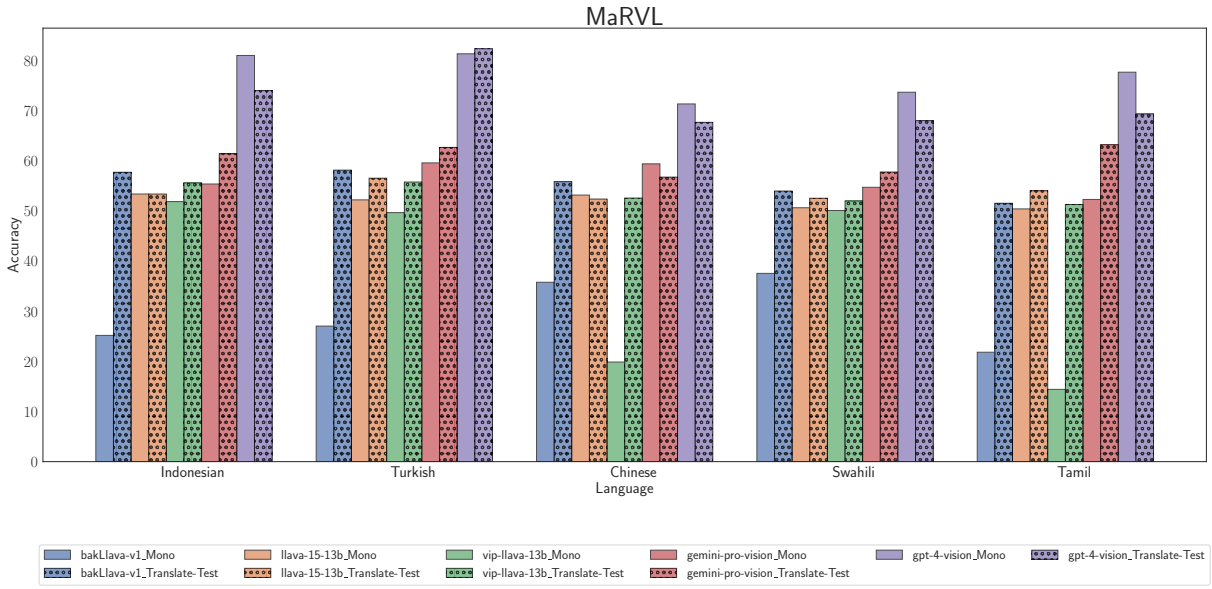

Figure 27: Accuracy scores for the LLaVA models, GPT4-Vision, and Gemini-Pro-Vision on MaRVL. We used two prompting strategies, monolingual and translate-test.

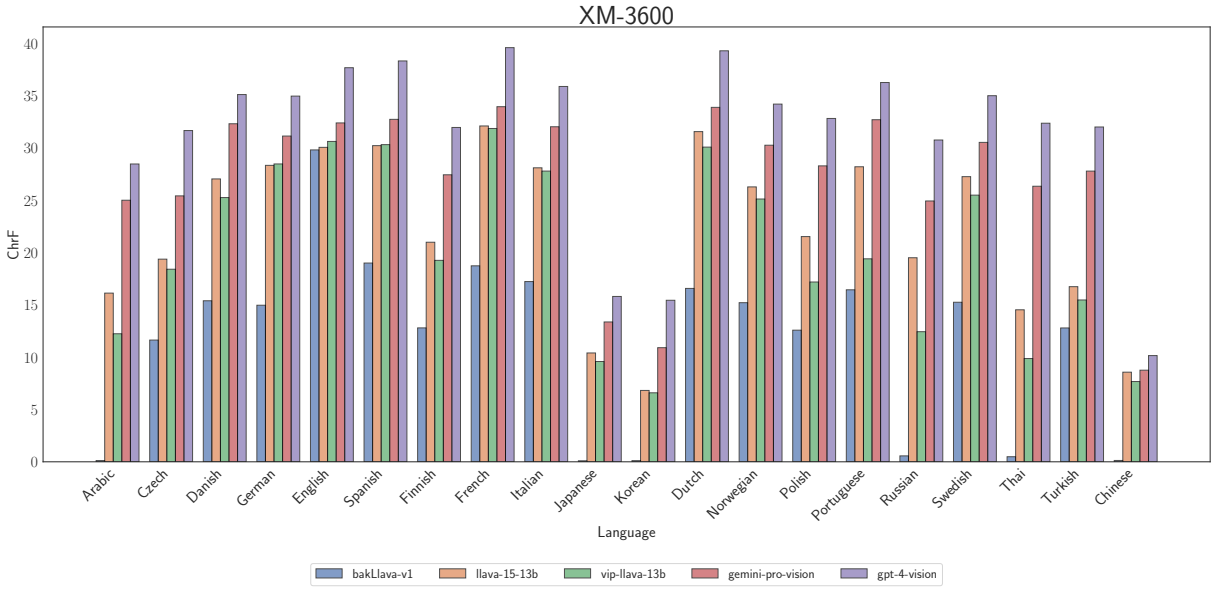

Figure 28: chrF scores for the LLaVA models, GPT4-Vision, and Gemini-Pro-Vision on XM-3600. We use monolingual prompting as the prompting strategy.

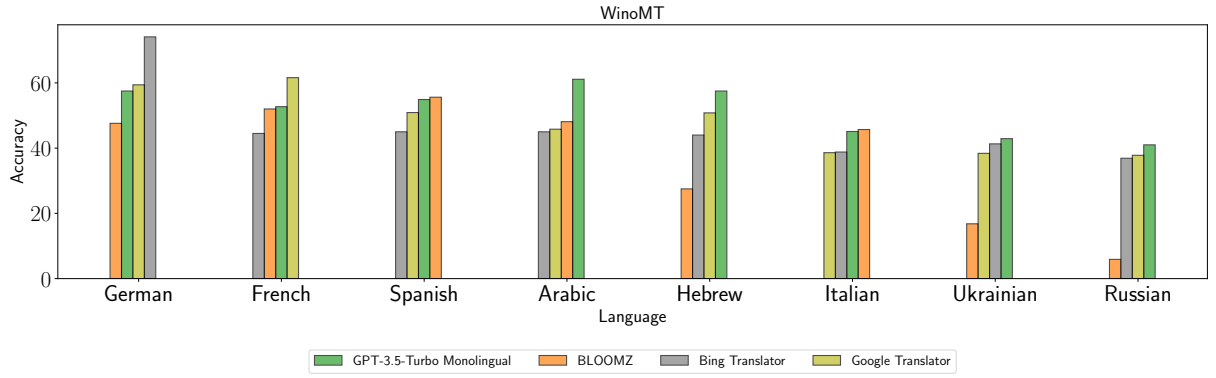

Figure 29: Results for WinoMT across all languages and models for monolingual prompting

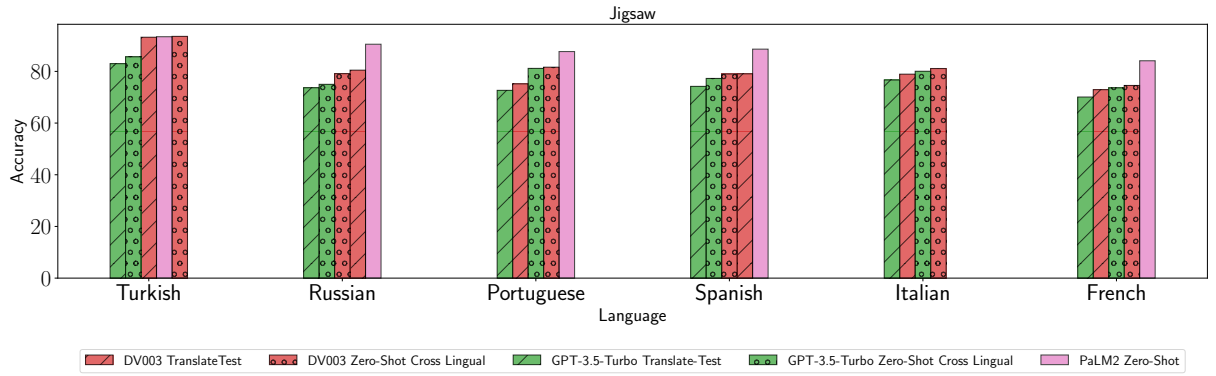

Figure 30: Results for Jigsaw across all languages and models for monolingual prompting

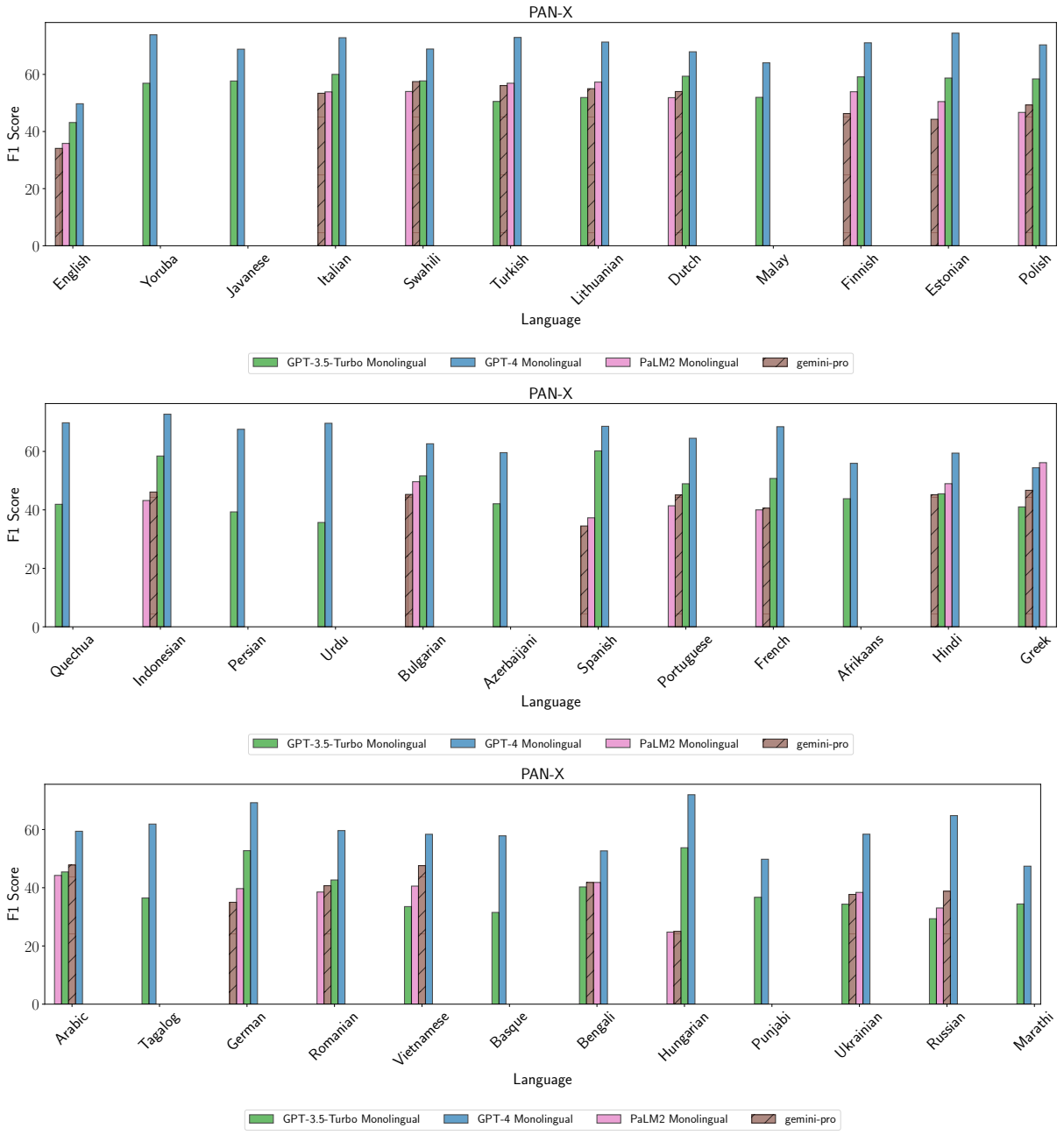

Figure 31: Results for PAN-X across all languages with monolingual prompting

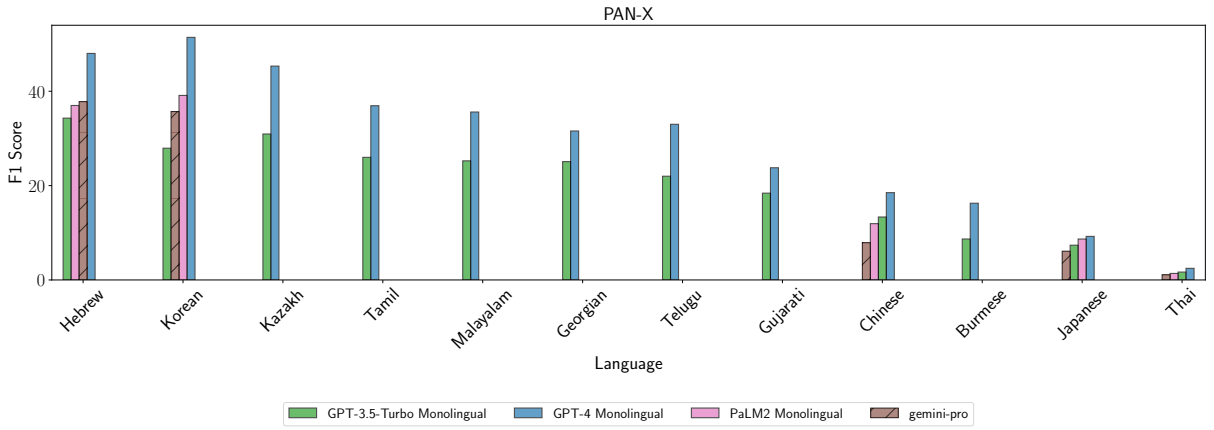

Figure 32: Results for PAN-X across all languages with monolingual prompting

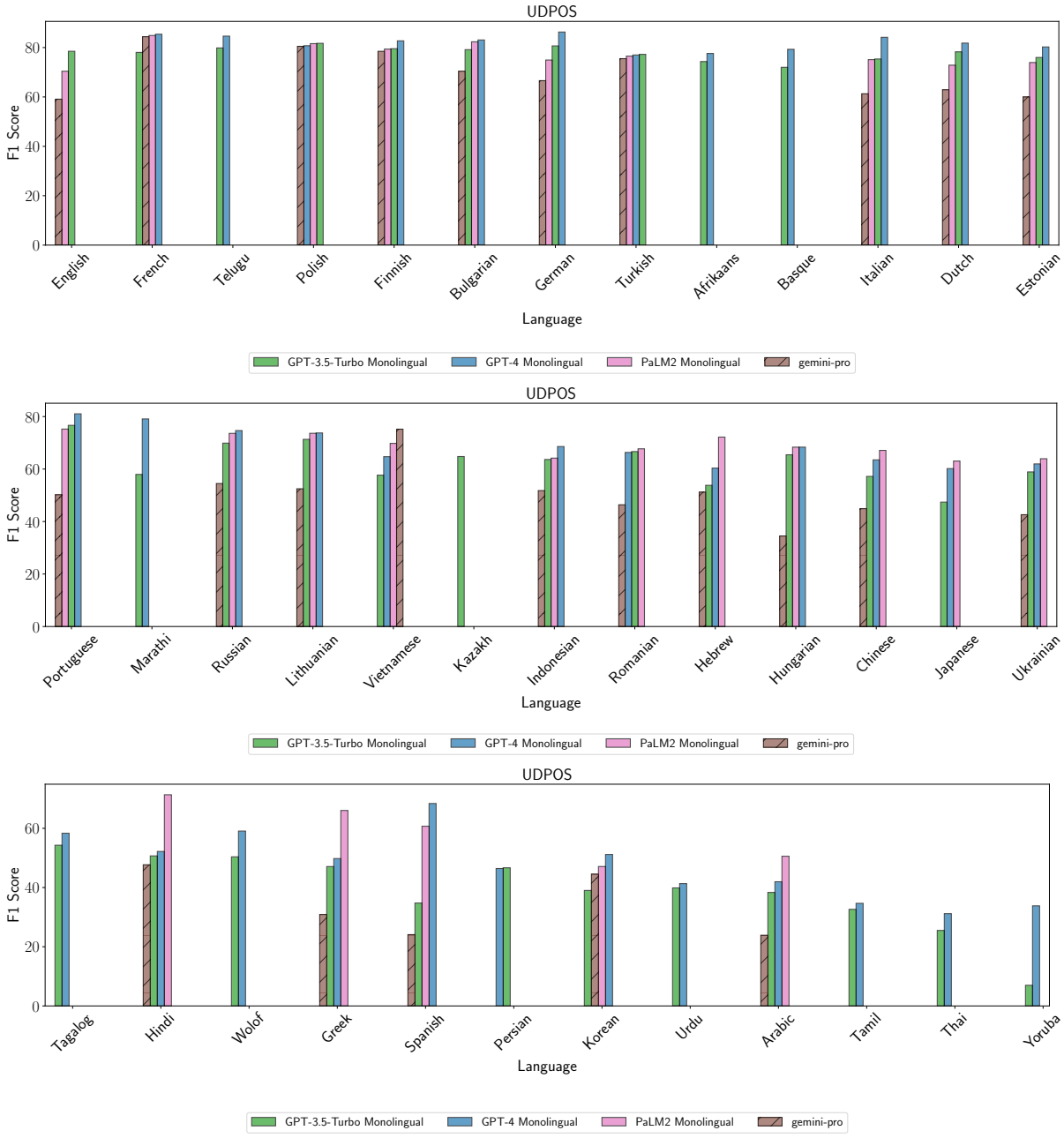

Figure 33: Results for UDPOS across all languages with monolingual prompting

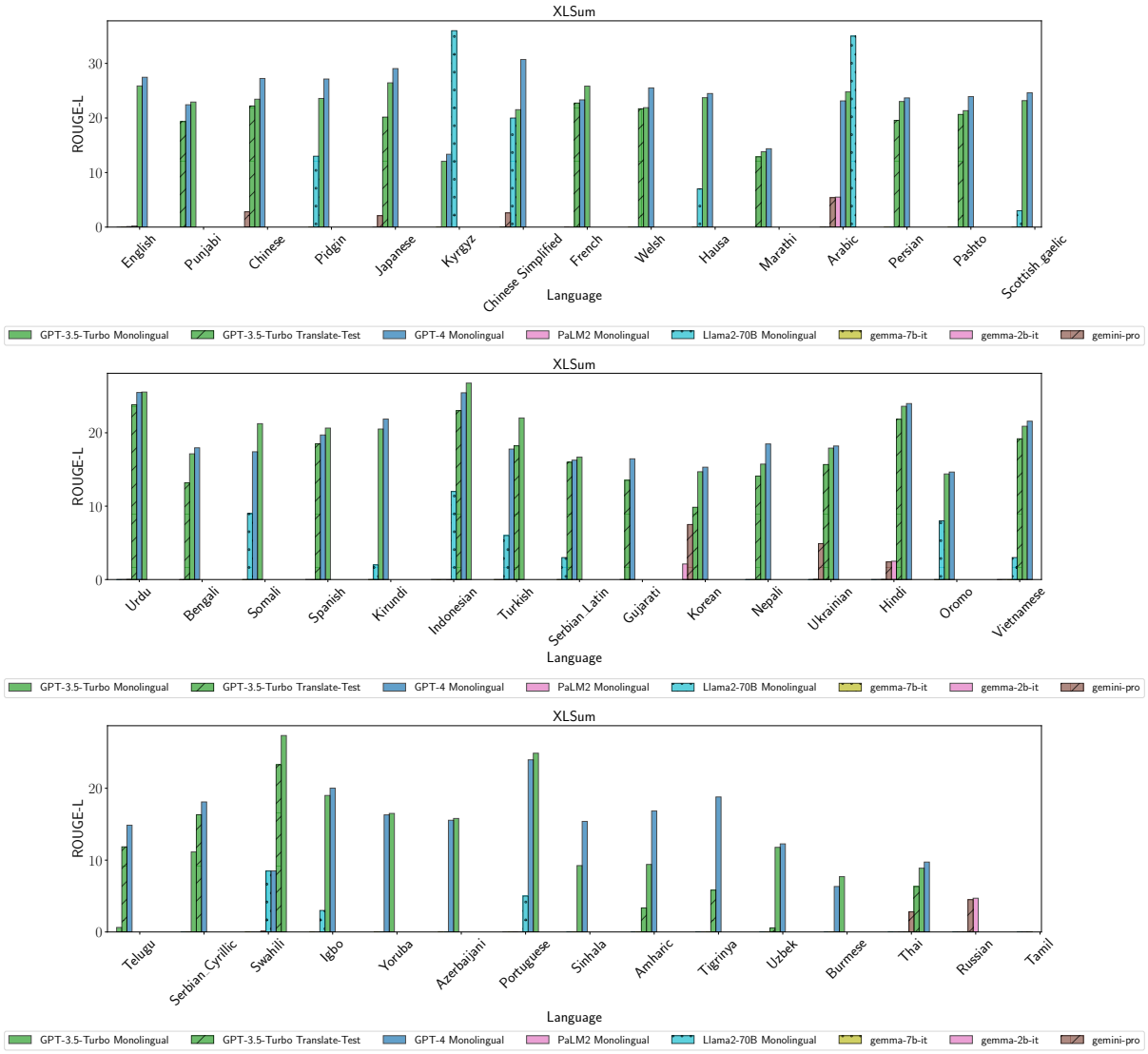

Figure 34: Results for XLSUM across all languages and models with monolingual prompting

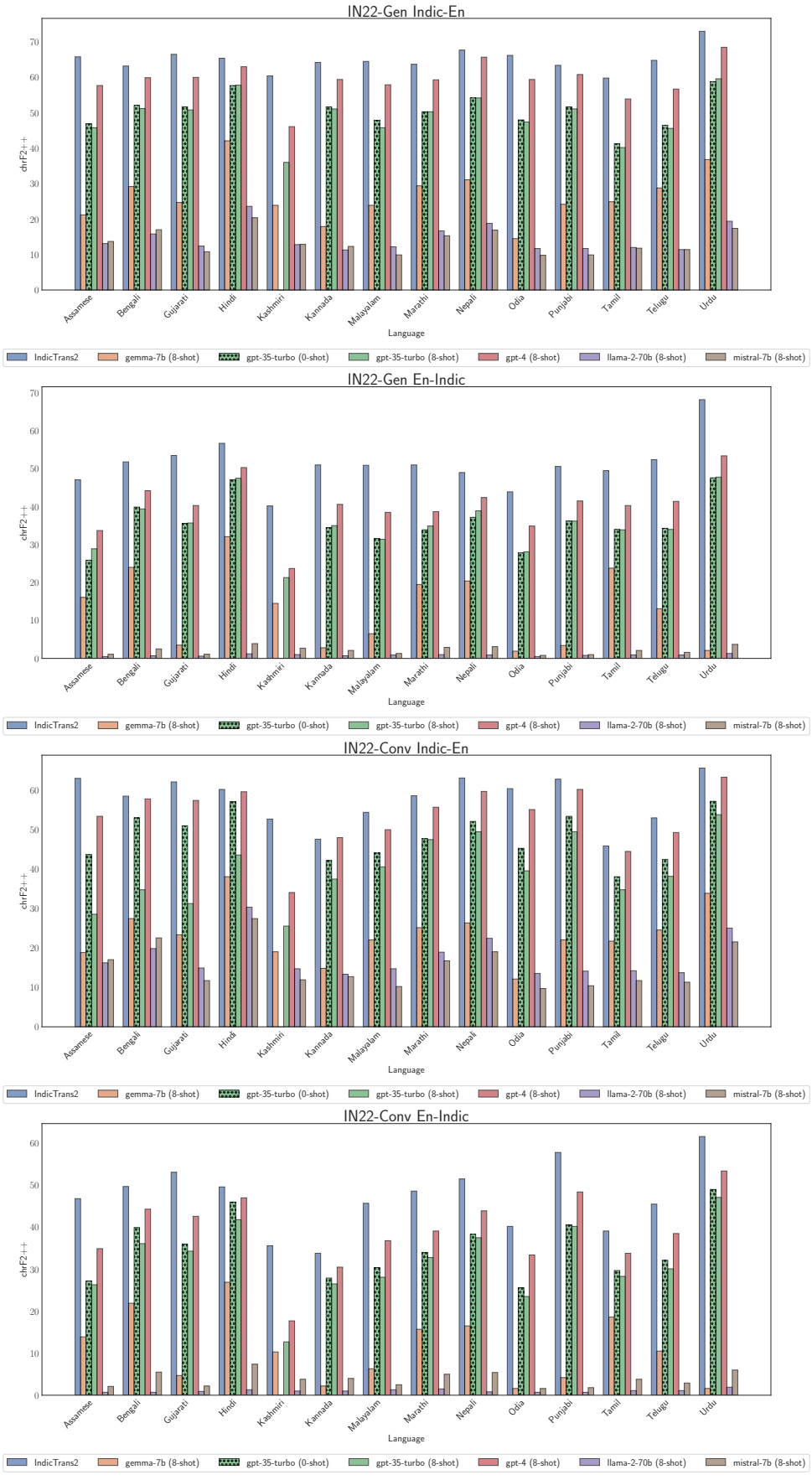

Figure 35: chrF++ scores of IN22. Note that, Kashmiri 0-shot was not covered in Gala et al. (2023)

Model	en	ar	bg	de	el	es	fr	hi	ru	sw	th	tr	ur	vi	zh	**avg**
Fine-tuned Baselines																
mBERT	80.8	64.3	68.0	70.0	65.3	73.5	73.4	58.9	67.8	49.7	54.1	60.9	57.2	69.3	67.8	65.4
mT5-Base	84.7	73.3	78.6	77.4	77.1	80.3	79.1	70.8	77.1	69.4	73.2	72.8	68.3	74.2	74.1	75.4
XLM-R Large	88.7	77.2	83.0	82.5	80.8	83.7	82.2	75.6	79.1	71.2	77.4	78.0	71.7	79.3	78.2	79.2
TuLRv6 - XXL	**93.3**	**89.0**	**90.6**	**90.0**	**90.2**	**91.1**	**90.7**	**86.2**	**89.2**	**85.5**	**87.5**	**88.4**	**82.7**	**89.0**	**88.4**	**88.8**
Prompt-Based Baselines																
BLOOMZ	67.5	60.7	46.5	54.0	47.4	61.2	61.4	56.8	53.3	50.4	43.8	42.7	50.0	61.0	56.7	54.2
XGLM	52.6	46.4	48.9	45.6	48.7	45.8	49.4	46.8	48.6	44.5	46.6	45.4	43.4	48.5	48.8	47.3
Llama 2 7B	56.3	39	45	45	39	50	50	37	48	33	35	40	36	41	45	38.9
Llama 2 13B	55	37	51	50	0	51	52	0	50	36	0	45.2	29	48	48	33.1
Llama 2 70B	63.3	35	55	58	41.3	55	55	31.1	54.5	39.1	34.55	49.1	42.0	48.8	51.0	43.3
Mistral 7B	41.1	38.5	41.8	41.6	38.1	44.7	49.1	35.7	40.1	33.8	36	35.3	34.8	35.6	39.1	39.0
Mistral 7B Instruct	43.4	35.0	40.3	38.5	35.2	40.6	41.4	34.6	43.0	33.5	33.1	39.8	34.3	33.2	40.0	37.8
Google Models																
PaLM 2	89.5	79.0	80.8	83.6	83.7	84.2	84.6	76.8	79.7	76.7	77.9	79.4	-	79.4	79.9	76.4
gemini-pro	79.0	67.8	75.0	75.8	72.7	76.5	73.5	68.1	72.3	67.1	62	72.2	-	67.9	66.1	71.1
Gemma 2B Instruct	50.7	48.5	47.9	45.5	47.1	46.9	44.2	48.5	49.9	38.9	46.7	45.3	45.0	49.6	52.4	47.1
Gemma 7B Instruct	56.9	46.4	52.3	52.6	49.8	52	52.7	45.7	50.4	43.3	48.7	46.8	43.8	49	47.4	49.2
Open AI Models																
`gpt-3.5-turbo`	76.2	59.0	63.5	67.3	65.1	70.3	67.7	55.5	62.5	56.3	54.0	62.6	49.1	60.9	62.1	62.1
`gpt-3.5-turbo` (TT)	76.2	62.7	67.3	69.4	67.2	69.6	69.0	59.9	63.7	55.8	59.6	63.8	54.0	63.9	62.6	64.3
`text-davinci-003`	79.5	52.2	61.8	65.8	59.7	71.0	65.7	47.6	62.2	50.2	51.1	57.9	50.0	56.4	58.0	59.3
`text-davinci-003` (TT)	79.5	65.1	70.8	71.7	69.3	72.2	71.8	63.3	67.3	57.3	62.0	67.6	55.1	66.9	65.8	67.1
`gpt-4-32k`	84.9	73.1	77.3	78.8	79.0	78.8	79.5	72.0	74.3	70.9	68.8	76.3	68.1	74.3	74.6	75.4

Table 2: Comparing performance of different models on all languages in XNLI. Metric: Accuracy. Unsupported languages are marked with '-'. Averages are calculated only from supported languages.

Model	as	bn	gu	hi	kn	ml	mr	or	pa	ta	te	**avg**
Fine-tuned Baselines												
MuRIL	**76.0**	**75.0**	**77.0**	**77.0**	**77.0**	**79.0**	**74.0**	**76.0**	**77.0**	**77.0**	**74.0**	**76.0**
Open AI Models												
`gpt-3.5-turbo`	49.5	53.6	50.6	55.5	53.9	48.4	49.9	47.4	53.6	48.2	47.4	50.7
`gpt-3.5-turbo` (TT)	54.3	61.6	61.8	59.6	60.8	59.9	58.7	58.5	62.3	58.3	60.8	59.7
`text-davinci-003`	48.6	52.6	51.2	56.9	49.1	48.2	49.4	46.4	50.4	45.5	47.2	49.6
`text-davinci-003` (TT)	56.0	66.0	64.7	62.6	63.9	61.8	60.9	60.8	64.7	61.8	63.1	62.4
`gpt-4-32k`	63.5	72.2	66.9	71.7	69.0	64.3	66.2	61.1	71.1	63.7	64.8	66.8
Open Source Models												
`Mistral 7B`	34.4	45.5	32.2	46	37.3	21.5	38.8	0	9.9	41.2	13.3	29.1
`Mistral 7B Instruct`	23.5	30.7	19.7	30.2	30	12.5	32.9	7	8	27.4	29.6	28.3
Google Models												
Gemma 2B Instruct	38.9	40.8	41.3	48.4	39.4	39.9	40.7	32.8	40.8	41.1	42	40.5
Gemma 7B Instruct	44.3	47.4	45.3	45.7	46.1	46.2	43.8	33.7	45.8	45.6	45.7	44.5

Table 3: Comparing performance of different models on all languages in IndicXNLI. Metric: Accuracy.

Model	en	de	es	fr	ja	ko	zh	**avg**
Fine-tuned Baselines								
mBERT	94.0	85.7	87.4	87.0	73.0	69.6	77.0	81.9
mT5-Base	95.4	89.4	89.6	91.2	79.8	78.5	81.1	86.4
XLM-R Large	94.7	89.7	90.1	90.4	78.7	79.0	82.3	86.4
TuLRv6 - XXL	**97.2**	**95.1**	**94.8**	**95.6**	**89.4**	**90.4**	**90.4**	**93.2**
Prompt-Based Baselines								
BLOOMZ	89.8	84.3	88.9	87.5	74.4	85.8	65.2	82.3
Llama 2 7B	68.4	65.1	67	67	56	53.8	60.5	62.5
Llama 2 13B	63.3	52.3	57.7	54	0	0	6	33.3
Llama 2 70B	53.2	63.8	65	65.6	46.5	27.7	54	53.7
Mistral 7B	64.2	68.6	67.4	63.6	53.4	49.3	53.3	60
Mistral 7B Instruct	65.3	60.7	64.6	64.5	54.8	54.1	55.1	59.9
Google Models								
PaLM 2	81.5	77.7	77.7	78.5	73.2	71.2	76.4	76.6
gemini-pro	80.0	76.9	76.4	75.7	67.3	65.7	72.4	73.5
Gemma 2B Instruct	54.9	54.8	53.8	54.9	53.4	55.1	53.8	54.4
Gemma 7B Instruct	57.9	51.6	57.0	52.3	45.6	45.8	48.3	51.2
Open AI Models								
`gpt-3.5-turbo`	72.4	70.6	72.0	72.1	67.2	66.5	69.2	70.0
`gpt-3.5-turbo` (TT)	72.4	70.8	69.7	70.1	61.9	62.5	63.1	67.2
`text-davinci-003`	72.5	70.6	72.7	70.7	60.6	61.8	60.8	67.1
`text-davinci-003` (TT)	72.5	69.8	70.1	71.3	65.4	65.8	65.2	68.6
`gpt-4-32k`	76.2	74.0	74.1	72.6	71.5	69.9	72.6	73.0

Table 4: Comparing performance of different models on all languages in PAWS-X. Metric: Accuracy.

Model	en	et	ht	id	it	qu	sw	ta	th	tr	**avg**
Fine-tuned Baselines											
mT5-Base	-	50.3	49.9	49.2	49.6	50.5	50.4	49.2	50.7	49.5	49.9
TuLRv6 - XXL	-	77.4	78.0	92.6	96.0	61.0	69.4	85.4	87.2	92.8	74.0
Prompt-Based Baselines											
BLOOMZ	88.0	48.0	55.0	86.0	74.0	50.0	60.0	67.0	50.0	54.0	63.2
XGLM	-	65.9	58.9	68.9	69.2	47.1	62.9	56.3	62.0	58.5	61.1
Llama 2 7B	74	50.6	51.2	59	70.6	50.4	50.6	49.6	52	52.6	56.0
Llama 2 13B	91	51.2	49.4	72.4	79.8	50.2	50.4	0	0	54	49.8
Llama 2 70B	94	46.6	43.6	40	52.6	37.2	34.2	21.4	24.6	44	43.8
Mistral 7B	91	55.8	58	86.2	51.8	49.7	51.2	60.4	65.6	72.2	63
Mistral 7B Instruct	92	53.6	52.4	62	72.4	50.4	49.8	29.1	50.8	53.7	57.7
Google Models											
PaLM 2	99.1	97.0	-	**98.0**	98.0	-	89.1	-	**95.0**	93.1	**95.6**
gemini-pro	99.0	96.2	-	95.8	97.6	-	90.0	-	**95.0**	96.0	**95.6**
Gemma 2B Instruct	61.0	50.0	50.0	50.0	51.2	50.0	50.0	50.0	49.8	51.3	51.3
Gemma 7B Instruct	92.0	60.6	48.6	69.2	73.8	49.6	49.6	50.6	61.4	59.0	61.4
Open AI Models											
`gpt-3.5-turbo`	97.8	90.6	72.0	90.4	95.2	54.6	82.0	59.0	77.6	91.0	81.0
`gpt-3.5-turbo` (TT)	97.8	88.2	79.4	90.8	94.4	50.0	77.6	87.0	82.2	87.8	83.5
`text-davinci-003`	98.2	87.8	75.0	91.4	96.0	54.8	63.6	53.8	66.6	87.8	77.5
`text-davinci-003` (TT)	98.2	89.6	82.8	93.0	94.6	50.0	82.8	87.0	84.8	89.8	85.3
`gpt-4-32k`	**99.6**	**98.8**	**93.2**	97.6	**99.8**	58.6	**94.4**	79.6	87.8	**97.4**	90.7
`gpt-4-32k` (TT)	**99.6**	94.4	85.8	96.0	98.2	**85.8**	83.4	**91.4**	87.8	92.2	90.6

Table 5: Comparing performance of different models on all languages in XCOPA. Metric: Accuracy. Unsupported languages are marked with '-'. Averages are calculated only from supported languages.

Model	en	ar	de	el	es	hi	ru	th	tr	vi	zh	**avg**
Fine-tuned Baselines												
mBERT	83.5 / 72.2	61.5 / 45.1	70.6 / 54.0	62.6 / 44.9	75.5 / 56.9	59.2 / 46.0	71.3 / 53.3	42.7 / 33.5	55.4 / 40.1	69.5 / 49.6	58.0 / 48.3	64.5 / 49.4
mT5-Base	84.6 / 71.7	63.8 / 44.3	73.8 / 54.5	59.6 / 35.6	74.8 / 56.1	60.3 / 43.4	57.8 / 34.7	57.6 / 45.7	67.9 / 48.2	70.7 / 50.3	66.1 / 54.1	67.0 / 49.0
XLM-R Large	86.5 / 75.7	68.6 / 49.0	80.4 / 63.4	79.8 / 61.7	82.0 / 63.9	76.7 / 59.7	80.1 / 64.3	74.2 / 62.8	75.9 / 59.3	79.1 / 59.0	59.3 / 50.0	76.6 / 60.8
TuLRv6 - XXL	90.1 / 80.6	**85.4 / 69.6**	**86.1 / 70.4**	**86.3 / 70.4**	**87.6 / 71.0**	**85.9 / 70.5**	**86.8 / 73.2**	**87.0 / 81.1**	**84.3 / 71.0**	**87.6 / 71.3**	79.2 / 73.2	**86.0 / 72.9**
Prompt-Based Baselines												
BLOOMZ	**92.1 / 83.8**	82.8 / 69.7	76.3 / 60.4	49.7 / 37.6	86.8 / 71.4	83.4 / 72.9	65.7 / 47.2	20.5 / 15.5	51.4 / 37.2	86.9 / 72.7	**82.4 / 78.6**	70.7 / 58.8
Llama 2 7B	19.1 / 0.1	1.9 / 0.1	12.0 / 0.0	1.7 / 0.1	13.8 / 0.1	0.7 / 0.0	4.6 / 0.6	0.6 / 0.0	4.9 / 0.1	7.0 / 0.0	7.0 / 0.5	6.7 / 0.1
Llama 2 13B	17.0 / 0.0	1.8 / 0.1	11.3 / 0.0	1.5 / 0.0	13.3 / 0.0	0.7 / 0.0	4.1 / 0.6	0.5 / 0.0	0.5 / 0.0	7.5 / 0.0	6.4 / 0.5	5.9 / 0.1
Llama 2 70B	19.8 / 0.0	1.9 / 0.0	10.7 / 0.0	1.7 / 0.0	13.9 / 0.0	6.3 / 0.0	4.6 / 0.6	0.4 / 0.0	4.8 / 0.0	7.3 / 0.0	7.1 / 0.6	7.1 / 0.1
Mistral 7B	33.4 / 0.7	22.1 / 2.7	28.9 / 0.8	21.3 / 2.9	33.5 / 1.2	17.8 / 2.3	27.2 / 1.3	15.8 / 3.4	25.9 / 1.6	35.4 / 3.5	16.6 / 1.3	25.9 / 2
Mistral 7B Instruct	34.3 / 0.7	84.7 / 1.8	28.1 / 0.5	11.5 / 0.8	31.4 / 0.8	17.7 / 1.8	27.1 / 2	11.6 / 1.6	23.1 / 1.8	31 / 2.3	14.1 / 0.9	23.3 / 1.3
Google Models												
PaLM 2	86.7 / 76.4	63.1 / 48.4	75.9 / 59.2	76.5 / 55.7	81.3 / 62.8	61.7 / 47.0	69.8 / 52.7	61.9 / 54.1	68.1 / 52.1	72.5 / 52.4	48.8 / 44.6	69.7 / 55.0
gemini-pro	83.1 / 72.2	68.8 / 52.7	73.7 / 56.7	67.9 / 52.4	77.1 / 59.6	71.6 / 57.2	62.9 / 47.3	68.9 / 62.8	67.9 / 51.4	77 / 59.4	59.4 / 55.9	71.1 / 57.5
Gemma 2B Instruct	24 / 6.8	14.8 / 2.2	18.2 / 4.9	17.7 / 7	19.7 / 3.4	10.8 / 3	14.6 / 2.6	17.6 / 9.3	15 / 3	22.9 / 4.9	12.5 / 4.5	17.1 / 4.5
Gemma 7B Instruct	73.3 / 56.3	42.4 / 24.3	60.6 / 41.8	44.4 / 28.3	64.6 / 38.2	41.4 / 25.6	51.4 / 31.2	42.2 / 29.7	44.7 / 23.2	64.7 / 41.5	40.5 / 33.4	52.5 / 34.6
Open AI Models												
`gpt-3.5-turbo`	79.3 / 58.7	59.6 / 35.1	70.6 / 46.6	49.0 / 22.8	70.3 / 40.8	54.0 / 29.0	58.0 / 31.3	41.9 / 30.4	61.8 / 35.0	69.1 / 42.4	50.4 / 48.3	60.4 / 38.2
`text-davinci-003`	77.2 / 61.8	36.8 / 22.5	55.2 / 39.7	31.8 / 19.7	61.8 / 41.3	19.9 / 10.0	29.4 / 17.6	11.5 / 8.7	44.8 / 29.2	41.7 / 25.4	35.6 / 32.8	40.5 / 28.1
`gpt-4-32k`	83.2 / 65.6	67.8 / 42.4	71.9 / 48.7	62.3 / 36.6	77.5 / 50.7	63.9 / 36.7	63.8 / 35.8	54.6 / 42.0	70.8 / 46.6	75.8 / 49.7	60.0 / 57.5	68.3 / 46.6

Table 6: Comparing performance of different models on all languages in XQUAD. Metric: F1 Score / Exact Match.

Model	en	ar	bn	fi	id	ko	ru	sw	te	**avg**
Fine-tuned Baselines										
mBERT	75.3 / 63.6	62.2 / 42.8	49.3 / 32.7	59.7 / 45.3	64.8 / 45.8	58.8 / 50.0	60.0 / 38.8	57.5 / 37.9	49.6 / 38.4	59.7 / 43.9
mT5-Base	71.8 / 60.9	67.1 / 50.4	40.7 / 22.1	67.0 / 52.2	71.3 / 54.5	49.5 / 37.7	54.9 / 32.6	60.4 / 43.9	40.6 / 31.1	58.1 / 42.8
XLM-R Large	71.5 / 56.8	67.6 / 40.4	64.0 / 47.8	70.5 / 53.2	77.4 / 61.9	31.9 / 10.9	67.0 / 42.1	66.1 / 48.1	70.1 / 43.6	65.1 / 45.0
TuLRv6 - XXL	**85.4 / 76.4**	**84.1 / 70.4**	86.9 / 79.6	**83.8 / 72.8**	**88.8 / 77.9**	**78.5 / 67.8**	**81.9 / 68.6**	**87.2 / 79.6**	85.2 / 71.6	**84.6 / 73.8**
Prompt-Based Baselines										
BLOOMZ	82.4 / 70.9	81.9 / 62.2	**87.8 / 82.3**	43.6 / 28.6	85.0 / 71.0	52.3 / 43.1	67.4 / 51.5	86.0 / 77.2	**90.3 / 81.6**	75.2 / 63.2
Llama 2 7B	21.0 / 0.2	0.2 / 0.0	0.0 / 0.0	10.6 / 0.3	12.7 / 0.0	0.6 / 0.0	7.7 / 0.5	14.5 / 0.0	0.0 / 0.0	7.5 / 0.1
Llama 2 13B	27.6 / 0.0	0.2 / 0.0	0.0 / 0.0	10.2 / 0.3	10.4 / 0.0	0.8 / 0.0	6.3 / 0.2	14.6 / 0.0	0.0 / 0.0	7.8 / 0.1
Llama 2 70B	28.7 / 0.2	0.2 / 0.0	0.0 / 0.0	10.4 / 0.3	11.3 / 0.0	0.8 / 0.0	9.6 / 0.6	15.5 / 0.0	0.0 / 0.0	8.5 / 0.1
Mistral 7B	44.9 / 1.6	40.6 / 2.9	36.1 / 1.8	35.8 / 0.8	35.9 / 0.5	41.8 / 0	28.7 / 0.6	27 / 0	13.9 / 0.45	33.8 / 1
Mistral 7B Instruct	40.6 / 0.7	31.5 / 2	15.2/0.9	33.5 / 0.5	39.3 / 1.1	33.7 / 0	26.5 / 0.6	29.9 / 0.4	11.2 / 0.45	29.6 / 0.8
Google Models										
PaLM 2	76.2 / 62.3	76.3 / 51.7	77.2 / 62.8	78.6 / 62.3	81.7 / 66.2	72.6 / 64.1	70.2 / 54.3	79.6 / 69.1	- / -	76.5 / 61.6
gemini-pro	71.2 / 59.3	63.7 / 54.1	75.1 / 66.4	68.8 / 60.7	75 / 64.2	74 / 65.6	60.8 / 48.9	79.7 / 73.1	- / -	71.0 / 61.5
Gemma 2B Instruct	51.9 / 29.5	43.6 / 15.9	20.5 / 4.4	49.1 / 30.8	43.6 / 20.4	39.4 / 22.8	39.7 / 15.6	57.4 / 44.7	- / -	43.1 / 23.0
Gemma 7B Instruct	68.2 / 51.6	61.7 / 37.6	52.1 / 32.7	68 / 53.5	71.3 / 53.5	64.5 / 52.5	52.5 / 31.4	75.1 / 62.9	- / -	64.2 / 47.0
Open AI Models										
`gpt-3.5-turbo`	54.8 / 30.7	50.9 / 24.2	60.7 / 32.7	66.6 / 49.0	67.2 / 43.4	59.7 / 45.3	45.8 / 20.0	64.3 / 47.7	70.9 / 53.1	60.1 / 38.4
`text-davinci-003`	73.7 / 59.1	56.2 / 38.7	16.1 / 10.6	70.3 / 58.8	68.6 / 51.2	40.6 / 32.2	42.3 / 28.9	74.1 / 62.3	5.8 / 3.0	49.8 / 38.3
`gpt-4-32k`	72.9 / 51.4	60.8 / 32.7	68.0 / 42.5	75.4 / 57.7	80.8 / 61.1	69.7 / 58.5	61.4 / 30.5	81.8 / 68.7	72.5 / 54.9	71.5 / 50.9

Table 7: Comparing performance of different models on all languages in TyDiQA. Metric: F1 Score / Exact Match. Unsupported languages are marked with '-'. Averages are calculated only from supported languages.

Model	**en**	**ar**	**de**	es	hi	vi	zh	**avg**
Fine-tuned Baselines								
mBERT	80.2 / 67.0	52.3 / 34.6	59.0 / 43.8	67.4 / 49.2	50.2 / 35.3	61.2 / 40.7	59.6 / 38.6	61.4 / 44.2
mT5-Base	81.7 / 66.9	57.1 / 36.9	62.1 / 43.2	67.1 / 47.2	55.4 / 37.9	65.9 / 44.1	61.6 / 38.6	64.4 / 45.0
XLM-R Large	83.5 / 70.6	66.6 / 47.1	70.1 / 54.9	74.1 / 56.6	70.6 / 53.1	74.0 / 52.9	62.1 / 37.0	71.6 / 53.2
TuLRv6 - XXL	**86.6 / 74.4**	**76.2 / 56.5**	**80.2 / 67.0**	**81.7 / 65.1**	**82.2 / 64.8**	**82.3 / 63.2**	**78.1 / 56.5**	**81.0 / 63.9**
Prompt-Based Baselines								
Llama 2 7B	69.6 / 50.7	5.2 / 1.2	44.5 / 27.3	51.9 / 29.8	0.5 / 0.0	42.5 / 27.4	7.0 / 4.2	31.6 / 20.1
Llama 2 13B	62.2 / 38.7	55.0 / 36.0	73.1 / 52.5	62.2 / 38.7	0.4 / 0.0	9.1 / 7.6	9.1 / 7.6	38.8 / 25.9
Llama 2 70B	56.1 / 36.4	4.1 / 0.5	42.1 / 24.1	34.5 / 12.9	0.8 / 0.0	26.9 / 12.6	5.3 / 3.3	24.2 / 12.8
Mistral 7B	35.3 / 1	23.2 / 0.8	25.1 / 0.5	34.3 / 0.8	29.9 / 2.4	40.1 / 1.2	1.9 / 0	27.2 / 1
Mistral 7B Instruct	32.4 / 1.1	17.1 / 0.6	29.6 / 0.4	32.5 / 0.6	19.5 / 1.2	36.4 / 1.7	1.48 / 0	24.14 / 0.8
Google Models								
PaLM 2	82.1 / 68.8	41.4 / 22.3	62.0 / 43.9	69.1 / 43.7	50.2 / 35.5	72.2 / 49.4	11.7 / 8.9	55.5 / 38.9
gemini-pro	78.2 / 66.4	38.9 / 21.7	61.4 / 46.3	69.9 / 48.4	63.3 / 47.3	71 / 51.3	10.3 / 7.5	56.1 / 41.3
Gemma 2B Instruct	43.4 / 20.7	16.5 / 7	31.1 / 16	36.6 / 8.7	18.4 / 8.7	32 / 13.9	5.1 / 1.6	26.2 / 10.9
Gemma 7B Instruct	72 / 53.5	30.3 / 16.1	56 / 40	58.4 / 34.2	40.7 / 24.5	60.5 / 38.4	9.1 / 5.6	46.7 / 30.3
Open AI Models								
`gpt-3.5-turbo`	72.8 / 53.2	48.5 / 23.9	51.0 / 29.6	53.8 / 29.4	50.7 / 28.9	58.9 / 35.1	56.7 / 29.4	56.1 / 32.8
`gpt-3.5-turbo` (TT)	72.8 / 53.2	37.8 / 18.4	44.3 / 26.2	54.1 / 31.8	37.3 / 20.0	41.6 / 22.5	36.5 / 17.2	46.4 / 27.0
`text-davinci-003`	74.8 / 59.0	38.4 / 21.7	57.7 / 38.1	62.9 / 37.8	24.9 / 14.1	47.7 / 29.7	32.3 / 31.7	48.4 / 33.1
`text-davinci-003` (TT)	74.8 / 59.0	48.2 / 25.6	53.5 / 33.9	62.9 / 40.9	49.2 / 28.7	51.0 / 30.4	45.2 / 24.1	55.0 / 34.7
`gpt-4-32k`	80.3 / 62.8	59.1 / 33.5	64.7 / 44.4	70.0 / 45.9	57.3 / 35.6	72.2 / 49.0	67.1 / 38.4	67.2 / 44.2

Table 8: Comparing performance of different models on all languages in MLQA. Metric: F1 Score / Exact Match.

Model	as	bn	gu	hi	kn	ml	mr	or	pa	ta	te	**avg**
Fine-tuned Baselines												
BLOOMZ	40.6 / 31.7	42.9 / 36.6	37.2 / 29.9	44.0 / 45.1	37.8 / 26.6	30.5 / 28.4	39.2 / 33.0	25.4 / 22.0	26.4 / 33.5	39.7 / 35.9	38.9 / 34.7	36.6 / 32.5
Google Models												
Gemma 2B Instruct	6.6 / 5.5	4.5 / 2.9	10.1 / 9.6	5.9 / 1.8	12.6 / 11.6	4 / 3	5.5 / 3.9	2.3 / 2.3	7.7 / 5.1	5.6 / 2.4	3.4 / 2.5	6.2 / 4.6
Gemma 7B Instruct	19.1 / 12.6	21.9 / 14.9	22.2 / 20.1	32 / 21.8	21.8 / 17.4	15.6 / 12.6	22.5 / 16.4	7.5 / 7.4	19.4 / 15	24.3 / 18.1	15.2 / 12.6	20.1 / 15.4
Open AI Models												
`gpt-3.5-turbo`	35.3 / 21.4	49.5 / 30.2	40.5 / 25.5	55.9 / 39.3	35.3 / 20.4	30.0 / 19.2	50.0 / 32.0	22.1 / 12.7	35.8 / 15.1	32.7 / 21.6	32.9 / 19.7	38.2 / 23.4
`text-davinci-003`	6.7 / 3.2	10.3 / 5.8	5.4 / 3.5	16.8 / 11.8	7.1 / 3.9	3.6 / 2.3	14.6 / 8.5	6.9 / 3.4	10.7 / 4.1	4.2 / 2.5	6.8 / 3.6	8.4 / 4.8
`gpt-4-32k`	**58.8 / 40.4**	**67.1 / 47.4**	**59.4 / 42.4**	**75.2 / 62.2**	**47.1 / 31.6**	**48.3 / 33.7**	**60.7 / 43.1**	**29.9 / 16.7**	**56.1 / 34.1**	**54.0 / 39.7**	**47.9 / 27.8**	**55.0 / 38.1**

Table 9: Comparing performance of different models on all languages in IndicQA. Metric: F1 Score / Exact Match.

Model	bem	fon	hau	ibo	kin	swa	twi	wol	yor	zul	**avg**
Open Source Baselines											
Llama 2 7B	0.4	0.5	0.6	0.7	0.3	0.7	0.9	0.4	2.9	1.3	0.9
Llama 2 13B	0.4	0.5	0.6	0.7	0.3	0.7	0.9	0.4	3	1.3	0.9
Llama 2 70B	0.4	0.5	0.6	0.7	0.3	0.7	0.9	0.4	3	1.3	0.9
Mistral 7B	1.9	0	1.7	1.4	1.9	4.3	12	0.8	1.4	2.9	2.83
Mistral 7B Instruct	1.9	5.6	0.4	0.1	1.9	2	0.3	1.1	0.5	1.8	1.56
Google Models											
Gemma 2B Instruct	0.9	0.0	0.0	0.0	0.2	0.4	0.4	0.0	0.5	0.0	0.2
Gemma 7B Instruct	0.9	0.0	0.4	0.0	0.3	0.5	0.0	0.0	0.0	0.0	0.2
Open AI Models											
`gpt-3.5-turbo`	4.4	10.5	11.2	20.5	10.4	18.5	20.3	4.9	16.4	8.8	12.6
`gpt-4-32k`	**17.7**	**15.0**	**32.9**	**44.7**	**31.5**	**38.7**	**30.8**	**11.6**	**28.2**	**27.5**	**27.9**

Table 10: Comparing performance of different models on all languages in AfriQA. Metric: F1 Score.

Model	en	af	ar	bg	de	el	es	et	eu	fa	fi	fr	he	hi	hu	id	it	ja	kk	
Fine-tuned Baselines																				
mBERT	96.4	86.7	50.0	84.7	88.7	80.9	86.6	79.9	62.1	65.5	73.3	81.2	55.5	66.0	78.6	74.2	87.8	47.2	70.4	
XLM-R Large	**97.0**	**89.2**	**63.0**	**88.3**	**91.2**	**86.5**	**89.2**	**87.3**	74.9	**70.8**	**82.7**	**86.7**	67.5	**75.2**	**83.4**	**75.7**	**89.2**	29.3	**78.3**	
Google Models																				
PaLM 2	70.4	-	50.6	82.3	74.9	66.0	60.7	73.9	-	-	79.4	84.9	**72.2**	71.3	68.3	64.2	75.1	63.1	-	
gemini-pro	59.1	-	23.9	70.4	66.6	30.9	24.1	60.0	-	-	78.5	84.4	51.2	47.7	34.5	51.8	61.3	-	-	
Open AI Models																				
`gpt-3.5-turbo`	78.5	74.3	38.3	79.1	80.7	47.1	34.8	76.0	72.0	46.7	79.5	78.0	53.8	50.7	65.4	63.6	75.4	47.4	64.8	
`gpt-4-32k`	84.1	77.6	42.0	83.1	86.3	49.8	68.4	80.2	**79.3**	46.4	**82.7**	85.4	60.4	52.2	68.3	68.6	84.1	**60.2**	71.8	

	ko	lt	mr	nl	pl	pt	ro	ru	ta	te	th	tl	tr	uk	ur	vi	wo	yo	zh	**avg**
Fine-tuned Baselines																				
mBERT	51.7	78.8	68.7	88.6	80.7	88.0	71.5	82.4	58.5	75.2	41.3	80.5	70.5	80.6	56.6	55.4	0.0	**56.6**	59.6	71.9
XLM-R Large	**57.1**	**84.2**	**81.8**	**89.5**	**86.8**	**90.2**	**82.6**	**87.3**	**64.0**	84.2	**48.5**	**92.4**	**81.2**	**85.8**	**70.8**	58.5	0.0	24.8	44.1	**76.2**
Google Models																				
PaLM 2	47.1	73.6	-	72.9	81.6	75.2	67.7	73.6	-	-	-	-	76.5	63.9	-	**69.8**	-	-	**67.1**	70.2
gemini-pro	44.5	52.5	-	62.9	80.5	50.2	46.4	54.5	-	-	-	-	75.5	42.6	-	75.2	-	-	44.9	55.0
Open AI Models																				
`gpt-3.5-turbo`	39.0	71.3	57.9	78.3	81.7	76.7	66.7	69.9	32.6	79.8	25.5	54.3	77.2	58.9	39.9	57.7	50.4	7.0	57.2	60.2
`gpt-4-32k`	51.2	73.7	79.1	81.8†	80.7	81.0	66.3†	74.7	34.7	**84.6**	31.2†	58.4†	77.0	61.9	41.3	64.7	**59.1**	33.8†	63.5	66.6

Table 11: Comparing performance of different models on all languages in UDPOS. Metric: F1 Score. All numbers are Monolingual results except the ones marked with † symbol which indicate Zero-Shot Cross-Lingual results (due to the absence of training data in those languages. Unsupported languages are marked with '-'. Averages are calculated only from supported languages.

Model	en	af	ar	az	bg	bn	de	el	es	et	eu	fa	fi	fr	gu	he	hi	hu	id	it	ja	jv	ka	kk
Fine-tuned Baselines																								
mBERT	**86.4**	76.1	42.9	65.5	76.7	69.7	79.5	70.9	**75.3**	75.8	**64.4**	40.0	76.6	79.6	51.3	**56.2**	65.9	76.1	61.0	**81.3**	**29.2**	62.4	65.1	50.3
XLM-R Large	85.4	**78.6**	47.3	**69.4**	**80.9**	**74.7**	**80.7**	**79.2**	71.8	**78.7**	61.6	55.2	**79.6**	**79.8**	**62.7**	55.5	**70.9**	**80.2**	51.8	80.3	18.5	61.9	**70.9**	**54.4**
Google Models																								
PaLM 2	35.8	-	44.2	-	49.6	41.8	39.7	56.2	37.3	50.5	-	-	53.9	40.0	-	37.0	49.0	24.8	43.2	53.9	8.7	-	-	-
gemini-pro	34.1	-	47.8	-	45.3	41.8	35.0	46.7	34.5	44.3	-	-	46.3	40.6	-	37.8	45.2	25.0	46.1	53.4	6.1	-	-	-
Open AI Models																								
`gpt-3.5-turbo`	43.2	43.8	45.4	42.1	51.6	40.3	52.7	41.0	60.2	58.7	31.5	39.3	59.1	50.7	18.4	34.3	45.5	53.7	58.4	60.0	7.4	57.7	25.1	30.9
`gpt-4-32k`	49.7	55.9	**59.4**	59.6	62.6	52.7	69.2	54.4	68.6	74.4	57.8	**67.6**	71.1	68.5	23.8	48.0	59.4	71.9	**72.7**	72.8	9.2	**68.8**	31.6	45.3

	ko	lt	ml	mr	ms	my	nl	pa	pl	pt	qu	ro	ru	sw	ta	te	th	tl	tr	uk	ur	vi	yo	zh	**avg**
Fine-tuned Baselines																									
mBERT	**59.5**	**75.8**	53.0	57.0	67.1	45.7	81.0	30.5	79.2	**80.4**	58.5	74.0	63.9	**71.4**	50.7	48.9	0.4	72.6	73.4	69.7	35.4	74.5	45.8	**42.5**	62.3
XLM-R Large	59.2	**75.8**	**60.2**	**63.4**	**68.5**	**55.2**	**83.2**	49.4	**79.3**	79.9	58.5	**78.7**	**71.9**	68.9	**58.4**	**53.8**	0.7	**74.7**	**80.3**	**78.0**	60.3	**78.3**	37.0	26.6	**65.2**
Google Models																									
PaLM 2	39.1	57.3	-	-	-	-	51.9	-	46.7	41.4	-	38.5	33.0	54.0	-	-	1.4	-	56.9	38.4	-	40.6	-	11.9	40.6
gemini-pro	35.7	55	-	-	-	-	54	-	49.3	45.1	-	40.7	38.8	57.5	-	-	1.1	-	56.1	37.7	-	47.6	-	7.9	39.9
Open AI Models																									
`gpt-3.5-turbo`	27.9	51.9	25.2	34.4	52.0	8.7	59.4	36.7	58.4	48.9	41.9	42.7	29.4	57.7	26.0	22.0	1.7	36.5	50.5	34.4	35.7	33.5	56.9	13.3	40.3
`gpt-4-32k`	51.4	71.3	35.6	47.4	64.1	16.3	67.9	**49.8**	70.3	64.5	**69.8**	59.6	64.8	68.9	36.9	33.0	**2.5**	61.9	72.9	58.4	**69.6**	58.4	**73.9**	18.5	55.5

Table 12: Comparing performance of different models on all languages in PAN-X. Metric: F1 Score. Unsupported languages are marked with '-'. Averages are calculated only from supported languages.

Model	ar	en	es	eu	hi	id	my	ru	sw	te	zh	**avg**
Prompt-Based Baselines												
BLOOMZ	79.7	95.7	87.3	70.5	79.9	85.6	49.9	67.3	65.3	67.4	90.0	76.2
XGLM	59.8	75.9	69.2	63.8	62.5	70.8	61.2	72.4	65.2	63.4	67.7	66.5
Google Models												
PaLM 2	1.7	53.1	52.4	-	0.7	4.2	-	28.5	47.1	-	23.9	18.8
gemini-pro	96.9	98.7	98.3	-	97.6	97.5	-	97.6	96.1	-	97.4	**97.5**
Gemma 2B Instruct	64.7	76.6	69.1	53.1	64.3	65.9	55.8	65.3	54.7	54.4	71.1	63.2
Gemma 7B Instruct	68.5	89.9	90.7	67.4	72.1	79.3	34.0	84.4	60.4	41.1	86.7	70.4
Open AI Models												
`gpt-3.5-turbo`	92.5	96.8	95.8	78.4	91.1	95.0	57.2	96.6	92.3	73.1	95.6	87.7
`gpt-3.5-turbo` (TT)	94.3	96.8	96.1	92.5	94.7	95.2	88.6	96.2	88.7	93.6	95.6	93.9
`text-davinci-003`	87.4	98.3	97.6	78.1	77.8	96.4	47.4	94.2	78.1	57.6	95.0	82.5
`text-davinci-003` (TT)	95.0	98.3	96.2	94.1	95.1	95.9	90.1	96.9	90.7	94.3	96.2	94.8
`gpt-4-32k`	**99.1**	**99.6**	**99.5**	**97.6**	**98.8**	**99.0**	77.6	99.1	**98.4**	93.4	**99.2**	96.5
`gpt-4-32k` (TT)	97.7	**99.6**	98.7	96.8	97.9	98.1	**93.2**	**99.2**	93.6	**96.4**	98.3	97.0

Table 13: Comparing performance of different models on all languages in XStoryCloze. Metric: Accuracy. Unsupported languages are marked with '-'. Averages are calculated only from supported languages.

Model	NLI En-Hi	Sentiment En-Es
Fine-tuned Baselines		
mBERT	63.1	**69.31**
Google Models		
PaLM 2	82.8	51.5
gemini-pro	80.8	29.4
Open AI Models		
`text-davinci-003`	72.1	68.8
`gpt-3.5-turbo`	78.8	68.0
`gpt-4-32k`	**90.4**	45.5

Table 14: Comparing performance of different models on code-mixing datasets from Khanuja et al. (2020b). Metric: Accuracy.

Model	en	am	ar	az	bn	cy	es	fa	fr	gd	gu	ha	hi	id	ig	ja	ko	ky	mr	my	ne	om	pa
Fine-tuned Baselines																							
mT5-RL	**35.1**	**24.6**	31.2	24.4	**24.2**	**29.9**	**27.4**	**33.5**	**31.3**	**26.3**	**22.7**	**35.3**	**33.5**	**34.4**	**25.6**	**38.7**	**27.1**	17.7	**22.9**	**17.3**	**27.9**	**22.3**	**28.8**
Prompt-Based Baselines																							
Llama 2 7B	0.0	0.0	0.0	**32.0**	0.0	23.8	0.0	0.0	0.0	7.1	0.0	7.7	0.0	2.7	10.5	0.0	0.0	0.0	0.0	0.0	0.0	1.1	0.0
Llama 2 13B	0.0	0.0	0.0	25.0	0.0	27.0	0.0	0.0	0.0	8.0	0.0	4.0	0.0	2.0	11.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
Llama 2 70B	0.0	0.0	**35.0**	0.0	0.0	0.0	0.0	0.0	0.0	3.0	0.0	7.0	0.0	12.0	3.0	0.0	0.0	**36.0**	0.0	0.0	0.0	8.0	0.0
Google Models																							
PaLM 2	0.2	-	5.5	-	-	-	0.0	-	0.0	-	-	-	2.5	0.0	-	0.0	2.12	-	-	-	-	-	-
gemini-pro	0.1	-	5.4	-	0.0	-	0.0	-	0.0	-	-	-	2.4	0.0	-	2.1	7.5	-	-	-	-	-	-
Gemma 7B Instruct	0.2	0.0	0.8	0.0	-	0.0	-	0.0	-	0.0	-	0.0	0.5	0.0	0.0	-	-	0.0	-	0.0	0.0	0.0	-
Open AI Models																							
`gpt-3.5-turbo`	25.9	9.4	24.8	15.8	17.1	21.9	20.6	23.0	25.8	23.2	0.0	23.7	23.6	26.8	19.0	26.4	14.7	12.1	13.8	7.7	15.7	14.4	22.9
`gpt-4-32k`	27.5	16.9	23.1	15.6	18.0	25.5	19.7	23.7	23.3	24.6	16.5	24.5	24.0	25.5	20.0	29.1	15.3	13.4	14.4	6.3	18.5	14.7	22.4

	pidgin	ps	pt	rn	ru	si	so	sr*	sr**	sw	ta	te	th	ti	tr	uk	ur	uz	vi	yo	zh-Hant	zh-Hans	**avg**
Fine-tuned Baselines																							
mT5-RL	**34.7**	**35.7**	**32.3**	**31.6**	0.0	**21.3**	**27.6**	**23.8**	**21.6**	**35.8**	0.0	**19.3**	**13.8**	**27.0**	**31.8**	**24.9**	**35.9**	**19.0**	28.8	**26.2**	**39.8**	**39.7**	**26.9**
Prompt-Based Baselines																							
Llama 2 7B	5.7	0.0	7.4	0.0	0.1	0.0	3.6	0.1	4.0	4.1	0.0	0.0	0.0	0.0	4.9	0.0	0.0	0.0	**34.1**	8.1	0.0	0.0	3.5
Llama 2 13B	6.0	0.0	7.0	4.0	0.0	0.0	3.0	0.0	4.0	3.7	0.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0	30.0	8.0	0.0	0.0	3.2
Llama 2 70B	13.0	0.0	5.0	2.0	0.0	0.0	9.0	0.0	3.0	8.5	0.0	0.0	0.0	0.0	6.0	0.0	0.0	0.0	3.0	0.0	0.0	20.0	3.9
Google Models																							
PaLM 2	-	-	0.0	-	**4.7**	-	-	-	-	0	-	-	0.0	-	-	-	-	-	0.0	-	-	-	1.2
gemini-pro	-	-	0.0	-	4.5	-	-	-	-	0.1	-	-	2.8	-	0.0	4.9	-	-	0.0	-	2.8	2.6	2.0
Gemma 7B Instruct	0.0	0.0	0.0	0.0	1.7	0.0	0.0	0.1	0.0	0.0	0.7	-	1.0	0.0	0.0	1.7	0.5	0.0	0.0	0.0	-	1.3	0.2
Open AI Models																							
`gpt-3.5-turbo`	23.6	21.3	24.9	20.5	0.0	9.2	21.2	11.1	16.7	27.3	0.0	0.6	8.9	0.0	22.0	17.9	25.6	11.8	20.9	16.5	23.4	21.5	17.2
`gpt-4-32k`	27.1	23.9	24.0	21.9	0.0	15.4	17.4	18.1	16.3	8.5	0.0	14.9	9.7	18.8	17.8	18.2	25.5	12.3	21.6	16.3	27.2	30.7	18.8

Table 15: Comparing performance of different models on all languages in XLSum. Metric: Rouge - L.
sr*: `serbian_cyrillic`, sr**: `serbian_latin`. Unsupported languages are marked with '-'. Averages are calculated only from supported languages.

Model	id	tr	zh	sw	ta	**avg**
LLaVA Models						
bakllava-v1	0.25	0.27	0.36	0.38	0.22	0.30
bakllava-v1 (TT)	0.58	0.58	0.56	0.54	0.52	0.56
llava-v1.5-13B	0.53	0.52	0.53	0.51	0.50	0.52
llava-v1.5-13B (TT)	0.53	0.57	0.52	0.53	0.54	0.54
vip-llava-13B	0.52	0.50	0.20	0.50	0.14	0.37
vip-llava-13B (TT)	0.56	0.56	0.53	0.52	0.51	0.54
Google Models						
`gemini-pro-vision`	0.55	0.60	0.59	0.55	0.52	0.56
`gemini-pro-vision (TT)`	0.61	0.63	0.57	0.58	0.63	0.60
OpenAI Models						
`gpt-4-vision`	0.81	0.81	0.71	0.74	0.78	0.77
`gpt-4-vision (TT)`	0.74	0.82	0.68	0.68	0.69	0.72

Table 16: Comparing performance of Multimodal models on MaRVL. Metric: Accuracy

Model	ar	cs	da	de	en	es	fi	fr	he	hu	it	ja
Prompt-Based Baselines												
Llama 2 70B	42.3	65.0	66.2	69.4	78.8	68.4	62.7	72.2	41.4	61.1	68.4	56.6
Mistral 7B Instruct	1	54.2	51.3	51.5	65.6	59.4	35.2	60	0	42	51.4	41.7
Google Models												
PaLM 2	86.9	87.8	88.1	87.6	**92.2**	86.0	86.6	88.2	0.0	87.0	86.3	84.1
gemini-pro	89.0	**90.4**	**89.8**	**88.8**	90.8	**89.3**	**89.3**	89.8	**88.0**	**88.2**	**89.3**	**86.9**
Gemma 2B Instruct	36.1	35.0	34.0	34.7	42.2	39.2	34.8	40.0	31.7	29.1	37.3	35.8
Gemma 7B Instruct	45.7	53.9	50.9	55.3	66.3	57.8	57.8	56.7	44.7	46.0	56.3	47.4
Open AI Models												
`gpt-3.5-turbo`	69.3	76.9	80.7	83.3	87.7	79.2	77.9	83.1	64.2	74.6	80.0	70.9
`gpt-4`	**91.8**	85.5	82.3	85.0	79.0	83.0	86.8	**90.9**	86.9	61.0	81.0	86.0
	ko	nl	no	pl	pt	ru	sv	th	tr	zh-Hans	zh-Hant	**avg**
Prompt-Based Baselines												
Llama 2 70B	56.3	66.2	65.7	61.7	70.2	67.0	67.4	38.9	47.3	62.4	59.3	61.5
Mistral 7B Instruct	41.4	54.5	51.4	42.8	54.4	62.1	54.9	0	31.4	60.4	56.4	43.2
Google Models												
PaLM 2	86.0	87.3	88.2	87.7	88.4	87.4	87.1	80.9	**85.0**	86.8	86.9	83.2
gemini-pro	87.9	**88.9**	**88.8**	**88.1**	**89.4**	89.1	**88.9**	83.4	84.1	**91.4**	**90.3**	**88.7**
Gemma 2B Instruct	36.4	35.9	34.0	35.0	27.7	40.0	36.2	36.7	32.0	41.2	39.2	35.8
Gemma 7B Instruct	46	49.1	53.4	51.1	58.4	56.2	52.6	47.6	41.4	57.7	57.1	52.6
Open AI Models												
`gpt-3.5-turbo`	67.1	80.4	79.0	74.7	83.0	78.4	81.7	55.7	69.9	77.6	76.3	76.2
`gpt-4`	**88.6**	85.3	83.1	85.2	78.8	**90.6**	80.0	**83.9**	78.4	89.0	89.8	84.0

Table 17: Comparing performance of different models on all languages in BeleBele. Metric: Accuracy.

Model	ar	cs	da	de	en	es	fi	fr	it	ja
LLaVA Models										
bakllava-v1	0.12	11.64	15.39	14.97	29.81	19.00	12.80	18.73	17.23	0.10
llava-v1.5-13B	16.13	19.37	27.04	28.34	30.06	30.22	20.99	32.10	28.10	10.40
vip-llava-13B	12.24	18.41	25.26	28.47	30.63	30.32	19.26	31.86	27.79	9.59
Google Models										
`gemini-pro-vision`	25.00	25.42	32.31	31.14	32.39	32.74	27.44	33.95	32.03	13.37
OpenAI Models										
`gpt-4-vision`	28.47	31.67	35.11	34.96	37.67	38.32	31.96	39.59	35.88	15.81
	ko	nl	no	pl	pt	ru	sv	th	tr	zh
LLaVA Models										
bakllava-v1	0.11	16.57	15.21	12.58	16.44	0.57	15.25	0.49	12.79	0.13
llava-v1.5-13B	6.82	31.56	26.28	21.53	28.20	19.51	27.26	14.53	16.74	8.57
vip-llava-13B	6.60	30.08	25.12	17.19	19.40	12.44	25.49	9.87	15.47	7.67
Google Models										
`gemini-pro-vision`	10.91	33.88	30.26	28.29	32.70	24.93	30.53	26.34	27.79	8.76
OpenAI Models										
`gpt-4-vision`	15.44	39.29	34.19	32.82	36.26	30.76	35.00	32.37	32.00	10.15

Table 18: Comparing performance of Multimodal models on XM3600. Metric: chrF

Model	es	fr	it	pt	ru	tr	**avg**
Prompt-Based Baselines							
PaLM (0-Shot)	79.83	78.99	-	77.58	80.35	84.1	80.17
PaLM (10-Shot Monolingual)	**91.23**	86.16	-	90.99	92.47	84.5	89.07
PaLM-2 (0-Shot)	88.6	84.11	-	87.68	90.5	93.42	88.86
PaLM-2 (10-Shot Monolingual)	89.68	**87.94**	-	**92.05**	**94.25**	**94.34**	**91.65**
OpenAI Models							
gpt-3.5-turbo (Crosslingual)	77.27	73.64	80.05	81.16	74.99	85.65	78.79
gpt-3.5-turbo (TT)	74.20	70.09	76.67	72.66	73.68	82.99	75.05
text-davinci-003 (Crosslingual)	79	74.55	**81.11**	81.63	79.13	93.55	81.50
text-davinci-003 (TT)	79.06	72.93	78.93	75.18	80.48	93.22	79.97

Table 19: Comparing performance of different models on all languages in Jigsaw. Metric: Accuracy. PaLM 2 does not support all languages; unsupported languages are marked with '-'. Averages for PaLM 2 are calculated only from supported languages.

	Google			Microsoft			Amazon			Systran			GPT Turbo 3.5			Bloomz		
	Acc	Δ_G	Δ_S	Acc	Δ_G	Δ_S	Acc	Δ_G	Δ_S	Acc	Δ_G	Δ_S	Acc	Δ_G	Δ_S	Acc	Δ_G	Δ_S
es	50.9	23.2	20.9	45	36.5	22.9	**57.2**	15.3	21.7	42.5	46.2	15.6	54.9	22.7	26.2	55.6	17.2	32.5
fr	**61.6**	6.1	22.3	44.5	34.2	15.8	54.2	16.4	15	43.4	41.8	-0.1	52.7	21.4	26.1	52	17.8	24.6
it	38.6	32.9	18.6	38.8	41.8	10.5	40.2	26.8	14.7	38.1	47.3	6.3	45.1	21.9	26.7	**45.7**	9	18.5
ru	37.8	36.7	11.4	36.9	42	8.4	39.8	34.8	9.4	37.3	44.1	9.2	**41**	31.6	10.2	5.9	INV	0
uk	38.4	43.5	10.7	41.3	46.8	11.9	-	-	-	28.9	22.4	12.9	**42.9**	34.2	12.1	16.8	22.7	2.2
he	50.8	11.7	35.5	44	22	29.8	48	13.6	45.9	43.1	26.9	23.1	**57.5**	7.6	40.8	27.5	31.4	5
ar	45.8	42.5	16.2	45	47.1	14.2	48.3	37.8	18.8	45.6	49.4	-4.1	**61.1**	13.9	27.9	48.1	23	25.6
de	59.4	12.5	12.6	**74.1**	0	8.8	62.4	12	16.7	48.5	34.5	10	57.5	19.5	14.2	47.6	56.2	6.6

Table 20: Performance of commercial MT systems and LLMs on the WinoMT corpus on 8 target languages. Results are categorized by language family. Acc indicates overall gender accuracy (% of instances the translation had the correct gender), Δ_G denotes the difference in performance (F1 score) between masculine and feminine scores, and Δ_S is the difference in performance (F1 score) between pro-stereotypical and anti-stereotypical gender role assignments (higher numbers in the two latter metrics indicate stronger biases). Numbers in bold indicate best accuracy for the language across all systems. Notes: [1. For Google, Microsoft, Amazon, and Systran we use the translations provided by (Stanovsky et al., 2019). Some values differ from the original paper due to updated Spcay modules. 2. For Ru in Bloomz, Precision in male predictions is 0 leading to Invalid (INV) in Δ_G]

Model	as	bn	gu	hi	ka	kn	ml	mr	np	or	pa	ta	te	ur
SOTA model														
IT2	46.8	49.7	53.1	49.6	35.6	33.8	45.7	48.6	51.5	40.2	57.8	39.1	45.5	61.6
Prompt-Based Baselines														
Llama 2 70B	0.7	0.7	0.9	1.3	1	1	1.3	1.5	0.8	0.7	0.7	1.1	1.1	1.9
Mistral 7B Instruct	2.1	5.5	2.2	7.4	3.8	4	2.5	5	5.4	1.6	1.8	3.8	2.9	6
Gemma 7B Instruct	13.9	21.9	4.7	26.9	10.3	2.2	6.3	15.7	16.5	1.6	4.2	18.6	10.5	1.6
OpenAI Models														
`gpt-3.5-turbo (0-shot)`	27.2	39.9	36	46	-	27.9	30.4	34	38.3	25.6	40.6	29.7	32.1	49
`gpt-3.5-turbo`	26.3	36.1	34.3	41.8	12.7	26.5	28.1	32.8	37.5	23.5	40.2	28.3	30.1	47.1
`gpt-4-32k`	34.9	44.3	42.6	47	17.7	30.5	36.8	39.1	43.9	33.4	48.4	33.8	38.5	53.4

Table 21: Performance of various models on IN22-Conv in the En-Indic direction. All the LLMs are prompted with 8 few-shot examples. However, we also report the 0-shot performance from Gala et al. (2023)

Model	as	bn	gu	hi	ka	kn	ml	mr	np	or	pa	ta	te	ur
SOTA model														
IT2	62.9	58.4	62	60.1	52.6	47.5	54.3	58.5	63	60.3	62.7	45.8	52.9	65.5
Prompt-Based Baselines														
Llama 2 70B	16.2	19.8	14.9	30.3	14.7	13.3	14.7	18.9	22.4	13.5	14.1	14.2	13.7	25
Mistral 7B Instruct	17	22.5	11.7	27.4	11.9	12.7	10.2	16.7	19	9.7	10.4	11.7	11.3	21.5
Gemma 7B Instruct	18.8	27.4	23.3	38	19	14.8	22	25.1	26.3	12.1	22	21.7	24.5	33.8
OpenAI Models														
`gpt-3.5-turbo (0-shot)`	43.6	52.9	50.9	57	-	42.1	44	47.6	52	45.2	53.3	38	42.4	57.1
`gpt-3.5-turbo`	28.5	34.7	31.2	43.5	25.5	37.4	40.5	47.4	49.4	39.5	49.4	34.7	38.1	53.7
`gpt-4-32k`	53.3	57.7	57.3	59.5	34	47.9	49.9	55.6	59.6	55	60.1	44.4	49.2	63.2

Table 22: Performance of various models on IN22-Conv in the Indic-En direction. All the LLMs are prompted with 8 few-shot examples. However, we also report the 0-shot performance from Gala et al. (2023)

Model	as	bn	gu	hi	ka	kn	ml	mr	np	or	pa	ta	te	ur
SOTA model														
IT2	47.1	51.8	53.5	56.7	40.2	51	50.9	51	49	43.9	50.6	49.5	52.4	68.2
Prompt-Based Baselines														
Llama 2 70B	0.5	0.7	0.6	1.2	1	0.7	0.9	1	0.9	0.5	0.8	0.9	0.9	1.3
Mistral 7B Instruct	1.1	2.5	1.1	3.9	2.7	2.1	1.3	2.9	3.1	0.8	1	2.1	1.6	3.7
Gemma 7B Instruct	16.1	24	3.5	32.1	14.5	2.8	6.5	19.5	20.4	1.9	3.4	23.8	13.1	2.1
OpenAI Models														
`gpt-3.5-turbo (0-shot)`	25.9	39.9	35.6	47.1	-	34.5	31.6	33.9	37.2	27.8	36.2	34	34.3	47.6
`gpt-3.5-turbo`	28.9	39.4	35.7	47.5	21.3	35	31.4	34.9	38.9	28.1	36.2	33.9	34	47.8
`gpt-4-32k`	33.7	44.2	40.3	50.3	23.7	40.6	38.5	38.7	42.4	34.9	41.5	40.3	41.4	53.4

Table 23: Performance of various models on IN22-Gen in the En-Indic direction. All the LLMs are prompted with 8 few-shot examples. However, we also report the 0-shot performance from Gala et al. (2023)

Model	as	bn	gu	hi	ka	kn	ml	mr	np	or	pa	ta	te	ur
SOTA model														
IT2	65.8	63.2	66.5	65.4	60.4	64.2	64.5	63.7	67.7	66.2	63.4	59.8	64.8	73
Prompt-Based Baselines														
Llama 2 70B	13.1	15.8	12.4	23.6	12.8	11.3	12.2	16.7	18.8	11.7	11.7	12	11.4	19.4
Mistral 7B Instruct	13.7	17	10.8	20.4	12.9	12.3	9.9	15.3	16.9	9.8	9.9	11.8	11.4	17.4
Gemma 7B Instruct	21.2	29.2	24.7	42.1	23.9	17.9	23.9	29.4	31.1	14.5	24.2	24.9	28.8	36.8
OpenAI Models														
`gpt-3.5-turbo (0-shot)`	46.9	52.1	51.7	57.7	-	51.7	47.8	50.3	54.2	48	51.7	41.3	46.5	58.8
`gpt-3.5-turbo`	45.8	51.2	50.8	57.8	36	51.1	45.8	50.3	54.2	47.4	51.1	40.2	45.6	59.6
`gpt-4-32k`	57.7	59.9	60	63	46.1	59.4	57.9	59.3	65.7	59.4	60.8	53.9	56.7	68.5

Table 24: Performance of various models on IN22-Gen in the Indic-En direction. All the LLMs are prompted with 8 few-shot examples. However, we also report the 0-shot performance from Gala et al. (2023)

Model	en	en-hi	fr	hi	ko	zh	**avg**
Fine-tuned Baselines							
mBART	**84.6**	**60.7**	**73.1**	**75.3**	**71.2**	**91.7**	**76.1**
Prompt-Based Baselines							
Llama 2 70B	59.0	44.0	51.8	38.4	51.1	73.0	52.9
Mistral 7B	49.8	38.1	44.3	35.5	41.4	51.8	43.5
Mistral 7B Instruct	48.2	38.1	42.4	31.9	41.4	50.8	42.1
Google Models							
PaLM 2	62.2	45.3	51.5	52.1	55.4	73.9	56.7
`gemini-pro`	64.4	49.3	52.2	53.0	53.2	73.6	57.6
Gemma 2B Instruct	49.3	33.5	52.2	38.7	51.4	58.4	47.2
Gemma 7B Instruct	50.3	33.7	52.0	39.5	52.1	58.0	47.8
Open AI Models							
`gpt-3.5-turbo`	71.0	50.2	60.3	57.3	59.9	81.4	63.4
`gpt-4-32k`	75.6	57.7	69.0	63.2	69.1	85.3	70.0

Table 25: Comparing performance of different models on all languages in X-RiSAWOZ. Metric: Dialogue Action Accuracy. The fine-tuned baseline is taken from Moradshahi et al. (2023).

Metric	Model	Languages						**avg**
		en	en-hi	fr	hi	ko	zh	
BLEU (↑)	Llama 2	12.8	6.6	13.6	3.4	2.1	1.7	6.7
	Mistral 7B	24.9	7.7	22.3	4.7	9.0	12.3	13.5
	Mistral 7B Instruct	30.7	5.5	23.8	3.4	9.6	15.1	14.7
	PaLM 2	35.9	21.9	**31.9**	**26.4**	21.7	24.3	27.0
	gemini-pro	**36.3**	**25.7**	33.3	16.7	23.8	**26.8**	**27.1**
	Gemma 2B Instruct	26.0	21.2	13.7	6.4	25.9	24.3	19.6
	Gemma 7B Instruct	30.7	21.6	17.6	10.6	**26.7**	24.4	21.9
	gpt-3.5-turbo	10.8	8.9	5.0	6.5	4.6	1.5	6.2
	gpt-4-32k	33.1	18.2	31.5	25.2	24.0	23.4	25.9
Slot Error Rate (↓)	Llama 2	11.1	39.1	17.9	46.6	43.7	31.3	31.6
	Mistral 7B	20.2	48.9	30.6	55.7	38.1	23.5	36.2
	Mistral 7B Instruct	18.6	53.4	31.3	61.9	42.7	22.8	38.4
	PaLM 2	10.4	31.3	18.9	35.9	22.8	7.2	21.1
	gemini-pro	10.8	35.2	15.7	13.3	13.7	8.3	16.2
	Gemma 2B Instruct	21.6	38.7	17.5	27.6	23.8	15.7	24.1
	Gemma 7B Instruct	15.7	36.3	15.2	27.6	21.3	13.3	21.6
	gpt-3.5-turbo	**5.2**	**21.2**	10.4	**26.7**	19.9	3.9	14.6
	gpt-4-32k	5.9	21.2	**8.8**	28.3	**18.2**	**3.3**	**14.3**
Success Rate (↑)	Llama 2	58.3	3.3	46.7	6.7	10.0	25.0	25.0
	Mistral 7B	25.0	1.7	11.7	1.7	3.3	20.0	10.6
	Mistral 7B Instruct	28.3	1.7	10.0	1.7	5.0	20.0	11.1
	PaLM 2	55.0	10.0	38.3	16.7	26.7	66.7	35.6
	gemini-pro	57.3	13.7	38.2	**55.4**	36.2	66.7	44.6
	Gemma 2B Instruct	22.7	1.7	31.3	1.7	35.0	56.3	24.8
	Gemma 7B Instruct	28.3	3.3	30.7	5.3	35.5	57.0	26.7
	gpt-3.5-turbo	**71.7**	**26.7**	63.3	10.0	33.3	**90.0**	49.2
	gpt-4-32k	66.7	25.0	**65.0**	36.7	**38.3**	83.3	**52.5**
API Accuracy (↑)	Llama 2	64.5	51.6	46.8	40.3	58.1	70.5	55.3
	Mistral 7B	3.2	3.2	3.2	1.6	1.6	1.6	2.4
	Mistral 7B Instruct	3.2	3.2	3.2	3.2	4.7	4.7	3.7
	PaLM 2	85.5	75.8	64.5	61.3	62.9	**98.4**	74.7
	gemini-pro	85.5	56.0	64.4	**85.2**	62.2	95.1	74.7
	Gemma 2B Instruct	3.2	1.6	4.7	1.6	4.7	22.3	6.4
	Gemma 7B Instruct	4.7	3.2	4.7	3.2	4.7	22.3	7.1
	gpt-3.5-turbo	90.3	74.2	64.5	35.7	82.3	88.5	72.6
	gpt-4-32k	**98.4**	**91.9**	**69.4**	80.7	**93.5**	98.4	**88.7**
Dialogue Action Accuracy (↑)	Llama 2	59.0	44.0	51.8	38.4	51.1	73.0	52.9
	Mistral 7B	49.8	38.1	44.3	35.5	41.4	51.8	43.5
	Mistral 7B Instruct	48.2	38.1	42.4	31.9	41.4	50.8	42.1
	PaLM 2	62.2	45.3	51.5	52.1	55.4	73.9	56.7
	gemini-pro	64.4	49.3	52.2	53.0	53.2	73.6	57.6
	Gemma 2B Instruct	49.3	33.5	52.2	38.7	51.4	58.4	47.2
	Gemma 7B Instruct	50.3	33.7	52.0	39.5	52.1	59.0	47.8
	gpt-3.5-turbo	71.0	50.2	60.3	57.3	59.9	81.4	63.4
	gpt-4-32k	**75.6**	**57.7**	**69.1**	**63.2**	**69.1**	**85.3**	**70.0**
Joint Goal Accuracy (↑)	Llama 2	64.8	45.3	49.5	38.8	39.4	71.7	51.6
	Mistral 7B	16.0	16.3	9.1	0.7	4.2	14.3	10.1
	Mistral 7B Instruct	12.7	15.6	8.8	0.7	3.6	13.7	9.2
	PaLM 2	65.8	49.5	52.4	48.9	53.4	77.5	57.9
	gemini-pro	67.9	55.7	52.7	53.6	54.0	79.3	60.5
	Gemma 2B Instruct	12.9	0.6	9.0	0.6	6.6	17.4	7.9
	Gemma 7B Instruct	11.2	0.7	10.9	1.7	7.3	17.6	8.4
	gpt-3.5-turbo	**77.2**	58.0	**58.3**	58.6	**56.7**	80.8	64.9
	gpt-4-32k	75.2	**62.5**	58.3	**63.5**	51.6	**87.6**	**66.5**

Table 26: Comparison of various task-specific metrics on X-RiSAWOZ. (↑) indicates metric is higher the better and (↓) the indicates lower the better. Best results for a particular model-language combination are bolded.

Model	ar	en	fi	id	ja	ko	ru	sw	te	th
GPT-4	-0.25	0.73	0.45	0.36	0.36	0.40	0.53	0.40	0.41	0.46
PaLM-2	0.55	0.64	0.07	0.16	0.72	0.60	0.61	0.23	NA	0.17

Table 27: Contamination values for the TydiQA dataset.

Model	de	en	es	fr	ja	ko	zh
GPT-4	0.77	0.72	0.66	0.71	0.55	0.44	0.65
PaLM-2	0.23	0.63	0.16	0.23	0.53	0.57	0.32

Table 28: Contamination values for the PAWS-X dataset.

Model	ar	bg	el	en	et	eu	fi	fr	hi	hu	it	ja	lt
GPT-4	0.41	0.05	0.17	0.21	0.16	0.13	0.04	0.09	0.32	0.2	0.03	0.23	0.08
PaLM-2	-0.05	-0.2	-0.12	0.04	-0.2	NA	-0.13	0.12	-0.17	-0.16	-0.03	-0.08	-0.13

Table 29: Contamination values for the UDPOS dataset (part-1)

Model	mr	pl	pt	ro	ru	ta	te	tr	uk	vi	wo	zh
GPT-4	0.32	-0.01	0.03	-0.25	0.15	0.11	-0.09	0.17	0.05	-0.01	-0.12	0.07
PaLM-2	NA	-0.21	-0.09	-0.13	-0.07	NA	NA	0.12	-0.23	-0.2	NA	-0.13

Table 30: Contamination values for the UDPOS dataset (part-2)

Model	et	ht	id	it	sw	ta	th	tr	vi	zh
GPT-4	NA	0.79	0.55	0.72	0.64	0.73	0.67	0.67	0.37	0.41
PaLM-2	-0.05	NA	-0.17	-0.07	-0.07	NA	-0.2	-0.24	-0.09	-0.13

Table 31: Contamination values for the X-COPA dataset.

Model	ar	bg	de	el	en	es	fr	hi	ru	sw	th	tr	ur	vi	zh
GPT-4	0.47	0.4	0.23	0.51	0.45	0.36	0.37	0.59	0.32	0.24	0.2	0.37	0.45	0.25	0.08
PaLM-2	0.47	0.31	0.17	0.47	0.37	0.13	0.27	0.33	0.25	0.15	0.09	0.31	NA	0.61	0.36

Table 32: Contamination values for the XNLI dataset.

Dataset	Gemma 7B Instruct	Llama 2 7B Instruct	Mistral 7B Instruct
PAWS-X	**0.0**	**0.0**	**0.0**
XCOPA	**0.0007**	**0.0**	**0.0**
XNLI	0.4162	0.0374	0.1148
XQUAD	0.0164	**0.0**	**0.0**
XRiSAWOZ	**0.0**	**0.0**	**0.0**
XstoryCloze	0.2917	0.0274	0.2743

Table 33: The statistical test was performed on a total of 5000 test points equally divided amongst all the languages of a given dataset. Our significance value is 0.001 which is calculated using $1/(1+r)$, where r is the number of permutations per shard (for us it is 700). If a value is less than 0.001, then that test set is contaminated for the given model. The it suffix for the above model stands for Instruction-Tuned variant of that said model.

Language	Language Family	Language Script	ISO code
Afrikaans	IE: Germanic	Latin	af
Amharic	Afro-Asiatic	Ge'ez (Ethiopic)	am
Arabic	Afro-Asiatic	Arabic	ar
Assamese	IE: Iranian	Brahmic	as
Azerbaijani	Turkic	Latin	az
Basque	Basque	Latin	eu
Bengali	IE: Iranian	Brahmic	bn
Bulgarian	IE: Balto-Slavic	Cyrillic	bg
Burmese	Sino-Tibetan	Brahmic	my
Mandarin	Sino-Tibetan	Chinese ideograms	zh
Dutch	IE: Germanic	Latin	nl
English	IE: Germanic	Latin	en
Czech	IE: Balto-Slavic	Latin	cs
Estonian	Uralic	Latin	et
Finnish	Uralic	Latin	fi
French	IE: Romance	Latin	fr
Georgian	Kartvelian	Georgian	ka
German	IE: Germanic	Latin	de
Greek	IE: Greek	Greek	el
Gujarati	IE: Iranian	Brahmic	gu
Hausa	Afro-Asiatic	Brahmic	ha
Hebrew	Afro-Asiatic	Hebrew	he
Hindi	IE: Iranian	Devanagari	hi
Urdu	IE: Iranian	Arabic	ur
Hungarian	Uralic	Latin	hu
Igbo	Niger-Congo	Latin	ig
Indonesian	Austronesian	Latin	id
Italian	IE: Romance	Latin	it
Japanese	Japonic	Japanese ideograms	ja
Javanese	Austronesian	Brahmic	jv
Kannada	Dravidian	Brahmic	kn
Kazakh	Turkic	Cyrillic	kk
Kirundi	Niger-Congo	Latin	rn
Korean	Koreanic	Hangul	ko
Kyrgyz	Turkic	Cyrillic	ky
Malay	Austronesian	Latin	ms
Malayalam	Dravidian	Brahmic	ml
Marathi	IE: Iranian	Devanagari	mr
Nepali	IE: Iranian	Devanagari	ne
Odia	IE: Iranian	Brahmic	or
Oromo	Afro-Asiatic	Latin	om
Pashto	IE: Iranian	Arabic	ps
Persian	IE: Iranian	Arabic	fa
Pidgin	IE: Germanic	Latin	pid
Portuguese	IE: Romance	Latin	pt
Punjabi	IE: Iranian	Gurmukhi	pa
Russian	IE: Balto-Slavic	Cyrillic	ru
Scottish_gaelic	IE: Celtic	Latin	gd
Serbian_Cyrillic	IE: Balto-Slavic	Cyrillic	sr
Serbian_Latin	IE: Balto-Slavic	Latin	sr
Sinhala	IE: Iranian	Brahmic	si
Somali	Afro-Asiatic	Latin	so
Spanish	IE: Romance	Latin	es
Swahili	Bantu	Latin	sw
Tagalog	Austronesian	Brahmic	tl
Tamil	Dravidian	Brahmic	ta
Telugu	Dravidian	Brahmic	te
Thai	Kra-Dai	Brahmic	th
Tigrinya	Afro-Asiatic	Ge'ez (Ethiopic)	ti
Turkish	Turkic	Latin	tr
Ukrainian	IE: Balto-Slavic	Cyrillic	uk
Uzbek	Turkic	Latin	uz
Vietnamese	Austro-Asiatic	Latin	vi
Welsh	IE: Celtic	Latin	cy
Yoruba	Niger-Congo	Latin	yo
Bemba	Bantu	Latin	bem
Fon	Niger-Congo	Latin	fon
Kinyarwanda	Bantu	Latin	rw
Twi	Kwa	Latin	tw
Wolof	Niger-Congo	Latin	wo
Zulu	Bantu	Latin	zu
Czech	IE: Balto-Slavic	Latin	cs
Danish	IE: Germanic	Latin	da
Norwegian	IE: Germanic	Latin	no
Polish	IE: Balto-Slavic	Latin	pl
Swedish	IE: Germanic	Latin	sv
English-Hindi	IE: Germanic	Latin	en-hi
English-Spanish	IE: Romance	Latin	en-es
Kashmiri	Dardic	Arabic	ks
Lithuanian	IE: Balto-Slavic	Latin	lt
Quechua	Quechuan	Latin	qu
Romanian	IE: Romance	Latin	ro
Haitian Creole	French Creole	Latin	ht

Table 34: List of Languages and their corresponding Language Families, Language Scripts and ISO Codes benchmarked in