Aya 23: Open Weight Releases to Further Multilingual Progress

Viraat Aryabumi*1, John Dang1, Dwarak Talupuru2, Saurabh Dash1, David Cairuz2, Hangyu Lin2, Bharat Venkitesh2, Madeline Smith1, Kelly Marchisio2, Sebastian Ruder2, Acyr Locatelli2, Julia Kreutzer1, Phil Blunsom2, Marzieh Fadaee1, Ahmet Üstün*1, and Sara Hooker*1

*1Cohere For AI, *2Cohere

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 model1 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
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.

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) \textbf{Outdated knowledge}: Having been pre-trained several years ago, mT5 is not as useful for interactions about events that occurred recently. 2) \textbf{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
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.

[^1]: https://docs.cohere.com/docs/command-r
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.


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.

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

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

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:** \(<\text{BOS_TOKEN}>\text{Hello, how are you?}\text{END_OF_TURN_TOKEN}>\)

**Completion:** \(<\text{START_OF_TURN_TOKEN}>\text{I am doing good!}\text{END_OF_TURN_TOKEN}>\)

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 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 ShareGPT⁵ and Dolly-15k [Conover et al., 2023b]. 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 use⁷ 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

⁵https://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]

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

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

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 $6 \times 10^{-4}$, an end LR of $6 \times 10^{-5}$ 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**.

---

8We 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.

9We 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 ↔ 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\(^\text{10}\) 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.\footnote{https://perspectiveapi.com/}

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...
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.

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.
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.**

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

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.

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

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.
### 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 Mistral-7B while Aya-24-35B wins 78.0%, 84.5% and 75.0% of the time against Mixtral-8x7B.

<table>
<thead>
<tr>
<th>Model</th>
<th>FLORES-200 (spBleu)</th>
<th>XLSum (RougeL)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>X→En</td>
<td>En→X</td>
</tr>
<tr>
<td>Bactrian-X-7B</td>
<td>25.9</td>
<td>16.6</td>
</tr>
<tr>
<td>Mistral-7B-Instruct-v0.2</td>
<td>31.1</td>
<td>21.0</td>
</tr>
<tr>
<td>Gemma-1.1-7B-it</td>
<td>32.0</td>
<td>25.6</td>
</tr>
<tr>
<td>Aya-101-13B</td>
<td>35.9</td>
<td>30.4</td>
</tr>
<tr>
<td>Aya-23-8B</td>
<td><strong>39.5</strong></td>
<td><strong>34.8</strong></td>
</tr>
<tr>
<td>Mixtral-8x7B-Instruct-v0.1</td>
<td>36.3</td>
<td>28.9</td>
</tr>
<tr>
<td>Aya-23-35B</td>
<td><strong>43.0</strong></td>
<td><strong>37.8</strong></td>
</tr>
</tbody>
</table>

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.
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.

### 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.

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

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.
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.

<table>
<thead>
<tr>
<th></th>
<th>Arabic</th>
<th>English</th>
<th>Hindi</th>
<th>Italian</th>
<th>Simplified Chinese</th>
<th>Ukrainian</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aya-101-13B</td>
<td>81.6</td>
<td>83.3</td>
<td>81.7</td>
<td>93.3</td>
<td>75.8</td>
<td>88.3</td>
<td>84.0</td>
</tr>
<tr>
<td>Aya-23-8B</td>
<td>42.5</td>
<td>56.1</td>
<td>51.7</td>
<td>51.7</td>
<td>55.8</td>
<td>53.6</td>
<td>51.9</td>
</tr>
<tr>
<td>Aya-23-35B</td>
<td>11.7</td>
<td>21.7</td>
<td>37.5</td>
<td>40.0</td>
<td>27.5</td>
<td>19.2</td>
<td>26.2</td>
</tr>
</tbody>
</table>

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.
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\textsuperscript{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


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.

\textsuperscript{13}\url{https://cohere.com/research/aya}


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


Liu, Yunxuan Li, Sarah Cogan, Jeremy Chen, Chao Jia, Chenjie Gu, Qiao Zhang, Jordan Grimm-
stad, Ale Jakse Hartman, Xavier Garcia, Thanumalayan Sankaranarayana Pillai, Jacob Devlin, 
Michael Laskin, Diego de Las Casas, Dasha Valter, Connie Tao, Lorenzo Blanco, Adrià Puig-
domènec 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 Stjern-
gren, 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 Samangoei, 
Bogdan Damoc, Alex Kaskasoi, Sébastien M. R. Arnold, Vijay Vasudevan, Shubham Agrawal, 
Jason Riesa, Dmitry Lepikhin, Richard Tanburn, Srivatsan Srinivasan, Hyontaek 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 Choudhery, Dipanjan Das, Dominika Rogozińska, Vitaliy Nikolaev, Pablo Sprech-
mann, Zachary Nado, Lukas Zilla, Flavien Prost, Luheng He, Marianne Monteiro, Gaurav Mishra, 
Chris Welty, Josh Newlan, Dawei Jia, Miltiadis Allamanis, Clara Huwyi Hu, Raoul de Liedekerke, 
Justin Gilmer, Carl Saroufim, Shruti Rijhwani, Shaobo Hou, Disha Shrivastava, Anirudh Bad-
depudi, Alex Goldin, Adnan Ozturel, Albin Cassirer, Yunhan Xu, Daniel Sohn, Devendra Sachan, 
Reinald Kim Amplayo, Craig Swanson, Dessie Petrova, Shashi Narayanan, 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 Sapturo, Kiran Vodrahalli, James Qin, Zeynep Cankara, Abhan-
shu Sharma, Nick Fernando, Will Hawkins, Belnham Neyshabur, Solomon Kim, Adrian Hutter, 
Priyanka Agrawal, Alex Castro-Ros, George van den Driessche, Tao Wang, Fan Yang, Shuo 
yin Chang, Paul Komarek, Ross McIlroy, Mario Lučić, Guodong Zhang, Wael Farhan, Michael 
Sharman, Paul Natsev, Paul Michel, Yamini Bansal, Siyuan Qiao, Kris CaO, Siyam Shakari, 
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 Tre-
bacz, Kevin Robinson, Yash Kataria, Sebastian Riedel, Paige Bailey, Kefan Xiao, Nimesh Ghe-
lani, Lora Aroyo, Ambrose Slone, Neil Houlsby, Xuehan Xiong, Zhen Yang, Elena Gribovskaya, 
Jonas Adler, Mateo Wirth, Lisa Lee, Music Li, Thais Kagoghara, Jay Pavagadhi, Sophie Bridgers, 
Anna Bortsova, Sanjay Ghemawat, Zafarali Ahmed, Tianqi Liu, Richard Powell, Vijay Bolina, 
Mariko Inuma, Polina Zablotskaia, James Besley, Da-Woon Chung, Timothy Dozat, Ramona 
Comanescu, Xiange Si, Jeremy Greer, Guolong Su, Martin Polacek, Raphaël Lopez Kaufman, 
Simon Tokumine, Hexiang Hu, Elena Buchatskaya, Yingjie Miao, Mohamed Elhawaty, Aditya 
Siddhant, Nanad 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 Pajaraskas, 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 Aispos, Léonard Hussenot, Livio Bal-
dini Soares, Kate Baumli, Michael B. Chang, Adrià Recasens, Ben Caine, Alexander Pritzel, 
Filip Pavetic, Fabio Pardo, Anita Gergely, Justin Frye, Vinay Ramesh, Dan Horgan, Kartikeya 
Badola, Nora Kassner, Subhrajit Roy, Ethan Dyer, Víctor Campos Campos, Alex Tomala, Yun-
Ghorbani, Charles Chen, Andy Crawford, Shalini Pal, Mukund Sridhar, Petru Gurita, Asier Mu-

jika, Igor Petrovski, Pierre-Louis Cedoz, Chenmei Li, Shiuyuan Chen, Niccolò Del Santo, Siddharth

Goyal, Jitesh Punjabi, Karthik Kappaganthu, Chester Kwak, Pallavi LV, Sarnishta Velury, Hi-

madri 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

Ädel, Sujeewan Rajayogam, Travis Choma, Fei Liu, Aditya Barua, Colin Ji, Ji Ho Park, Vincent

Hellendoorn, Alex Bailey, Taylan Bilal, Huanjie Zhou, Mehrdad Khatir, Charles Sutton, Woj-

ciech Rzadkowski, Fiona Macintosh, Konstantin Shagin, Paul Medina, Chen Liang, Jinjing Zhou,

Pararth Shah, Yingying Bi, Attila Dankovics, Shipra Banga, Sabine Lehmman, Marissa Bredeisen,

Zifan Lin, John Eric Hoffmann, Jonathan Lai, Raynaold Chung, Kai Yang, Nihal Balani, Arthur

Brażinskias, Andrei Sozanschi, Matthew Hayes, Héctor Fernández Alcalde, Peter Makarov, Will

Chen, Antonio Stella, Liselotte Snijders, Michael Mandl, Ante Kärrman, Pawel Nowak, Xinya

Wu, Alex Dyck, Krishnan Vaidyanathan, Raghaevan, Jessica Mallet, Mitch Rudominer, Eric

Johnston, Sushil Mittal, Akhil Udathu, Janara Christensen, Vishal Verma, Zach Irving, Andreas

Santucci, Gamaleldin Elsaied, Elnau 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 Rajasheker, Lara Tumeh, Eyal Ben-David, Rishub Jain, Jonathan

Usato, Romina Datta, Oskar Bunyan, Shimi 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 Vieil-

lard, Jane Park, Jiageng Zhang, Jeff Stanway, Drew Garmon, Abhijit Karmarkar, Zhe Dong,

Jong Lee, Aviral Kumar, Luwei Zhou, Jonathan Evans, William Isaac, Geoffrey Irving, Edward

Loper, Michael Fink, Isha Arkatkar, Nanxin Chen, Izhak Shafran, Ivan Petryanenko, Zhe

Chen, Johnson Jia, Anselm Levskaya, Zhenkai Zhu, Peter Grabowski, Yu Mao, Alberto Magni,

Kaisheng Yao, Javier Snaider, Norman Casagrande, Evan Palmer, Paul Sugarman, Alfonso Cas-

ťano, Irene Giannoumis, Wooyeon Kim, Mikołaj Rybiński, Ashwin Sreevatsa, Jennifer Prendki,

David Soregel, Adrian Goedemeyn, Willi Gierke, Mohsen Safari, Meenu Gaba, Jeremy Wies-

ner, 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, Rildo Aguilar, Keith Pallo, Abhishek Chakladar, Ginger Perng, Elena Allica

Abellan, Mingyang Zhang, Ishita Dasgupta, Kate Kushman, Ivo Penchev, Alena Repina, Xiuhui

Wu, Tom van der Weide, Priya Ponnappalli, Caroline Kaplan, Jiri Simsa, Shuangfeng Li, Olivier

Dousse, Fan Yang, Jeff Piper, Nathan Le, Rama Pasumarthi, Nathan Lintz, Anitha Vijayakumar,

Daniel Andor, Pedro Valenzuela, Minnie Lu, 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 Singh, Dayou Du, Dan

McKinnon, Natasa Antropova, Tolga Bolukbasi, Orgad Keller, David Reid, Daniel Finchselstein,

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 Albert, Chu-Cheng Lin, Colin Evans, Alex 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, Meghanah Thotakuri, Tom Natan, Matthieu Geist, Ser tan Girgin, Hui Li,

Jiayu Ye, Oifer Roval, Reiko Tojo, Michael Kwong, James Lee-Thorp, Christopher Yew, Danila

Sinopahnikov, Sabela Ramos, John Mellor, Abhishek Sharma, Kathy Wu, David Miller, Nicolas

Sonnerat, Denis Vmukov, Rory Greig, Jennifer Beattie, Emily Caveness, Libin Bai, Julian Eisen-
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 Ramachandrune, Xiangkai Zeng, Ben Bariach, Laura Weidinger, Omar Subramanya, Sissie Hsiao, Demis Hassabis, Koray Kavukcuoglu, Adam Sadowsky, Quoc Le, Trevor Strohman, Yonghui Wu, Slav Petrov, Jeffrey Dean, and Oriol Vinyals. Gemini: A family of highly capable multimodal models, 2024.


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


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


## Languages in Aya 23 Model Family

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

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.