Aya 23: Open Weight Releases to Further Multilingual Progress

Viraat Aryabumi*[1], John Dang[1], Dwarak Talupuru[2], Saurabh Dash[1], David Cairuz[2], Hangyu Lin[2], Bharat Venkitesh[2], Madeline Smith[1], Kelly Marchisio[2], Sebastian Ruder[2], Acyr Locatelli[2], Julia Kreutzer[1], Phil Blunsom[2], Marzieh Fadaee[1], Ahmet Üstün*[1], and Sara Hooker*[1]

[1]Cohere For AI, [2]Cohere

Corresponding authors: Viraat Aryabumi `<viraat@cohere.com >`, Ahmet Üstün `<ahmet@cohere.com>`, Sara Hooker `<sarahooker@cohere.com>`

Abstract

This technical report introduces **Aya** 23, a family of multilingual language models. **Aya** 23 builds on the recent release of the **Aya** model [Üstün et al., 2024], focusing on pairing a highly performant pre-trained model with the recently released **Aya** collection [Singh et al., 2024]. The result is a powerful multilingual large language model serving 23 languages, **expanding state-of-art language modeling capabilities to approximately half of the world's population.** The **Aya** model covered 101 languages whereas 23 is an experiment in depth vs breadth, exploring the impact of allocating more capacity to fewer languages that are included during pre-training. **Aya** 23 outperforms both previous massively multilingual models like **Aya** 101 for the languages it covers, as well as widely used models like Gemma, Mistral and Mixtral on an extensive range of discriminative and generative tasks. We release the open weights for both the 8B and 35B models as part of our continued commitment for expanding access to multilingual progress.

Aya-23-8B: `https://huggingface.co/CohereForAI/aya-23-8B`
Aya-23-35B: `https://huggingface.co/CohereForAI/aya-23-35B`

1 Introduction

In this work we introduce **Aya** 23, a family of multilingual instruction-tuned language models supporting 23 languages based on Cohere's Command model[1] and the **Aya** multilingual instruction-style collection. To date, the majority of progress in large language modeling has been English-centric, leading to models which perform poorly outside of a handful of languages. This can result in cliffs in model performance in languages not included in pre-training [Schwartz et al., 2022; Kotek et al., 2023; Khandelwal et al., 2023; Vashishtha et al., 2023; Khondaker et al., 2023], the

*First authors.
[1]https://cohere.com/command

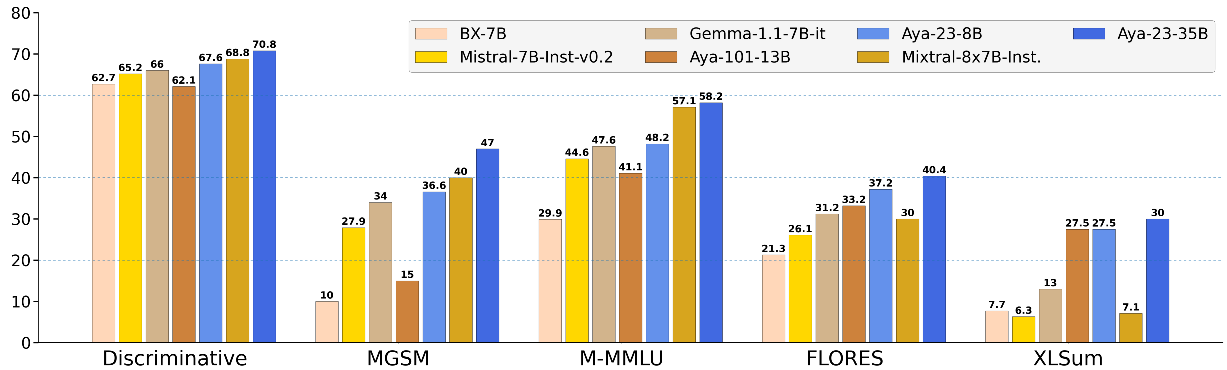


Figure 1: **Multilingual benchmark results** covering 5 task categories from 8 datasets for **Aya** 23 models against massively multilingual **Aya**-101-13B and widely used open weight models of similar size such as Bacterian-X-7B, Gemma-1.1-7B-it, Mistral-7B-Inst-v0.2 and Mixtral-8x7B-Inst.

introduction of security flaws for all users, [Yong et al., 2023a; Nasr et al., 2023; Li et al., 2023b; Lukas et al., 2023; Deng et al., 2023] and a growing divide in the cost of technology due to high latencies for generations outside of English [Held et al., 2023; Durmus et al., 2023; Nicholas & Bhatia, 2023; Ojo et al., 2023; Ahia et al., 2023].

Multilingual efforts including the release of Aya 101 [Üstün et al., 2024], BLOOMZ [Muennighoff et al., 2023] and mT0 [Muennighoff et al., 2023] models have made great strides in expanding access to modern natural language processing technologies for the world. However, there still remains significant room for improvement relative to first-class citizen languages like English and Chinese. Two major hurdles in the development of powerful multilingual models are (1) the lack of robust multilingual pretrained models, and (2) the scarcity of instruction-style training data covering a diverse set of languages.

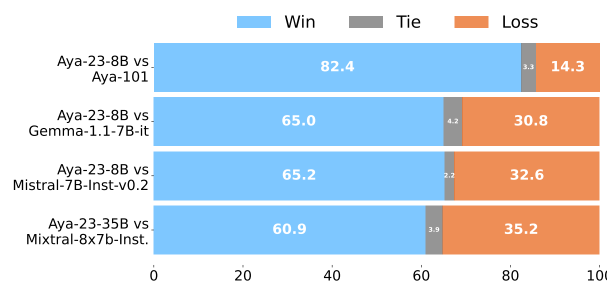


Figure 2: **Average win-rates (%)** across 10 languages for **Aya** 23 models against widely used open weight models of similar size.

The **Aya** initiative[2] was created to address the aforementioned data scarcity issues by creating and releasing the largest multilingual instruction-style dataset [Singh et al., 2024] to date, along with the **Aya** 101 model [Üstün et al., 2024]. **Aya** 101 was a step forward in massively multilingual language modeling, creating a 101 languages state-of-the-art instruction fine-tuned LLM. However, **Aya** 101 was by necessity built upon the mT5 [Xue et al., 2020] pre-trained base model given it was one of the few pre-trained models that had been trained on 101 languages. mT5 is relatively outdated given the rapid advances in LLM technology since its release in 2019. Its major limitations are: **1) Outdated knowledge:** Having been pre-trained several years ago, mT5 is not as useful for interactions about events that occurred recently. **2) Inadequate Performance:** There are many stronger models now compared to when mT5 was released, such as the Command R+[3], Command

[2] https://cohere.com/research/aya
[3] https://docs.cohere.com/docs/command-r-plus

R[4], Llama series [Touvron et al., 2023a;b], Mistral models [Jiang et al., 2023; 2024] and Gemma models [Gemma-Team, 2024].

Furthermore, **Aya** 101 was a 13-billion parameter model designed for ***breadth***, expanding coverage to nearly double that achieved by previous models with 101 languages. Due to the well-documented *curse of multilinguality* [Arivazhagan et al., 2019; Conneau et al., 2019; Pfeiffer et al., 2022], models attempting to serve such a broad variety of languages often lag in generative performance on any given language relative to models dedicated to serving a more focused subset, because of the need to share model capacity so widely. For **Aya** 23, we instead balance ***breadth*** and ***depth***, exploring the impact of allocating more capacity to fewer languages (23 languages) that are included during pre-training, alleviating the "curse" and leading to large gains over the original **Aya** 101 and widely used models such as Gemma [Gemma-Team, 2024], Mistral [Jiang et al., 2023], and Mixtral [Jiang et al., 2024] for the corresponding 23 languages.

In this technical report, we assess the performance of **Aya** 23 models following the comprehensive multilingual evaluation framework proposed by Üstün et al. [2024]. In our evaluation, we focus on 23 languages that are covered by the new **Aya** model family.

These 23 languages are: *Arabic, Chinese (simplified & traditional), Czech, Dutch, English, French, German, Greek, Hebrew, Hindi, Indonesian, Italian, Japanese, Korean, Persian, Polish, Portuguese, Romanian, Russian, Spanish, Turkish, Ukrainian* and *Vietnamese*. Our choice of languages was guided to align with the languages present in pre-training of Command R, due to known difficulties of introducing new languages after pre-training [Zhao et al., 2024; Yong et al., 2023b].

We release **Aya** 23 in two model sizes: 8-billion (8B) and 35-billion (35B) parameters. **Aya**-23-35B achieves the highest results across all the evaluation tasks and languages covered, while **Aya**-23-8B demonstrates *best-in-class* multilingual performance which is crucial given that model sizes above 13B parameters limit model usability on consumer-grade hardware. We note that relative to **Aya** 101, **Aya** 23 improves on discriminative tasks by up to 14%, generative tasks by up to 20%, and multilingual MMLU by up to 41.6%. Furthermore, **Aya** 23 achieves a 6.6x increase in multilingual mathematical reasoning compared to **Aya** 101. Across **Aya** 101, Mistral, and Gemma, we report a mix of human annotators and LLM-as-a-judge comparisons. Across all comparisons, the **Aya**-23-8B and **Aya**-23-35B are consistently preferred. By releasing the weights of the **Aya** 23 model family, we hope to empower researchers and practitioners to advance multilingual models and applications.

2 Pre-trained Models

The **Aya** 23 model family is based on the Cohere Command series models which are pre-trained using a data mixture that includes texts from 23 languages. In particular, **Aya**-23-35B is a further fine-tuned version of Cohere Command R. For pre-trained models, a standard decoder-only Transformer architecture is used with the following setup:

1. **Parallel Attention and FFN layers**: Similar to PALM-2 [Anil et al., 2023] we use a parallel block architecture that leads to a significant improvement in training efficiency without hurting model quality, especially in tensor-parallel (TP) settings.

[4] https://docs.cohere.com/docs/command-r

	Embedding dims	Num layers	FFN hidden dims	Num heads	Num KV heads	Head size	Vocab size	Embedding parameters	Non-embedding parameters
Aya-23-8B	4096	32	22528	32	8	128	256000	1,048,576,000	6,979,457,024
Aya-23-35B	8012	40	45056	64	64	128	256000	2,097,152,000	32,883,679,232

Table 1: Architecture parameters for **Aya** 23 model family

2. **SwiGLU Activation:** We found SwiGLU [Shazeer, 2020] to have higher downstream performance than other activations. We scale the dimensions of FFN layers to retain approximately the same number of trainable parameters compared to non-SwiGLU activation functions.

3. **No bias:** Similar to PALM2 [Anil et al., 2023], we remove all biases from dense layers to improve the training stability.

4. **RoPE:** We use rotary positional embeddings [Su et al., 2021] to provide better long context extrapolation. Furthermore, it also achieves better downstream task performance for short context lengths compared to other relative positional encoding methods such as ALiBi [Press et al., 2021].

5. **Tokenizer:** We use a BPE tokenizer of size 256k. We perform NFC normalization and digits are split into individual tokens. The tokenizer is trained on a subset of our pre-training datasets balanced to ensure efficient representations across languages.

6. **Grouped Query Attention (GQA): Aya**-23-8B uses grouped-query attention [Joshua Ainslie, 2023] where each KV head shares multiple Q heads to reduce inference-time memory footprint.

All base models are trained using Fax [Yoo et al., 2022], a Jax-based distributed training framework on TPU v4 chips [Jouppi et al., 2023]. A combination of parallelism strategies is used to ensure high training throughput. We split the available device mesh into data and model parallel submeshes. The model parameters and optimizer states are sharded on the model submesh and replicated along data submesh. This avoids increasing the communication costs during the forward and backward passes by limiting the number of chips holding the shards of the model and the optimizer state. We refer to Table 1 for all key architecture parameters.

3 Instruction Fine-Tuning

3.1 Data mixture

We adopt the multilingual instruction data described in Üstün et al. [2024] for fine-tuning the pre-trained models. Given the scarcity of multilingual instruction data, these fine-tuning datasets combine a range of approaches to improve the availability of data. This includes relying on extensive efforts to aggregate and prune *multilingual templates* and hard-to-find *human annotations* curated by fluent speakers of various languages. Moreover, it also extends to data augmentation strategies such as *machine translation* and leveraging *synthetic data* generation coupled with translation.

We briefly describe each source below:

Prompt:	<BOS_TOKEN><|START_OF_TURN_TOKEN|><|USER_TOKEN|> Hello, how are you?<|END_OF_TURN_TOKEN|>
Completion:	<|START_OF_TURN_TOKEN|><|CHATBOT_TOKEN|> I am doing good!<|END_OF_TURN_TOKEN|>

Table 2: Example prompt-completion pair with the chat-format for the **Aya**-23 models. The formatting allows indication of roles (`user`, `chatbot`), and delineation of turns.

1. **Multilingual Templates**: We use structured text to transform specific NLP datasets into instruction and response pairs. This set of data includes samples from the xP3x dataset [Üstün et al., 2024], the data provenance collection [Longpre et al., 2023b], and the **Aya** collection [Singh et al., 2024]. The final collection consists of 55.7M examples which consists of zero and few-shot examples, covering 23 languages and 161 different datasets [Üstün et al., 2024].

2. **Human Annotations:** The **Aya** dataset [Singh et al., 2024] has a total of 204K human-curated prompt-response pairs written by native speakers in 65 languages. We filter this data for 23 languages we train on, resulting in 55K samples.

3. **Translated Data:** We use the translated subset of **Aya** collection [Singh et al., 2024] which open-sources translations of widely used English instruction datasets [Longpre et al., 2023b] filtered for the languages we train on. This collection includes, among others, translations of HotpotQA [Yang et al., 2018] and Flan-CoT-submix [Longpre et al., 2023a]. We randomly sample a subset of up to 3,000 instances for each language for each dataset to preserve instance-level diversity. We filter this data to the 23 languages we train on, resulting in a subset of 1.1M examples.

4. **Synthetic Data:** We construct synthetic fine-tuning data similar to Üstün et al. [2024] using human-annotated prompts from ShareGPT[5] and Dolly-15k [Conover et al., 2023b].[6] Unlike Üstün et al. [2024], we use Cohere's Command R+ to natively generate multilingual responses for the translated ShareGPT and Dolly prompts in all 23 languages, resulting in 1.63M examples. We note that Cohere's terms of use[7] prohibit training on model generations. However, we received a special exception for these releases of **Aya** models.

The **Aya** fine-tuning mix emphasizes available supervised datasets with self-reported commercially permissive licenses. We use the filtering tools from the Data Provenance Initiative [Longpre et al., 2023b] to ensure appropriate provenance.

3.2 Training details

For instruction fine-tuning, we fine-tune the base models for 13,200 update steps using an 8192 context length with data packing enabled, corresponding to approximately 10.5M training samples. We use the Adam optimizer [Kingma & Ba, 2014] with a cosine schedule learning rate, with a peak

[5] `https://sharegpt.com`; we do not use the original synthetic completions from ShareGPT dataset as they are generated from user-shared conversations with ChatGPT. We filter the prompts, following the same method as Üstün et al. [2024]

[6] We held out 200 selected prompts from Dolly-15k for open-ended evaluation following to Üstün et al. [2024]

[7] `https://cohere.com/terms-of-use`

Task	Dataset	Metric		Unseen Task	Languages
Discriminative Tasks					
Coreference Resolution	XWinograd [Muennighoff et al., 2023]	0-shot	Acc.	✔	6
Sentence Completion	XCOPA [Ponti et al., 2020]	0-shot	Acc.	✔	11
	XStoryCloze [Lin et al., 2021]	0-shot	Acc.	✔	10
Language Understanding	M-MMLU [Dac Lai et al., 2023]	5-shot	Acc.	✔	14
Generative Tasks					
Translation	FLORES-200 [Goyal et al., 2021; NLLB-Team et al., 2022]	0-shot	spBLEU	✗	23
Summarization	XLSum [Hasan et al., 2021]	0-shot	RougeL	✗	15
Mathematical Reasoning	MGSM [Shi et al., 2023]	5-shot	Acc.	✔	7
Open-Ended Generation	Dolly Human-edited & Machine-translated [Singh et al., 2024]	0-shot	win-rate	✗	5

Table 3: Datasets considered for evaluation. `Unseen Task` refers to tasks entirely excluded from training, which includes the 4 discriminative tasks. Additionally, we include multilingual MMLU as an unseen dataset. The seen tasks refer to the generative tasks where supervised training is performed and instances are held-out (`validation` and `test` splits) for evaluation. We limit the evaluation languages only to the ones that are included in 24 languages, except for the first 3 datasets (XWinograd, XCOPA, XStoryCloze) where we use all the available languages.

LR of 6×10^{-4}, an end LR of 6×10^{-5} and a batch size of 64. For all training runs, we use TPUv4 with up to 128 pod slices.

Similar to other instruction-tuned models [Gemini Team et al., 2024], the examples used to instruction-tune **Aya** 23 are formatted using special tokens to include extra information (an example is shown in Table 2). The formatting allows indication of roles (`user`, `chatbot`), and delineation of turns. This formatting is used both during instruction-tuning and inference. While it is possible to obtain coherent generations without using the formatting, generation quality suffers without it. While we use the chat formatting, the model is a single-turn instruction-following model and is not optimized explicitly for chat mode use.

4 Multilingual Evaluation

To measure our models' performance, we follow the comprehensive evaluation framework introduced in Üstün et al. [2024]. Different from Üstün et al. [2024], we use `eval-harness` [Gao et al., 2023] to evaluate all the models for discriminative tasks, multilingual MMLU, and MGSM.[8] This includes assessing performance on:

1. **Completely unseen discriminative tasks**: We evaluate on XWinograd [Muennighoff et al., 2023], XCOPA [Ponti et al., 2020], and XStoryCloze [Lin et al., 2021].[9] We use zero-shot evaluation. Note that these evaluation tasks are completely unseen and there is no dataset in the training mixture from the same task categories.

2. **General purpose language understanding**: We use Multilingual MMLU [Dac Lai et al., 2023] where the dataset is not seen during the training (5-shot evaluation) to evaluate **Aya**

[8] We only update the `eval-harness` code base to enable `bos_token` for **Aya** models similar to Gemma [Gemma-Team, 2024] to align with the tokenizer and data format.

[9] We omit XNLI [Conneau et al., 2018] due to the low performance for all the models evaluated, compared to Üstün et al. [2024]. We relate this to the different prompts used in `eval-harness`.

models' general language understanding. The dataset is a version of English MMLU [Hendrycks et al., 2020] translated into 31 languages using ChatGPT. The original English MMLU contains 13,062 questions consisting of 57 different tasks, covering a wide range of topics including STEM, humanities, and the social sciences. We use the 14 languages that are covered by **Aya** 23 models for evaluation.

3. **Multilingual mathematical reasoning**: We use Multilingual Grade School Math (MGSM) Benchmark [Shi et al., 2023] to evaluate multilingual mathematical reasoning. MGSM consists of 250 problems from the GSM8K benchmark [Cobbe et al., 2021], which are human-translated into 10 languages. We pick the subset of MGSM languages, which are covered by **Aya** 23 models. We use questions with answers followed by CoT prompt (`5-shot`) in the same language (`native_cot`) and `strict-match` score as the evaluation metric following Shi et al. [2023].

4. **Generative tasks**: We evaluate model performance in machine translation and summarization on FLORES-200 [NLLB-Team et al., 2022] and XLSum [Hasan et al., 2021] respectively. For FLORES, we use all 21 languages (X $\leftrightarrow$ English) and for XLSum, we use 15 languages based on language coverage of **Aya** 23 models.

5. **Preference evaluation**: We assess the open-ended generation capabilities of the models through human- and LLM-simulated evaluation using the (1) **dolly-machine-translated** test set Singh et al. [2024] which is a held-out test set of 200 instances from the Dolly-15k dataset [Conover et al., 2023b] translated into 101 languages. This test set was curated by multiple annotators to avoid the inclusion of any culturally specific or geographic references, intending to minimize estimations of performance that require specific cultural or geographic knowledge. We also evaluate on the (2) **dolly-human-edited** test set Singh et al. [2024] consisting of improved versions of the **dolly-machine-translated** test set for 6 languages (`French`, `Spanish`, `Serbian`, `Russian`, `Arabic`, `Hindi`) post-edited by professional compensated human annotators to correct any possible translation issues.

   For open-ended evaluation, we rely on both LLM-simulated win-rates and human evaluation. We detail the protocol for each briefly below:

   (a) **LLM-simulated win-rates**: Consistent with Üstün et al. [2024] and other recent works [Rafailov et al., 2023; Dubois et al., 2023; Kim et al., 2023], we use GPT-4[10] as a proxy judge. We measure pairwise win rates between **Aya** 23 models with **Aya** 101, Gemma-1.1-7b-it, and Mixtral-8x7b-Instruct-v0.1 on 10 languages (`English`, `Chinese`, `Turkish`, `Spanish`, `Russian`, `Hindi`, `French`, and `Arabic`, `Japanese`, `Portuguese`). We use the same prompt for eliciting GPT-4 preferences as specified by Üstün et al. [2024]. For languages where there is **dolly-human-edited** coverage, we default to these prompts given that they were edited for translation-induced issues by professional annotators.

   (b) **Human evaluation of preferences**: We ask compensated professional annotators in five languages (`Russian`, `Hindi`, `French`, `Spanish`, `English`) to select their preferred model completions for the **dolly-human-edited** test set and original English Dolly test prompts, respectively. The annotation setup (raters, instructions) is the same setup used by Üstün et al. [2024]. Each pair of generations is rated once; ties ("both bad" or "both good") are allowed but discouraged.

6. **Safety, Toxicity & Bias**: We evaluate the safety of model generations under adversarial prompts from the multilingual AdvBench [Yong et al., 2023a] benchmark representing multiple

[10] We use `gpt-4-turbo` as LLM judge: https://platform.openai.com/docs/models/gpt-4-turbo-and-gpt-4

angles of harm, such as crime, physical harm, and misinformation. GPT-4 is used as an automatic evaluator for harmfulness on 120 test prompts. The reliability of GPT-4 for this evaluation was previously confirmed by Üstün et al. [2024]. In addition, we measure toxicity and bias towards identity groups with the multilingual identity description prompts from Üstün et al. [2024]. We sample $k = 25$ model completions for each prompt, and evaluate their toxicity with Perspective API.[11]

4.1 Model Comparisons

We evaluate against multiple open-source massively multilingual models to ensure a comprehensive evaluation. We select models based on architecture, size, base model type, and the extent of coverage of languages. The selected models cover a range of sizes (7B to 46B), base models (mT5, Llama, Gemma, Mistral), languages, and training regimes (SFT and preference tuning).

Details of each model are below:

- **Aya-101-13B** [Üstün et al., 2024] is a 13B parameter mT5 model [Muennighoff et al., 2023] fine-tuned on xP3x [Üstün et al., 2024], **Aya** collection [Singh et al., 2024], Data Provenance collection [Longpre et al., 2023b], and ShareGPT-Command [Üstün et al., 2024] for 101 languages. **Aya** 101 is a state-of-art massively multilingual instruction-tuned LLM that covers the largest number of languages in our comparison.
- **Bactrian-X-7B** [Li et al., 2023a] is a LLaMA-7B model [Touvron et al., 2023a] fine-tuned on the Bactrian-X dataset which contains 3.4M pairs of instructions and responses in 52 languages. This dataset was automatically constructed by translating the Alpaca [Taori et al., 2023] and Dolly [Conover et al., 2023a] datasets using the Google Translate API.
- **Mistral-7B-Instruct-v0.2** [Jiang et al., 2023] is an open-source instruct fine-tuned edition of the Mistral-7B pre-trained model. The model is trained on instruction datasets publicly available on the HuggingFace repository.
- **Gemma-1.1-7B-it** [Gemma-Team, 2024] is a 7B parameter instruction fine-tuned model trained with Gemini models' architectures, data, and training recipes [Gemini-Team et al., 2024] on 6T tokens of data from web documents, mathematics, and code that are primarily English. In addition to the supervised fine-tuning, this model is also further fine-tuned using RLHF on collected pairs of preferences from human annotators.
- **Mixtral-8x7B-Instruct-v0.1** [Jiang et al., 2024] is a sparse mixture-of-experts model with 46.7B total parameters (active 12.9B parameters per token) that is instruction fine-tuned and preference-tuned using DPO [Rafailov et al., 2023]. The model supports five languages—English, French, Italian, German, and Spanish.

We do not compare our models to mT0 [Muennighoff et al., 2023] and Okapi [Dac Lai et al., 2023] models, as they have been shown to be significantly outperformed by the **Aya**-101-13B model [Üstün et al., 2024] which we do compare to as a baseline representative of the state-of-art in massively multilingual LLMs. We note that some of the models we evaluate such as Mistral and Gemma, do

[11] https://perspectiveapi.com/

Model	Held out tasks (Accuracy %)			
	XCOPA	XSC	XWG	Avg
Bactrian-X-7B	55.3	59.0	73.7	62.7
Mistral-7B-Instruct-v0.2	55.5	60.4	79.5	65.2
Gemma-1.1-7B-it	59.3	**63.1**	75.5	66.0
Aya-101-13B	59.7	60.4	66.3	62.1
Aya-23-8B	**59.8**	62.3	**80.7**	**67.6**
Mixtral-8x7B-Instruct-v0.1	59.9	63.4	83.1	68.8
Aya-23-35B	**62.8**	**65.1**	**84.4**	**70.8**

Table 4: Results for **discriminative unseen (held-out) task** evaluation. Results are reported as the zero-shot performance averaged across all languages of XCOPA, XStoryCloze, and XWinoGrad.

not explicitly claim to support multiple languages, however in practice, they are heavily used by multilingual users relative to explicitly multilingual models like mT0 [Muennighoff et al., 2023] and BLOOMZ [Dac Lai et al., 2023]. Furthermore, we also find that these models achieve considerable performance in many multilingual tasks as shown in our evaluation.

5 Results

5.1 Discriminative Tasks

Since all discriminative tasks were unseen during training, we measure zero-shot performance during evaluation. For these tasks, we use all the languages available in the evaluation datasets. In Table 4, we report average scores across all languages for XCOPA, XStoryCloze, and XWinoGrad along with an overall average across all tasks. We observe that across all tasks **Aya**-23-35B outperforms all baselines with an average of 70.8%.. Relative to other large models of comparable size, **Aya**-23-35B also outperforms Mixtral-8x7B-Instruct-v0.1 (70.8 vs 68.8).

Aya-23-8B achieves the best score within its class in terms of model size, with an average score of 67.6 compared to the next-best model Gemma-1.1-7B-it, which reaches an average score of 66. **Aya**-23-8B also outperforms Bactrian-X-7B, Mixtral-7B-Inst-v0.2, and **Aya**-101-13B.[12]

The significant performance improvements exhibited by **Aya**-23-8B and **Aya**-23-35B over the other models including **Aya**-101-13B, highlight the importance of a high-quality pre-trained base model and an emphasis on a smaller set of languages to achieve a strong performance by avoiding *the curse of multilinguality* [Conneau et al., 2019].

5.1.1 Multilingual MMLU

Table 5 presents multilingual MMLU [Hendrycks et al., 2020] results for all models on 14 languages which is a subset of multilingual MMLU languages [Dac Lai et al., 2023] that are covered by **Aya** 23 models. We use 5-shot evaluation following the English MMLU benchmark [Beeching et al., 2023].

[12]Note that our evaluation framework along with the zero-shot prompts for these tasks differs from Üstün et al. [2024], which leads to a difference in **Aya**-101-13B performance compared to the original paper.

	ar	de	es	fr	hi	id	it	nl	pt	ro	ru	uk	vi	zh	Avg
Bactrian-X-7B	26.9	32.1	32.6	31.2	27.5	28.6	31.1	31.8	31.4	30.6	29.7	28.7	26.4	29.3	29.9
Mistral-7B-Instruct-v0.2	32.7	48.5	50.6	49.7	30.8	43.6	48.8	48.1	50.2	46.7	46.5	46.0	38.4	43.9	44.6
Gemma-1.1-7B-it	40.8	49.7	**51.8**	**51.6**	40.1	48.3	50.0	48.4	51.1	47.4	47.2	46.0	46.2	**47.7**	47.6
Aya-101-13B	39.8	42.6	42.2	42.5	38.4	41.9	41.2	42.3	41.5	40.4	41.8	41.0	40.1	40.4	41.1
Aya-23-8B	**45.1**	**50.0**	50.9	51.0	**39.7**	**48.8**	**50.7**	**49.7**	**50.8**	**49.9**	**47.8**	**46.8**	**46.5**	47.1	**48.2**
Mixtral-8x7B-Instruct-v0.1	41.8	**63.7**	**65.2**	**64.9**	37.8	55.4	**64.3**	**62.2**	**63.7**	**60.6**	**59.0**	**57.8**	48.8	54.7	57.1
Aya-23-35B	**53.9**	60.4	61.6	62.0	**47.8**	**58.9**	61.5	60.3	62.0	59.7	57.8	56.3	**55.3**	**57.5**	**58.2**

Table 5: **Multilingual MMLU** (`5-shot`) results for **Aya** 23 models and **Aya** 101, Bactrian-X, Gemma-7B, Mistral-7B and Mixtral-8x7B in 14 languages.

	de	en	es	fr	ja	ru	zh	Avg
Bactrian-X-7B	5.6	7.2	5.6	6.0	4.0	4.0	4.8	5.3
Mistral-7B-Instruct-v0.2	34.4	31.2	29.2	32.8	6.0	31.6	30.4	27.9
Gemma-1.1-7B-it	35.6	45.2	38.4	**41.6**	6.0	**39.2**	32.0	34.0
Aya-101-13B	9.6	10.0	8.4	8.8	4.0	10.8	4.8	8.1
Aya-23-8B	**40.4**	**48.0**	**45.2**	38.8	**12.8**	38.0	**32.8**	**36.6**
Mixtral-8x7B-Instruct-v0.1	58.8	60.0	55.2	52.8	**24.4**	56.0	44.4	50.2
Aya-23-35B	**61.6**	**68.4**	**58.4**	**55.6**	22.8	**58.0**	**50.8**	**53.7**

Table 6: **Multilingual Grade School Math benchmark (MGSM)** results for baselines and **Aya** models. We use questions with answers followed by CoT prompt (`5-shot`) in the same language (`native_cot`) as the dataset and `strict-match` score as the evaluation metric.

Similar to zero-shot unseen tasks, **Aya**-23-8B performs overall best among comparable "smaller" models, achieving an average of 48.2% accuracy across all languages and the highest score in 11 languages out of 14 for its class. At the larger model scale, **Aya**-23-35B outperforms Mixtral-8x7B-Inst on average (58.2 vs 57.1). Here, Mixtral performs slightly better in relatively high resource languages, however, especially for non-European languages such as Arabic, Hindi, and Vietnamese, **Aya**-23-35B scores significantly higher with a 12.1%, 10.0% and 6.5% respective accuracy increase for these 3 languages.

5.2 Multilingual Mathematical Reasoning

On MGSM, **Aya** 23 models outperform all in-class baselines, indicating strong mathematical reasoning ability across languages. **Aya**-23-8B achieves a score of 36.6 averaged over 7 languages compared to Gemma-1.1-7b-it's score of 34.0 which is the next-best model in its class. Notably, **Aya**-23-8B achieves a 4.5x increase in performance compared to **Aya**-101-13B (36.6 vs 8.1), showing the significant impact of the high-quality pre-trained model once more. For the larger scale models, **Aya**-23-35B outperforms Mixtral-8x7B-Instruct-v0.1 by achieving a score of 53.7 compared to 50.2. When looking at individual language scores, **Aya** 23 models outperform the strongest in-class models for every language with the exception of French and Russian for **Aya**-23-8B, and Japanese for **Aya**-23-35B.

Model	Generative Tasks		
	FLORES-200 (spBleu)		XLSum (RougeL)
	X→ En	En → X	
Bactrian-X-7B	25.9	16.6	7.7
Mistral-7B-Instruct-v0.2	31.1	21.0	6.3
Gemma-1.1-7B-it	32.0	25.6	13.0
Aya-101-13B	35.9	30.4	**27.5**
Aya-23-8B	**39.5**	**34.8**	**27.5**
Mixtral-8x7B-Instruct-v0.1	36.3	28.9	7.1
Aya-23-35B	**43.0**	**37.8**	**30.9**

Table 7: **Translation (FLORES)** and **multilingual summarization (XLSum)** results for baselines and **Aya** models. For XLSUM, we evaluate models on 15 languages that are included in **Aya** 23, and for FLORES we use all 22 languages paired with English.

5.3 Generative Tasks

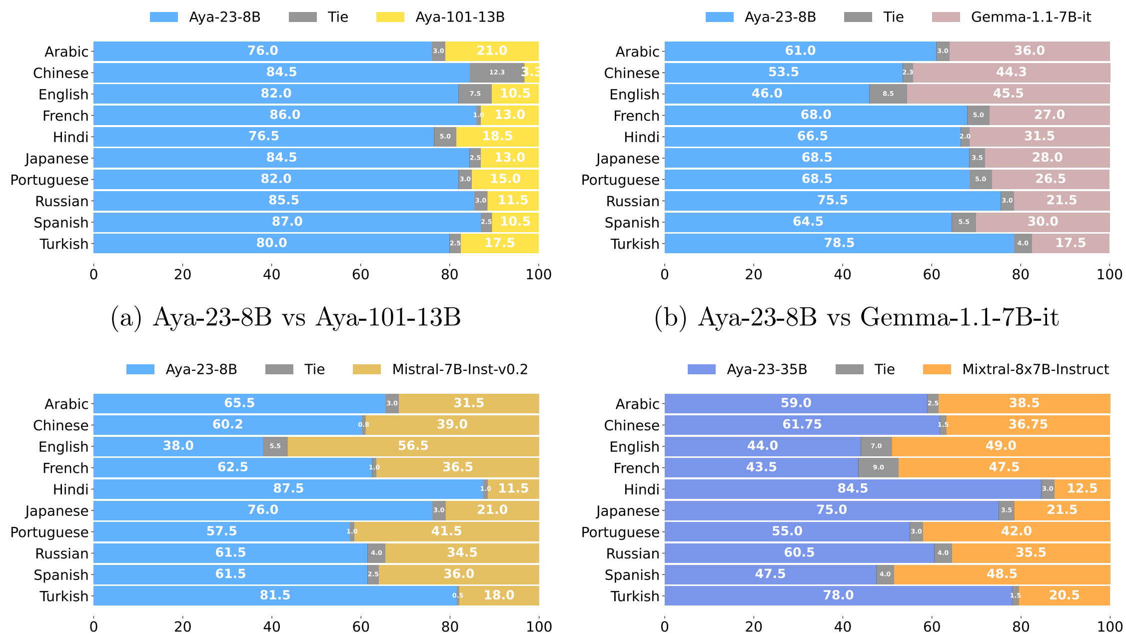
Table 7 presents the results for translation (FLORES) and multilingual summarization (XLSum). For FLORES, we use all 23 languages paired with English (X↔EN). For XLSum, we use 15 languages that are available and covered by **Aya** 23 models. In this evaluation, **Aya** 23 models achieve significantly higher results than other models with similar sizes. **Aya**-23-8B achieves an average spBleu score of 37.2, outperforming the second best model **Aya**-101-13B by 4 points. In XLSum, **Aya**-23-8B and **Aya**-101-13B are on par with an average RougeL score of 27.5 surpassing the next-best model Gemma-1.1 by 14.5 points.

For large model size, **Aya**-23-35B outperforms Mixtral-8x7B by 7.8 spBleu (40.4 vs 32.6) in translation and 23.8 (30.9 vs 7.1) in summarization. We find that both Mistral-7B and Mixtral-8x7B tend to generate English responses to the prompt although the context is in the target language, leading to poor performance in multilingual summarization.

5.4 Simulated Win Rates and Human Eval

GPT-4 Win Rates We perform automatic model ranking using GPT-4 as a judge comparing generations for 200 held-out prompts from **dolly-human-edited** and **dolly-machine-translated** [Singh et al., 2024]. **Aya** 23 models exhibit superior win rates averaged over all languages against the strongest in-class baseline models as shown in Figure 1. **Aya**-23-8B outperforms Aya-101-13B, Mistral-7B-Instruct-v0.2, and Gemma-1.1-7B-it achieving average win rates of 82.4%, 65.2%, and 65.0% respectively. **Aya**-23-35B outperforms Mixtral-8x7B-Instruct-v0.1 with an average win-rate of 60.9%.

Figure 3 shows win rates broken down for 10 languages, against the strongest models of similar size. **Aya** 23 models achieve superior win rates across all languages against all in-class baseline models with the exception of English for Mistral-7B-Instruct-v0.2 for **Aya**-23-8B and English/French/Spanish for Mixtral-8x7B-Instruct-0.1 for **Aya**-23-35B. Especially for non-European languages such as Turkish, Hindi, and Japanese **Aya** 24 models outperform comparison models by a significant margin: **Aya**-23-8B wins 81.5%, 87.5%, and 76.0% of the time against Mistal-7B while **Aya**-24-35B wins 78.0%, 84.5% and 75.0% of the time againist Mixtral-8x7B.



(a) Aya-23-8B vs Aya-101-13B

(b) Aya-23-8B vs Gemma-1.1-7B-it

(c) Aya-23-8B vs Mistral-7B-Instruct-v0.2

(d) Aya-23-35B vs Mixtral-8x7B-Instruct-v0.1

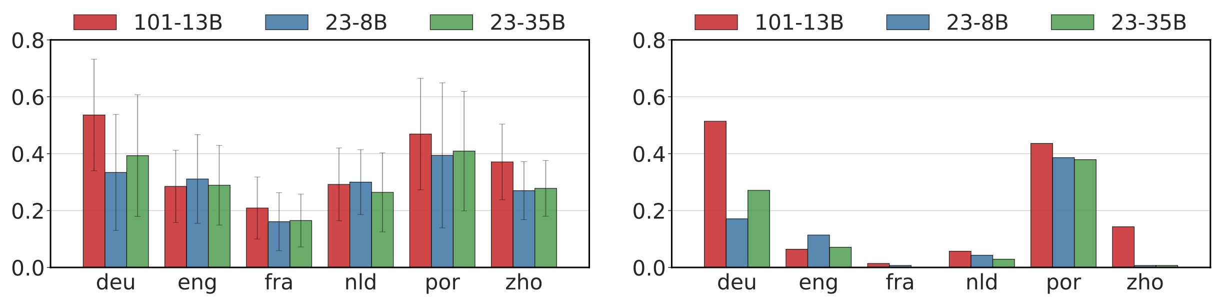
Figure 3: **LLM-as-a-judge evaluation (% win rates)** for 10 languages comparing **Aya**-23 models with similar size models for 10 languages. We use `gpt-4-turbo` for these evaluation as the judge LLM.

Finally, among models that include a similar instruction fine-tuning mixture, **Aya**-23-8B is heavily preferred to **Aya**-101-13B in all 10 languages, showing the significant impact of a stronger pre-trained model.

	English	French	Hindi	Russian	Spanish	**Avg**
Aya-101-13B	**44.0**	33.8	37.0	31.0	32.0	35.6
Aya-23-8B	43.0	**56.1**	**43.0**	**59.5**	**52.5**	**50.8**
Aya-101-13B	35.5	30.0	34.3	28.0	26.0	30.8
Aya-23-35B	**58.5**	**60.0**	**50.5**	**63.5**	**55.5**	**57.6**
Aya-23-8B	36.5	42.7	25.6	39.5	**41.2**	37.1
Aya-23-35B	**40.0**	**48.7**	**33.7**	**47.0**	39.2	**41.7**

Table 8: **Human evaluation results (% win rates)** for pairwise comparisons between each pair of models. The remaining percentages are ties. The respective higher average win-rates are boldfaced.

Human Evaluation Table 8 presents win rates resulting from human preference ratings, comparing the **Aya** 23 models with **Aya**-101-13B. We observe that with the stronger pre-trained model, **Aya** 23 family models consistently outperform the mT5-based **Aya**-101-13B on all evaluated languages. In particular, **Aya**-23-8B, despite its smaller size wins against **Aya**-101-13B for 50.8% of prompts on average across languages. Furthermore, **Aya**-23-35B achieves 57.6% win-rate against **Aya**-101-13B.



(a) Expected maximum toxicity

(b) Toxicity probability

Figure 4: Toxicity analysis of **Aya** models (101: **Aya**-101, 23-8B: **Aya**-23-8B, 23-35B: **Aya**-23-35B) generations when prompted with sentences for identity groups such as gender, ethnicity, and religion.

We note that human evaluation has been conducted using intermediate checkpoints of **Aya** 23 models before finalizing our model training due to the required time and cost for these evaluations. We expect higher win-rates for the final **Aya** 23 models against **Aya**-101-13B for human evaluation, based on GPT4 win-rates and our internal comparison.

	Arabic	English	Hindi	Italian	Simplified Chinese	Ukrainian	**Avg**
Aya-101-13B	81.6	83.3	81.7	93.3	75.8	88.3	84.0
Aya-23-8B	42.5	56.1	51.7	51.7	55.8	53.6	51.9
Aya-23-35B	**11.7**	**21.7**	**37.5**	**40.0**	**27.5**	**19.2**	**26.2**

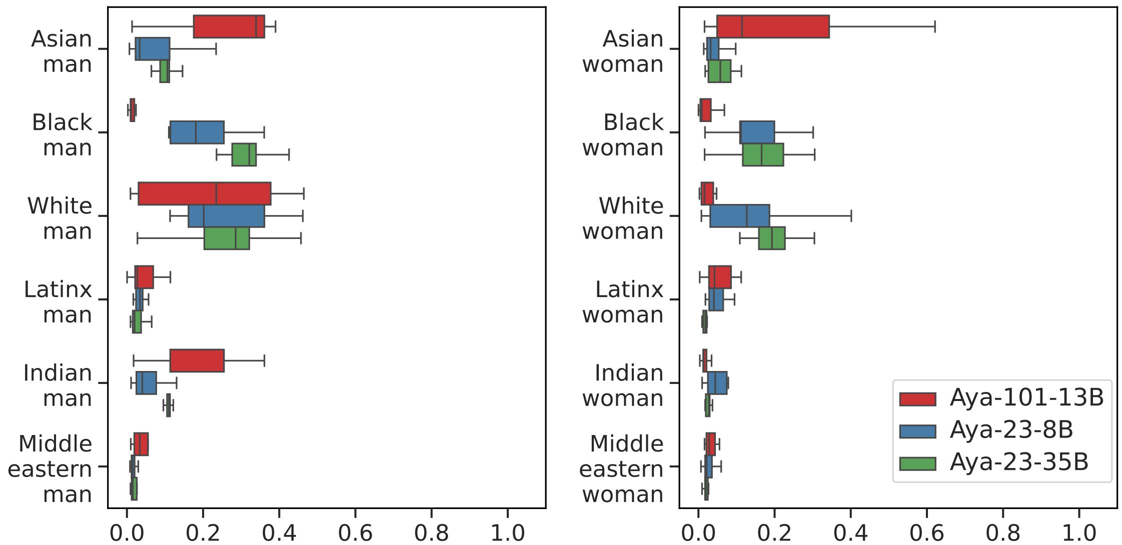
Table 9: Multilingual AdvBench results: percentage of harmful responses as judged by GPT-4. Lower is better.

5.5 Safety, Toxicity & Bias

Safety Table 9 reports the percentage of harmful model completions for the 120 adversarial test split prompts from multilingual AdvBench for 6 languages, as judged by GPT-4.

Comparing **Aya** 23 models with the **Aya**-101-13B model previously benchmarked in [Üstün et al., 2024], we find that the rate of harmful responses is lower for all languages, and on average reduced by at least half. The larger capacity of the **Aya**-23-35B model further helps to lower the harmfulness of the responses, especially for Arabic and Italian, presumably due to a beneficial effect of improved cross-lingual transfer. In terms of quality, we notice that in particular the refusal responses are more eloquent, diverse, and elaborate than those of the **Aya**-101-13B model which is a reflection of the improved generation quality assessed above.

It is important to note that none of the three models have undergone any targeted safety alignment in the multilingual fine-tuning stage beyond learning from incidental safety examples in synthetically generated examples from Command R+. These scores therefore reflect how much alignment would still be needed for the specific safety cases captured in AdvBench, rather than how much they are already aligned.



(a) Racial Groups (Man) (b) Racial Groups (Woman)

Figure 5: Perspective API toxicity scores for **Aya**-101, **Aya**-23-7B and **Aya**-23-35B generations given input prompts in `English` for racial identity groups.

Toxicity & Bias Figure 4 shows the expected maximum toxicity and toxicity probability for model completions of the identity group descriptions prompts. We observe that both **Aya** 23 models generally have lower expected maximum toxicity and a lower toxicity probability than the **Aya**-101-13B model. This holds true for all languages except English, where the toxicity is slightly higher for the new **Aya** 23 models. Inspecting English generations further, Figure 5 details the toxicity in descriptions of different racial groups and genders. We note that **Aya** 23 models tend to produce less toxic generations describing Asians, Latinx, but have a much higher chance to produce toxic descriptions of Blacks and Whites, especially for women.

6 Conclusion

While language technologies have made rapid strides in recent years, this progress has been predominantly concentrated in the English language. Given the increasing importance of cross-cultural communication for a broad range of social, economic, and political activities, there is a growing imperative to broaden this progress to other languages so that language technologies can better reflect the reality of the world and more effectively contribute to its more equitable development. We introduce a new family of multilingual models, **Aya** 23, to advance our mission of using multilingual technologies to empower a multilingual world. Our extensive evaluation demonstrates the high performance of these models on a broad range of multilingual benchmarks and human evaluation. By releasing these model weights, we hope this work will contribute to furthering future research towards this critical mission.

6.1 Limitations

While **Aya** 23 greatly improves performance for the subset of 23 languages chosen and are far more comprehensive in coverage than most open weight releases, we recognize that this subset is only a tiny fraction of the world's linguistic diversity; of the world's approximately 7,000 languages [eth, 2023], only half of them are captured in any sort of written form [Adda et al., 2016]. Of this half, only a few hundred are included on the internet in machine readable corpora [Adda et al., 2016]. More work is needed to improve both coverage and performance simultaneously.

Additionally, it is important to acknowledge that the languages covered by these models are still limited to those present during pre-training, with a particular bias towards languages prevalent in certain regions of the world. Specifically, the pre-training coverage underrepresents languages spoken in Asia and Africa. This limitation is a critical area that requires ongoing effort and attention. We aim to address this gap and improve language inclusivity as part of the broader **Aya** Initiative[13], with a dedicated focus on these underrepresented languages.

Building upon the foundation laid by the original **Aya** model, which prioritized breadth, future work will concentrate on enhancing coverage and performance for these remaining languages. This includes developing tailored language models, improving data collection and representation, and addressing any cultural and linguistic nuances to ensure equitable and effective language technologies for all.

7 Acknowledgements

We thank the Hugging Face team for helping us with our open-weights release including Younes Belkada, Matthew Carrigan, Lysandre Debut, Clémentine Fourrier, Nathan Habib, Quentin Lhoest, Omar Sanseviero, Daniel van Strien, and Arthur Zucker. We thank Aakanksha for sharing their evaluation code for FLORES and XLSum, and Zheng-Xin Yong for the toxicity evaluation.

Thanks to colleagues who have supported various aspects of this project: Linus Chui, Manoj Govindassamy, Yina Moe-Lange, Morgan Norman, Shubham Shukla.

References

Ethnologue. https://www.ethnologue.com/insights/how-many-languages/, 2023. Accessed: 2023-06-17.

Gilles Adda, Sebastian Stüker, Martine Adda-Decker, Odette Ambouroue, Laurent Besacier, David Blachon, Hélène Bonneau-Maynard, Pierre Godard, Fatima Hamlaoui, Dmitry Idiatov, Guy-Noël Kouarata, Lori Lamel, Emmanuel-Moselly Makasso, Annie Rialland, Mark Van de Velde, François Yvon, and Sabine Zerbian. Breaking the unwritten language barrier: The bulb project. *Procedia Computer Science*, 81:8–14, 2016. ISSN 1877-0509. doi: https://doi.org/10.1016/j.procs.2016.04.023. URL https://www.sciencedirect.com/science/article/pii/S1877050916300370. SLTU-2016 5th Workshop on Spoken Language Technologies for Under-resourced languages 09-12 May 2016 Yogyakarta, Indonesia.

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

[13]https://cohere.com/research/aya

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. Palm 2 technical report. *arXiv*, abs/2305.10403, 2023.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, et al. Massively multilingual neural machine translation in the wild: Findings and challenges. *arXiv preprint arXiv:1907.05019*, 2019.

Edward Beeching, Clémentine Fourrier, Nathan Habib, Sheon Han, Nathan Lambert, Nazneen Rajani, Omar Sanseviero, Lewis Tunstall, and Thomas Wolf. Open llm leaderboard. https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard, 2023.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R Bowman, Holger Schwenk, and Veselin Stoyanov. Xnli: Evaluating cross-lingual sentence representations. pp. 2475–2485, October-November 2018. doi: 10.18653/v1/D18-1269. URL https://aclanthology.org/D18-1269.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. pp. 8440–8451, July 2019. doi: 10.18653/v1/2020.acl-main.747. URL https://aclanthology.org/2020.acl-main.747.

Mike Conover, Matt Hayes, Ankit Mathur, Xiangrui Meng, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, et al. Free dolly: Introducing the world's first truly open instruction-tuned llm. *Databricks*, 2023a.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. Free dolly: Introducing the world's first truly open instruction-tuned llm, 2023b. URL https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm.

Viet Dac Lai, Chien Van Nguyen, Nghia Trung Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan A Rossi, and Thien Huu Nguyen. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. *arXiv e-prints*, pp. arXiv–2307, 2023.

Yue Deng, Wenxuan Zhang, Sinno Jialin Pan, and Lidong Bing. Multilingual jailbreak challenges in large language models. *arXiv preprint arXiv:2310.06474*, 2023.

Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. Alpacafarm: A simulation framework for methods that learn from human feedback. *arXiv preprint arXiv:2305.14387*, 2023.

Esin Durmus, Karina Nyugen, Thomas I. Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. Towards measuring the representation of subjective global opinions in language models. *arXiv*, abs/2306.16388, 2023.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. A framework for few-shot language model evaluation. 12 2023. doi: 10.5281/zenodo.10256836. URL https://zenodo.org/records/10256836.

Gemini-Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, 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, Jack Krawczyk, Cosmo Du, Ed Chi, Heng-Tze Cheng, Eric Ni, Purvi Shah, Patrick Kane, Betty Chan, Manaal Faruqui, Aliaksei Severyn, Hanzhao Lin, YaGuang Li, Yong Cheng, Abe Ittycheriah, Mahdis Mahdieh, Mia Chen, Pei Sun, Dustin Tran, Sumit Bagri, Balaji Lakshminarayanan, Jeremiah Liu, Andras Orban, Fabian Güra, Hao Zhou, Xinying Song, Aurelien Boffy, Harish Ganapathy, Steven Zheng, HyunJeong Choe, Ágoston Weisz, Tao Zhu, Yifeng Lu, Siddharth Gopal, Jarrod Kahn, Maciej Kula, Jeff Pitman, Rushin Shah, Emanuel Taropa, Majd Al Merey, Martin Baeuml, Zhifeng Chen, Laurent El Shafey, Yujing Zhang, Olcan Sercinoglu, 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, Gaurav Singh Tomar, Evan Senter, Martin Chadwick, Ilya Kornakov, Nithya Attaluri, Iñaki Iturrate, Ruibo

Liu, Yunxuan Li, Sarah Cogan, Jeremy Chen, Chao Jia, Chenjie Gu, Qiao Zhang, Jordan Grimstad, Ale Jakse Hartman, Xavier Garcia, 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, Ravi Addanki, Antoine Miech, Annie Louis, Denis Teplyashin, Geoff Brown, Elliot Catt, 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, Vitaliy 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, 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, 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 Campos, Alex Tomala, Yun-hao 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, 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, Sidharth Mudgal, Romina Stella, Kevin Brooks, Gautam Vasudevan, Chenxi Liu, Mainak Chain, Nivedita Melinkeri, Aaron Cohen, Venus Wang, Kristie Seymore, Sergey Zubkov, Rahul Goel, Summer Yue, Sai Krishnakumaran, Brian Albert, Nate Hurley, Motoki Sano, Anhad Mohananey, Jonah Joughin, Egor Filonov, Tomasz Kępa, Yomna Eldawy, Jiawern Lim, Rahul Rishi, Shirin Badiezadegan, Taylor Bos, Jerry Chang, Sanil Jain, Sri Gayatri Sundara Padmanabhan, Subha Puttagunta, Kalpesh Krishna, Leslie Baker, Norbert Kalb, Vamsi Bedapudi, Adam Kurzrok, Shuntong Lei, Anthony Yu, Oren Litvin, Xiang Zhou, Zhichun Wu, Sam Sobell, Andrea Siciliano, Alan Papir, Robby Neale, Jonas Bragagnolo, Tej Toor, Tina Chen, Valentin Anklin, Feiran Wang, Richie Feng, Milad Gholami, Kevin Ling, Lijuan Liu, Jules Walter, Hamid Moghaddam, Arun Kishore, Jakub Adamek, Tyler Mercado, Jonathan Mallinson, Siddhinita Wandekar, Stephen Cagle, Eran Ofek, Guillermo Garrido, Clemens Lombriser, Maksim Mukha, Botu Sun, Hafeezul Rahman Mohammad, Josip Matak, Yadi Qian, Vikas Peswani, Pawel Janus, Quan Yuan, Leif Schelin, Oana David, Ankur Garg, Yifan He, Oleksii Duzhyi, Anton Älgmyr, Timothée Lottaz, Qi Li, Vikas Yadav, Luyao Xu, Alex Chinien, Rakesh Shivanna, Aleksandr Chuklin, Josie Li, Carrie Spadine, Travis Wolfe, Kareem Mohamed, Subhabrata Das, Zihang Dai, Kyle He, Daniel von Dincklage, Shyam Upadhyay, Akanksha Maurya, Luyan Chi, Sebastian Krause, Khalid Salama, Pam G Rabinovitch, Pavan Kumar Reddy M, Aarush Selvan, Mikhail Dektiarev, Golnaz Ghiasi, Erdem Guven, Himanshu Gupta, Boyi Liu, Deepak Sharma, Idan Heimlich Shtacher, Shachi Paul, Oscar Akerlund, François-Xavier Aubet, Terry Huang, Chen Zhu, Eric Zhu, Elico Teixeira, Matthew Fritze, Francesco Bertolini, Liana-Eleonora Marinescu, Martin Bölle, Dominik Paulus, Khyatti Gupta, Tejasi Latkar, Max Chang, Jason Sanders, Roopa Wilson, Xuewei Wu, Yi-Xuan Tan, Lam Nguyen Thiet, Tulsee Doshi, Sid Lall, Swaroop Mishra, Wanming Chen, Thang Luong, Seth Benjamin, Jasmine Lee, Ewa Andrejczuk, Dominik Rabiej, Vipul Ranjan, Krzysztof Styrc, Pengcheng Yin, Jon Simon, Malcolm Rose Harriott, Mudit Bansal, Alexei Robsky, Geoff Bacon, David Greene, Daniil Mirylenka, Chen Zhou, Obaid Sarvana, Abhimanyu Goyal, Samuel Andermatt, Patrick Siegler, Ben Horn, Assaf Israel, Francesco Pongetti, Chih-Wei "Louis" Chen, Marco Selvatici, Pedro Silva, Kathie Wang, Jackson Tolins, Kelvin Guu, Roey Yogev, Xiaochen Cai, Alessandro Agostini, Maulik Shah, Hung Nguyen, Noah Ó Donnaile, Sébastien Pereira, Linda Friso, Adam Stambler, Adam Kurzrok, Chenkai Kuang, Yan Romanikhin, Mark Geller, ZJ Yan, Kane Jang, Cheng-Chun Lee, Wojciech Fica, Eric Malmi, Qijun Tan, Dan Banica, Daniel Balle, Ryan Pham, Yanping Huang, Diana Avram, Hongzhi Shi, Jasjot Singh, Chris Hidey, Niharika Ahuja, Pranab Saxena, Dan Dooley, Srividya Pranavi Potharaju, Eileen O'Neill, Anand Gokulchandran, Ryan Foley, Kai Zhao, Mike Dusenberry, Yuan Liu, Pulkit Mehta, Ragha Kotikalapudi, Chalence Safranek-Shrader, Andrew Goodman, Joshua Kessinger, Eran Globen, Prateek Kolhar, Chris Gorgolewski, Ali Ibrahim, Yang Song, Ali Eichenbaum, Thomas Brovelli, Sahitya Potluri, Preethi Lahoti, Cip Baetu, Ali

Ghorbani, Charles Chen, Andy Crawford, Shalini Pal, Mukund Sridhar, Petru Gurita, Asier Mujika, Igor Petrovski, Pierre-Louis Cedoz, Chenmei Li, Shiyuan Chen, Niccolò Dal Santo, Siddharth Goyal, Jitesh Punjabi, Karthik Kappaganthu, Chester Kwak, Pallavi LV, Sarmishta Velury, Himadri Choudhury, Jamie Hall, Premal Shah, Ricardo Figueira, Matt Thomas, Minjie Lu, Ting Zhou, Chintu Kumar, Thomas Jurdi, Sharat Chikkerur, Yenai Ma, Adams Yu, Soo Kwak, Victor Ähdel, Sujeevan Rajayogam, Travis Choma, Fei Liu, Aditya Barua, Colin Ji, Ji Ho Park, Vincent Hellendoorn, Alex Bailey, Taylan Bilal, Huanjie Zhou, Mehrdad Khatir, Charles Sutton, Wojciech Rzadkowski, Fiona Macintosh, Konstantin Shagin, Paul Medina, Chen Liang, Jinjing Zhou, Pararth Shah, Yingying Bi, Attila Dankovics, Shipra Banga, Sabine Lehmann, Marissa Bredesen, Zifan Lin, John Eric Hoffmann, Jonathan Lai, Raynald Chung, Kai Yang, Nihal Balani, Arthur Bražinskas, Andrei Sozanschi, Matthew Hayes, Héctor Fernández Alcalde, Peter Makarov, Will Chen, Antonio Stella, Liselotte Snijders, Michael Mandl, Ante Kärrman, Paweł Nowak, Xinyi Wu, Alex Dyck, Krishnan Vaidyanathan, Raghavender R, Jessica Mallet, Mitch Rudominer, Eric Johnston, Sushil Mittal, Akhil Udathu, Janara Christensen, Vishal Verma, Zach Irving, Andreas Santucci, Gamaleldin Elsayed, Elnaz Davoodi, Marin Georgiev, Ian Tenney, Nan Hua, Geoffrey Cideron, Edouard Leurent, Mahmoud Alnahlawi, Ionut Georgescu, Nan Wei, Ivy Zheng, Dylan Scandinaro, Heinrich Jiang, Jasper Snoek, Mukund Sundararajan, Xuezhi Wang, Zack Ontiveros, Itay Karo, Jeremy Cole, Vinu Rajashekhar, Lara Tumeh, Eyal Ben-David, Rishub Jain, Jonathan Uesato, Romina Datta, Oskar Bunyan, Shimu Wu, John Zhang, 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, Jane Park, Jiageng Zhang, Jeff Stanway, Drew Garmon, Abhijit Karmarkar, Zhe Dong, Jong Lee, Aviral Kumar, Luowei Zhou, Jonathan Evens, William Isaac, Geoffrey Irving, Edward Loper, Michael Fink, Isha Arkatkar, Nanxin Chen, Izhak Shafran, Ivan Petrychenko, Zhe Chen, Johnson Jia, Anselm Levskaya, Zhenkai Zhu, Peter Grabowski, Yu Mao, Alberto Magni, Kaisheng Yao, Javier Snaider, Norman Casagrande, Evan Palmer, Paul Suganthan, 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, Marcus Wu, Ricardo Aguilar, Keith Pallo, Abhishek Chakladar, Ginger Perng, Elena Allica Abellan, Mingyang Zhang, Ishita Dasgupta, Nate Kushman, Ivo Penchev, Alena Repina, Xihui Wu, Tom van der Weide, Priya Ponnapalli, Caroline Kaplan, Jiri Simsa, Shuangfeng Li, Olivier Dousse, Fan Yang, Jeff Piper, Nathan Ie, Rama Pasumarthi, Nathan Lintz, Anitha Vijayakumar, Daniel Andor, Pedro Valenzuela, Minnie Lui, Cosmin Paduraru, Daiyi Peng, Katherine Lee, Shuyuan Zhang, Somer Greene, Duc Dung Nguyen, Paula Kurylowicz, Cassidy Hardin, Lucas Dixon, Lili Janzer, Kiam Choo, Ziqiang Feng, Biao Zhang, Achintya Singhal, 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, Ken Franko, Anna Bulanova, Rémi Leblond, Shirley Chung, Harry Askham, Luis C. Cobo, Kelvin Xu, Felix Fischer, Jun Xu, Christina Sorokin, Chris Alberti, Chu-Cheng Lin, Colin Evans, Alek Dimitriev, Hannah Forbes, Dylan Banarse, Zora Tung, Mark Omernick, Colton Bishop, Rachel Sterneck, Rohan Jain, Jiawei Xia, Ehsan Amid, Francesco Piccinno, Xingyu Wang, Praseem Banzal, Daniel J. Mankowitz, Alex Polozov, Victoria Krakovna, Sasha Brown, MohammadHossein Bateni, Dennis Duan, Vlad Firoiu, Meghana Thotakuri, Tom Natan, Matthieu Geist, Ser tan Girgin, Hui Li, Jiayu Ye, Ofir Roval, Reiko Tojo, Michael Kwong, James Lee-Thorp, Christopher Yew, Danila Sinopalnikov, Sabela Ramos, John Mellor, Abhishek Sharma, Kathy Wu, 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, Tian Huey Teh, Jason Sanmiya, Evgeny Gladchenko, Nejc Trdin, 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, Diane Wu, Denese Owusu-Afriyie, Cosmo Du, Chloe Thornton, Jordi Pont-Tuset, Pradyumna Narayana, Jing Li, Saaber Fatehi, John Wieting, Omar Ajmeri, Benigno Uria, Yeongil Ko, Laura Knight, Amélie Héliou, Ning Niu, Shane Gu, Chenxi Pang, Yeqing Li, Nir Levine, Ariel Stolovich, Rebeca Santamaria-Fernandez, Sonam Goenka, Wenny Yustalim, Robin Strudel, Ali Elqursh, Charlie Deck, Hyo Lee, Zonglin Li, Kyle Levin, Raphael Hoffmann, Dan Holtmann-Rice, Olivier Bachem, Sho Arora, Christy Koh, Soheil Hassas Yeganeh, Siim Põder, Mukarram Tariq, Yanhua Sun, Lucian Ionita, Mojtaba Seyedhosseini, Pouya Tafti, 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, Vinod Koverkathu, Christopher A. Choquette-Choo, Yunjie Li, TJ Lu, Abe Ittycheriah, Prakash Shroff, Mani Varadarajan, Sanaz Bahargam, Rob Willoughby, David Gaddy, 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, 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, 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, Lynette Webb, Sahil Dua, Dong Li, 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, Evgenii Eltyshev, 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, Christof Angermueller, Xiaowei Li, Anoop Sinha, 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, Ken Durden, Harsh Mehta, Nikola Momchev, Elahe Rahimtoroghi, Maria Georgaki, Amit Raul, Sebastian Ruder, Morgan Redshaw, Jinhyuk Lee, Denny Zhou, Komal Jalan, Dinghua Li, Blake Hechtman, Parker Schuh, Milad Nasr, Kieran Milan, Vladimir Mikulik, Juliana Franco, Tim Green, Nam Nguyen, Joe Kelley, Aroma Mahendru, Andrea Hu, Joshua Howland, Ben Vargas, Jeffrey Hui, Kshitij Bansal, Vikram Rao, Rakesh Ghiya, Emma Wang, Ke Ye, Jean Michel Sarr, Melanie Moranski Preston, Madeleine Elish, Steve Li, Aakash Kaku, Jigar Gupta, Ice Pasupat, Da-Cheng Juan, Milan Someswar, Tejvi M., Xinyun Chen, Aida Amini, Alex Fabrikant, Eric Chu, Xuanyi Dong, Amruta Muthal, Senaka Buthpitiya, Sarthak Jauhari, Nan Hua, Urvashi Khandelwal, Ayal Hitron, Jie Ren, Larissa Rinaldi, Shahar Drath, Avigail Dabush, Nan-Jiang Jiang, Harshal Godhia, Uli Sachs, Anthony Chen, Yicheng Fan, Hagai Taitelbaum, Hila Noga, Zhuyun Dai, James Wang, Chen Liang, Jenny Hamer, Chun-Sung Ferng, Chenel Elkind, Aviel Atias, Paulina Lee, Vít Listík, Mathias Carlen, Jan van de Kerkhof, Marcin Pikus, Krunoslav Zaher, Paul Müller, Sasha Zykova, Richard Stefanec, Vitaly Gatsko, Christoph Hirnschall, Ashwin Sethi, Xingyu Federico Xu, Chetan Ahuja, Beth Tsai, Anca Stefanoiu, Bo Feng,

Keshav Dhandhania, Manish Katyal, Akshay Gupta, Atharva Parulekar, Divya Pitta, Jing Zhao, Vivaan Bhatia, Yashodha Bhavnani, Omar Alhadlaq, Xiaolin Li, Peter Danenberg, Dennis Tu, Alex Pine, Vera Filippova, Abhipso Ghosh, Ben Limonchik, Bhargava Urala, Chaitanya Krishna Lanka, Derik Clive, Yi Sun, Edward Li, Hao Wu, Kevin Hongtongsak, Ianna Li, Kalind Thakkar, Kuanysh Omarov, Kushal Majmundar, Michael Alverson, Michael Kucharski, Mohak Patel, Mudit Jain, Maksim Zabelin, Paolo Pelagatti, Rohan Kohli, Saurabh Kumar, Joseph Kim, Swetha Sankar, Vineet Shah, Lakshmi Ramachandruni, Xiangkai Zeng, Ben Bariach, Laura Weidinger, Amar Subramanya, Sissie Hsiao, Demis Hassabis, Koray Kavukcuoglu, Adam Sadovsky, Quoc Le, Trevor Strohman, Yonghui Wu, Slav Petrov, Jeffrey Dean, and Oriol Vinyals. Gemini: A family of highly capable multimodal models, 2024.

Gemma Gemini Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

Gemma-Team. Gemma: Open models based on gemini research and technology, 2024.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzman, and Angela Fan. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *arXiv*, abs/2106.03193, 2021.

Tahmid Hasan, Abhik Bhattacharjee, Md Saiful Islam, Kazi Samin, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. XL-Sum: Large-Scale Multilingual Abstractive Summarization for 44 Languages. pp. 4693–4703, August 2021. doi: 10.48550/arXiv.2106.13822. URL https://aclanthology.org/2021.findings-acl.413.

William Held, Camille Harris, Michael Best, and Diyi Yang. A material lens on coloniality in nlp. *arXiv*, abs/2311.08391, 2023.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2020.

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. Mistral 7b, 2023.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mixtral of experts. *arXiv*, abs/2401.04088, 2024.

Michiel de Jong Yury Zemlyanskiy Federico Lebrón Sumit Sanghai Joshua Ainslie, James Lee-Thorp. Gqa: Training generalized multi-query transformer models from multi-head checkpoints, 2023.

Norman P. Jouppi, George Kurian, Sheng Li, Peter Ma, Rahul Nagarajan, Lifeng Nai, Nishant Patil, Suvinay Subramanian, Andy Swing, Brian Towles, Cliff Young, Xiang Zhou, Zongwei Zhou, and

David Patterson. Tpu v4: An optically reconfigurable supercomputer for machine learning with hardware support for embeddings, 2023.

Khyati Khandelwal, Manuel Tonneau, Andrew M. Bean, Hannah Rose Kirk, and Scott A. Hale. Casteist but not racist? quantifying disparities in large language model bias between india and the west. *ArXiv*, abs/2309.08573, 2023. URL https://api.semanticscholar.org/CorpusID:262013517.

Md Tawkat Islam Khondaker, Abdul Waheed, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. Gptaraeval: A comprehensive evaluation of chatgpt on arabic nlp. *arXiv*, abs/2305.14976, 2023.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, et al. Prometheus: Inducing fine-grained evaluation capability in language models. *arXiv preprint arXiv:2310.08491*, 2023.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Hadas Kotek, Rikker Dockum, and David Q. Sun. Gender bias and stereotypes in large language models. *Proceedings of The ACM Collective Intelligence Conference*, 2023. URL https://api.semanticscholar.org/CorpusID:261276445.

Haonan Li, Fajri Koto, Minghao Wu, Alham Fikri Aji, and Timothy Baldwin. Bactrian-x: Multilingual replicable instruction-following models with low-rank adaptation. *arXiv*, abs/2305.15011, 2023a.

Haoran Li, Yulin Chen, Jinglong Luo, Yan Kang, Xiaojin Zhang, Qi Hu, Chunkit Chan, and Yangqiu Song. Privacy in large language models: Attacks, defenses and future directions. *ArXiv*, abs/2310.10383, 2023b. URL https://api.semanticscholar.org/CorpusID:264145758.

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. Few-shot learning with multilingual language models. *arXiv*, abs/2112.10668, 2021.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V. Le, Barret Zoph, Jason Wei, and Adam Roberts. The flan collection: Designing data and methods for effective instruction tuning. *arXiv*, abs/2301.13688, 2023a.

Shayne Longpre, Robert Mahari, Anthony Chen, Naana Obeng-Marnu, Damien Sileo, William Brannon, Niklas Muennighoff, Nathan Khazam, Jad Kabbara, Kartik Perisetla, et al. The data provenance initiative: A large scale audit of dataset licensing & attribution in ai. *arXiv preprint arXiv:2310.16787*, 2023b.

Nils Lukas, A. Salem, Robert Sim, Shruti Tople, Lukas Wutschitz, and Santiago Zanella-B'eguelin. Analyzing leakage of personally identifiable information in language models. *2023 IEEE Symposium on Security and Privacy (SP)*, pp. 346–363, 2023. URL https://api.semanticscholar.org/CorpusID:256459554.

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. Crosslingual generalization through multitask finetuning. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15991–16111, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.891. URL https://aclanthology.org/2023.acl-long.891.

Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A. Feder Cooper, Daphne Ippolito, Christopher A. Choquette-Choo, Eric Wallace, Florian Tramèr, and Katherine Lee. Scalable extraction of training data from (production) language models. *arXiv*, abs/2311.17035, 2023.

Gabriel Nicholas and Aliya Bhatia. Lost in translation: Large language models in non-english content analysis. *arXiv*, abs/2306.07377, 2023.

NLLB-Team, 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. No language left behind: Scaling human-centered machine translation. 2022.

Jessica Ojo, Kelechi Ogueji, Pontus Stenetorp, and David I. Adelani. How good are large language models on african languages? *arXiv*, abs/2311.07978, 2023.

Jonas Pfeiffer, Naman Goyal, Xi Lin, Xian Li, James Cross, Sebastian Riedel, and Mikel Artetxe. Lifting the curse of multilinguality by pre-training modular transformers. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3479–3495, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.255. URL https://aclanthology.org/2022.naacl-main.255.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. Xcopa: A multilingual dataset for causal commonsense reasoning. pp. 2362–2376, November 2020. doi: 10.18653/v1/2020.emnlp-main.185. URL https://aclanthology.org/2020.emnlp-main.185.

Ofir Press, Noah A. Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. *CoRR*, abs/2108.12409, 2021. URL https://arxiv.org/abs/2108.12409.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*, 2023.

Reva Schwartz, Apostol Vassilev, Kristen Greene, Lori Perine, Andrew Burt, Patrick Hall, et al. Towards a standard for identifying and managing bias in artificial intelligence. *NIST special publication*, 1270(10.6028), 2022.

Noam Shazeer. GLU variants improve transformer. *CoRR*, abs/2002.05202, 2020. URL https://arxiv.org/abs/2002.05202.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=fR3wGCk-IXp.

Shivalika Singh, Freddie Vargus, Daniel Dsouza, Börje F. Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura OMahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Souza Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergün, Ifeoma Okoh, Aisha Alaagib, Oshan Mudannayake, Zaid Alyafeai, Vu Minh Chien, Sebastian Ruder, Surya Guthikonda, Emad A. Alghamdi, Sebastian Gehrmann, Niklas Muennighoff, Max Bartolo, Julia Kreutzer, Ahmet Üstün, Marzieh Fadaee, and Sara Hooker. Aya dataset: An open-access collection for multilingual instruction tuning. *arXiv preprint arXiv:2402.06619*, 2024.

Jianlin Su, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *CoRR*, abs/2104.09864, 2021. URL https://arxiv.org/abs/2104.09864.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. Stanford alpaca: An instruction-following llama model. 2023.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *arXiv*, abs/2302.13971, 2023a.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. *arXiv*, abs/2307.09288, 2023b.

Aniket Vashishtha, Kabir Ahuja, and Sunayana Sitaram. On evaluating and mitigating gender biases in multilingual settings. *arXiv*, abs/2307.01503, 2023.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mt5: A massively multilingual pre-trained text-to-text transformer. pp. 483–498, June 2020. doi: 10.18653/v1/2021.naacl-main.41. URL https://aclanthology.org/2021.naacl-main.41.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2369–2380, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1259. URL https://aclanthology.org/D18-1259.

Zheng-Xin Yong, Cristina Menghini, and Stephen H. Bach. Low-resource languages jailbreak GPT-4. *arXiv*, abs/2310.02446, 2023a.

Zheng Xin Yong, Hailey Schoelkopf, Niklas Muennighoff, Alham Fikri Aji, David Ifeoluwa Adelani, Khalid Almubarak, M Saiful Bari, Lintang Sutawika, Jungo Kasai, Ahmed Baruwa, Genta Winata, Stella Biderman, Edward Raff, Dragomir Radev, and Vassilina Nikoulina. BLOOM+1: Adding language support to BLOOM for zero-shot prompting. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11682–11703, Toronto, Canada, July 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.653. URL https://aclanthology.org/2023.acl-long.653.

Joanna Yoo, Kuba Perlin, Siddhartha Rao Kamalakara, and João G. M. Araújo. Scalable training of language models using jax pjit and tpuv4, 2022.

Jun Zhao, Zhihao Zhang, Luhui Gao, Qi Zhang, Tao Gui, and Xuanjing Huang. Llama beyond english: An empirical study on language capability transfer. *arXiv*, abs/2401.01055, 2024.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. Aya model: An instruction finetuned open-access multilingual language model, 2024.

A Languages in Aya 23 Model Family

Code	Language	Script	Family	Subgrouping	Native speakers
ar	Arabic	Arabic	Afro-Asiatic	Semitic	380 million
cs	Czech	Latin	Indo-European	Balto-Slavic	10.7 million
de	German	Latin	Indo-European	Germanic	95 million
el	Greek	Greek	Indo-European	Graeco-Phrygian	13.5 million
en	English	Latin	Indo-European	Germanic	500 million
es	Spanish	Latin	Indo-European	Italic	500 million
fa	Persian	Arabic	Indo-European	Iranian	72 million
fr	French	Latin	Indo-European	Italic	74 million
he	Hebrew	Hebrew	Afro-Asiatic	Semitic	5 million
hi	Hindi	Devanagari	Indo-European	Indo-Aryan	350 million
id	Indonesian	Latin	Austronesian	Malayo-Polynesian	43 million
it	Italian	Latin	Indo-European	Italic	65 million
jp	Japanese	Japanese	Japonic	Japanesic	120 million
ko	Korean	Hangul	Koreanic	Korean	81 million
nl	Dutch	Latin	Indo-European	Germanic	25 million
pl	Polish	Latin	Indo-European	Balto-Slavic	40 million
pt	Portuguese	Latin	Indo-European	Italic	230 million
ro	Romanian	Latin	Indo-European	Italic	25 million
ru	Russian	Cyrillic	Indo-European	Balto-Slavic	150 million
tr	Turkish	Latin	Turkic	Common Turkic	84 million
uk	Ukrainian	Cyrillic	Indo-European	Balto-Slavic	33 million
vi	Vietnamese	Latin	Austroasiatic	Vietic	85 million
zh	Chinese	Han & Hant	Sino-Tibetan	Sinitic	1.35 billion

Table A1: 23 languages supported in **Aya** 23 models, each language's corresponding script, family, subgrouping, and approximate number of native speakers. The number of native speakers for each language is taken from the Wikipedia page of the respective language accessed on May 22, 2024.