Reka Core, Flash, and Edge: A Series of Powerful Multimodal Language Models

Aitor Ormazabal **Che Zheng** **Cyprien de Masson d'Autume** **Dani Yogatama**

Deyu Fu **Donovan Ong** **Eric Chen** **Eugenie Lamprecht** **Hai Pham** **Isaac Ong**

Kaloyan Aleksiev **Lei Li** **Matthew Henderson** **Max Bain** **Mikel Artetxe**

Nishant Relan **Piotr Padlewski** **Qi Liu** **Ren Chen** **Samuel Phua**

Yazheng Yang **Yi Tay** **Yuqi Wang** **Zhongkai Zhu** **Zhihui Xie**

Abstract

We introduce Reka Core, Flash, and Edge, a series of powerful multimodal language models trained from scratch by Reka.[1] Reka models are able to process and reason with text, images, video, and audio inputs. This technical report discusses details of training some of these models and provides comprehensive evaluation results. We show that **Reka Edge** and **Reka Flash** are not only state-of-the-art but also outperform many much larger models, delivering outsized values for their respective compute class. Meanwhile, our most capable and largest model, **Reka Core**, approaches the best frontier models (OpenAI, 2023; Google, 2023; Anthropic, 2024) on both automatic evaluations and blind human evaluations. On image question answering benchmarks (e.g., MMMU, VQAv2), **Core** performs competitively to GPT4-V. Meanwhile, on multimodal chat, **Core** ranks as the second most preferred model under a blind third-party human evaluation setup, outperforming other models such as Claude 3 Opus. On text benchmarks, **Core** not only performs competitively to other frontier models on a set of well-established benchmarks (e.g., MMLU, GSM8K) but also outperforms GPT4-0613 on human evaluation. On video question answering (Perception-Test), **Core** outperforms Gemini Ultra. Models are shipped in production at `chat.reka.ai`. A showcase of *non cherry picked* qualitative examples can also be found at showcase.reka.ai.

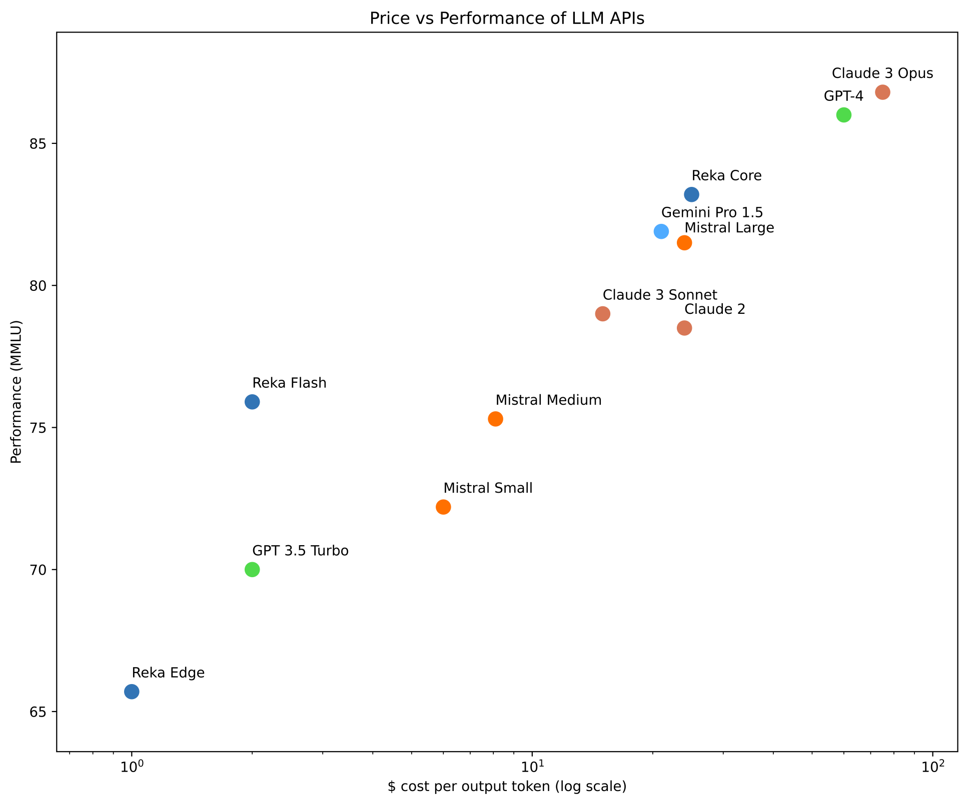
[1]Please cite this report as authored by Reka team.

1 Introduction

This technical report details comprehensive evaluations of the Reka models (Core, Flash, Edge) on language and vision tasks along with discussions on development, benchmark design, and the training pipeline.

Reka **Edge** and **Flash** are dense models with 7B and 21B parameters, respectively. Our evaluation shows that these models are state-of-the-art for their compute class, often surpassing models much larger. Meanwhile, the current version of Reka **Core** approaches many of the best frontier models (OpenAI, 2023; Google, 2023; Google et al., 2023; Anthropic, 2024). It excels in both automated base model evaluations and blind third-party human evaluations. Figure 1 compares Reka models against proprietary large language models (LLM) APIs. We plot the price against performance, using MMLU score as an approximate indicator of model quality. All Reka models are positioned either on or beyond the Pareto frontier.

Figure 1: Price per performance (MMLU score) of different LLM APIs.



Reka Core approaches the performance levels of GPT-4V (OpenAI, 2024) on MMMU (Yue et al., 2024), VQAv2, and third-party multimodal chat evaluation. Meanwhile, Reka Core surpasses all Claude 3 models (Opus, Sonnet, Haiku) (Anthropic, 2024) on multimodal chat human evaluation. On video question answering (Perception-test (Pătrăucean et al., 2023)), both Reka Flash and Core outperform Gemini Ultra (Google, 2023). On language benchmarks, Reka Core achieves 83.2 MMLU score and competitive GSM8K, HumanEval, and GPQA scores compared to other frontier models. On text-only chat, blind human evaluation shows that Reka Core outperforms GPT-4 (0613) and ranks third on our internal ELO leaderboard (right after GPT-4 Turbo and Claude 3 Opus).

Meanwhile, our **Edge** (7B) model surpasses the current state-of-the-art models of this compute class, outperforming both Gemma 7B (Gemma et al., 2024) and Mistral 7B (Jiang et al., 2023). Additionally, the **Flash** (21B) model, aside from outperforming GPT-3.5 Turbo, also outperforms much larger state-of-the-art models such as Grok-1 (xAI, 2023), Mistral Medium (Touvron et al., 2023) and Gemini Pro 1.0 (Google, 2023). On multimodal evaluations, **Flash** outperforms both Claude 3 Opus and Sonnet (Anthropic, 2024) on multimodal chat and matches the Sonnet model on MMMU (Yue et al., 2024). All in all, the **Edge** & **Flash** models are extremely powerful models on a compute-class basis.

In addition to comprehensive evaluations and benchmark evaluations on both language and vision (video + image) tasks, this report also shares some interesting technical details and behind-the-scenes of training large multimodal models as a startup. Areas discussed include infrastructure, data pipeline, compute, annotation pipelines, and more. Finally, artifacts of our models (playground/chat, developer platform) can be found in the following resource table (Table 1).

Table 1: Resource tree of Reka artifacts.

What	Where?
Playground (chat app)	chat.reka.ai
Qualitative Examples (static, non-cherry picked)	showcase.reka.ai
API platform (sign up, manage credits)	platform.reka.ai
Discord (questions)	discord
Homepage	reka.ai

2 Model

This section briefly describes the technical details behind these models.

2.1 Training Data

The training data comprises a mixture of publicly available and proprietary/licensed datasets with a dataset knowledge cutoff of November 2023. The dataset ingested by our model comprises of text, images, videos, and audio clips. Reka Flash and Reka Edge were trained on approximately 5 trillion and 4.5 trillion extensively deduplicated and filtered language tokens, respectively. While the classification of corpora is not strictly defined to one class or category, approximately 25% of our pretraining data is code related, and 30% are STEM related. Approximately 25% of the data is web crawl. About 10% of our data has some relation to math. Overall mixture rates generally follow a principle of prioritizing unique tokens but are hand-adjusted using signal from a limited number of small scale ablations.

Table 2: Statistics of Reka suite of multimodal language models. **Note: Reka Core has not finished training and is still improving.**

Model	Model Size	Text tokens	Context	Long-context	Knowldge Cutoff
Edge	7B dense	4.5T	8K	64K	Nov 2023
Flash	21B dense	5T	8K	128K	Nov 2023
Core	-	-	8K	128K	Nov 2023

Multilingual Data: Approximately 15% of our pretraining data is explicitly (and deliberately) multilingual, comprising 32 diverse languages tier-weighted (roughly by frequency in the wild). Beyond these explicitly up-weighted languages, we also train on the entire multilingual Wikipedia comprising of 110 languages so we expect a baseline performance for most languages. It is worth noting that these tiers reflect pretraining capability and not necessarily downstream post-training induced capabilities of the final model. To be concrete, these are meaningful to estimate the potential of a particular language, given suitable supervised fine tuning data. Languages included during pretraining are shown below.

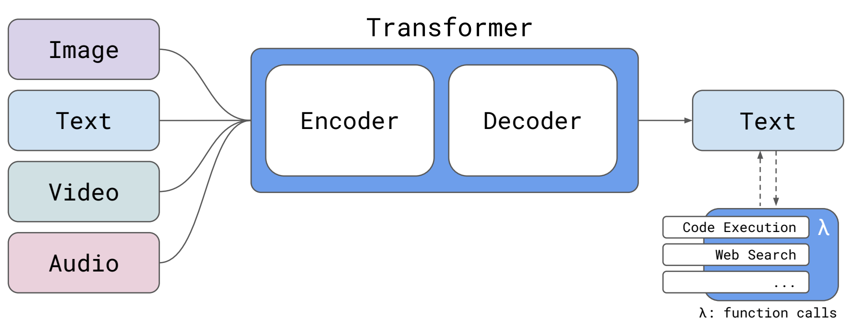
Table 3: Tiered languages in pretraining mixture.

Pretraining Tier	Languages
P1 languages	German, Chinese, Japanese, French, Korean, Spanish, Italian, Arabic, Hindi
P2 languages	Indonesian, Vietnamese, Thai, Czech, Dutch, Finnish, Bulgarian, Portuguese, Tamil, Persian, Greek, Russian
Additional languages	Turkish, Telugu, Burmese, Swahili, Urdu, Estonian, Malay, Basque, Swedish, Norwegian

Multimodal Data: The multimodal training data comprises large collections of images, videos, documents, and webpages. The chosen data mixture is carefully optimized for quality, diversity, and scale.

2.2 Architecture & Modeling

Figure 2: **Architectural overview for Reka Core, Flash & Edge models:** a modular encoder-decoder transformer supporting multimodal input (image, text, video & audio). The text output can invoke function calls, such as web search and code execution, then return the results.



This section introduces training details, model architecture, and context length details.

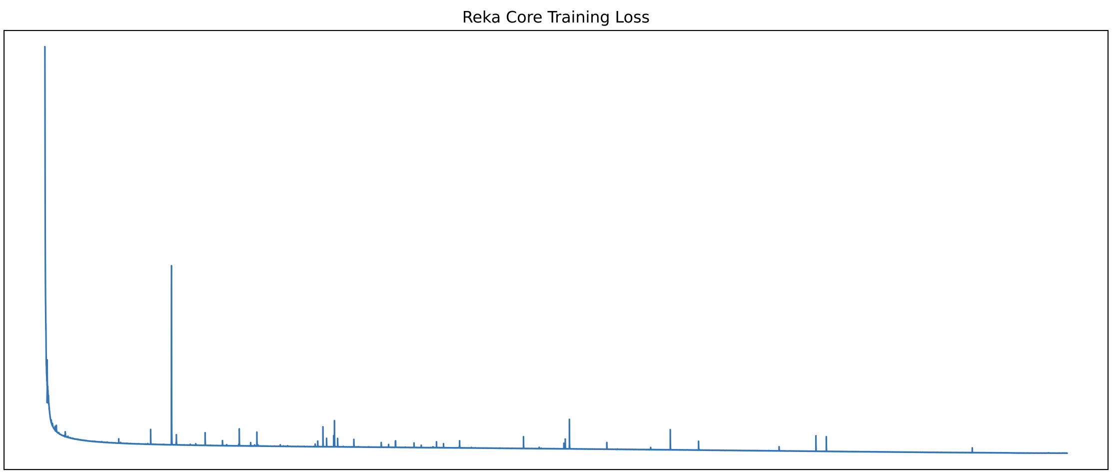
Architecture & Training. Our overall architecture (Figure 2) is a modular encoder-decoder architecture supporting text, image, video, and audio inputs. For now, our model only supports text outputs. The backbone Transformer model is based on the 'Noam' architecture, i.e., it uses SwiGLU (Shazeer, 2020), Grouped Query Attention (Ainslie et al., 2023; Shazeer, 2019), Rotary positional embeddings (Su et al., 2021) and RMSNorm (Zhang and Sennrich, 2019). Architecturally, this is similar to the PaLM architecture (Chowdhery et al., 2022) but without parallel layers. Reka Flash and Edge uses a sentencepiece vocab of 100K based on *tiktoken* (e.g., GPT-4 tokenizer). We add sentinel tokens for masking spans, i.e., <EXTRA_ID_0> and other special use cases such as tool-use that are beyond the scope of this technical report. Pretraining uses a curriculum that goes through multiple stages with different mixture distributions, context lengths, and objectives. The current version of this model is a dense model. Models are trained with BFLOAT16.

Context Length. Our standard models have a context length of 8K for our regular models. Reka Flash and Reka Core have 128K for long context models for retrieval and long document tasks. All our models pass needle-in-the-haystack (passkey retrieval) for the context they support. Based on these tests, our 128K models seem to extrapolate to 256K context length (but not beyond). For long context training, in addition to instruction tuning data we collect, we synthetically create supervised fine tuning data using our own suite of models by conditioning on long documents found in pretraining corpus using a technique we call *reverse instruction tuning from long documents*.

2.3 Compute & Infrastructure

Our family of Reka models was trained predominantly on Nvidia H100s using Pytorch (Paszke et al., 2019). Our setup comprises of clusters from a mixture of vendors with our peak compute being approximately 2.5K H100s and 2.5K A100s. Our peak number of clusters is 6. About more than 90% of our compute came online in mid-December 2023. Reka Flash and Edge were trained on several hundreds of H100s across a period of several weeks. Our pretraining process was relatively smooth with very few loss spikes despite very aggressive learning rates[2] even for much larger models. Figure 3 shows the training loss for Reka Core. To improve the I/O of our clusters, especially for scalable training with multimodal inputs, we used the Ceph filesystem for distributed and scalable data storage across nodes which improved I/O substantially but came with maintenance overheads.

Figure 3: Training loss for Reka Core.



2.3.1 Hardware lottery and node stability.

Generally, we find great unreliability when it comes to GPU nodes which often fail due to hardware errors or connection issues. Moreover, reliability among providers is generally of high variance. For more details, refer to Tay (2024). To expand upon Tay (2024), we report the average number of node failures across four anonymized providers, as shown in Table 4. Since the likelihood of node failures is influenced by the number of nodes concurrently used for training, we report estimated failure rates for different configurations.

Chaotic and Stable phases Aside from variances across clusters and providers, providers could also have high variance across time periods. For example, many compute providers have clusters that behave very differently in the first few weeks of handover or whenever the cluster undergoes a big change. Hence, we also compare the node failure rates during both the early phase and stable phase. More often than not, aside from early phase of handing over a cluster, provisioning new nodes can also introduce a new chaotic era that can last a few days or weeks. In general, we determined that a key factor influencing the difference between the early and stabilized phases is whether the cluster was actively used for distributed training by previous customers.

[2] Models trained at the edge of stability turn out stronger. See https://x.com/m__dehghani/status/1686056450081337344.

Table 4: Average number of node failures (on a weekly basis) across four anonymized compute providers. Since node failures depend on the number of nodes used concurrently, we report estimated failure rates for different configurations. Many compute providers have clusters that behave very differently in the first few weeks of handover. Hence, we also report the difference in node failure rate in both the early phase and stable phase. Chips refer to either H100s or A100s.

Provider	Number of chips used	Number of node failures (per week)
Provider A	2000 chips	3
Provider A (early phase)	2000 chips	20+
Provider B	300 chips	0.2
Provider B (early phase)	300 chips	0.2
Provider C (stable phase)	300 chips	3
Provider C (stable phase)	100 chips	3
Provider C (early phase)	100 chips	30+
Provider D	300 chips	2

Inference and serving. We built a custom inference stack for text and multi-modality running on a combination of A10s and A100s. We use Kubernetes as the underlying orchestration engine and manage several large clusters across different regions.

2.4 Post-Training

This section describes the post-training process which involves aligning, instruction tuning the model.

SFT and RLHF. After pretraining, our models are then instruction tuned (Wei et al., 2021; Ouyang et al., 2022; Chung et al., 2024) for multiple epochs using strong regularization. As for SFT data, we train on a mixture of datasets that include our proprietary and publicly available data. After SFT, models are then aligned with RLHF, specifically PPO (Schulman et al., 2017), using the same family of Reka models as the reward model. Our models go through a couple of rounds of RLHF in total. Moreover, our post-training process considers tool-use, function calling and web search, which is out of scope for this technical report.

Annotation Pipelines for Data Collection and Human Evaluation. We collect data using external data collection companies and provide them with a user interface for annotating both text-only and multimodal data. We create an annotation UI for both collecting data and/or sending examples to human raters for blind human evaluation. This software also supports annotating for individual pointwise quality and also side-by-side (pairwise) evaluations. Our annotation software supports images, videos, and text-only prompts and responses. It also supports the annotation of multi-turn dialogues.

3 Evaluation

This section discusses the results of extensive evaluations of Reka models.

3.1 Base Model Evaluation

We conduct a series of language-only and multimodal (image, video input) evaluations.

Language Model Evaluation. We compare our models on four language model evaluations: 1) MMLU (general language understanding and question answering) (Hendrycks et al., 2021), 2) GSM8K (reasoning and arithmetic) (Cobbe et al., 2021), HumanEval (code generation) (Chen et al., 2021) and GPQA (graduate-level question answering) (Rein et al., 2023). All numbers from baselines are reported numbers in other works. MMLU is evaluated with 5-shot direct prompting for all models. For GSM8K, most models use 8-shot chain-of-thought (Wei et al., 2022) and majority voting (maj@8). For HumanEval, this is evaluated in 0-shot setup. All results from other models are reported from other works.

Multimodal (Image/Video) Evaluation. We compare our models using visual question answering datasets, i.e., MMMU (Yue et al., 2024), VQAv2 (Goyal et al., 2017), and Perception-Test (Pătrăucean et al., 2023) for video question answering. For Reka models, all results are 0-shot.

Table 5: Comparisons of our Reka Flash and Reka Core against other frontier models. Dashes (−) refer to either model not supporting modality or unavailable benchmark scores.

Model / Eval	Reka Core v0.5	Reka Flash v1.5	GPT-4	Claude 3 Opus	Claude 3 Sonnet	Gemini Ultra	Gemini Pro 1.5
MMLU (*Knowledge*)	83.2	75.9	86.4	86.8	79.0	83.7	81.9
GSM8K (*Reasoning*)	92.2	85.8	92.0	95.0	92.3	94.4	91.7
HumanEval (*Coding*)	76.8	72.0	76.5	84.9	73.0	74.4	71.9
GPQA (*main*) (*Hard QA*)	38.2	34.0	38.1	50.2	39.1	35.7	41.5
MMMU (*Image QA*)	56.3	53.3	56.8	59.1	53.1	59.4	58.5
VQAv2 (*Image QA*)	78.1	78.4	77.2	−	−	77.8	73.2
Perception-test (*Video QA*)	59.3	56.4	−	−	−	54.7	51.1[3]

Results. Table 5 reports comparisons of Reka Core against other frontier-class models. Overall, Reka Core performs competitively with other frontier-class models. On most metrics (with the exception of MMLU), it is comparable to GPT-4[4]. In terms of overall performance and with respect to the Claude 3 series, it falls somewhere in between Opus and Sonnet. When compared to Gemini models, Reka Core has mixed outcomes, i.e., winning some and losing some. Reka Core outperforms Gemini Pro 1.5 on several benchmarks (MMLU, GSM8K, HumanEval) but is outperformed on GPQA and MMMU. Notably, Reka Core and Flash outperform Gemini Ultra (and Pro 1.5) on video question answering. Reka Core is still improving so we expect better results in the near future.

3.2 Chat Model Evaluation

We conduct a blind evaluation with human raters from a third party data provider company. We consider two setups: 1) multimodal chat, where the user asks a question about an image, and 2) text-only chat. We

[3]We report Pro 1.0 performance here since Pro 1.5 did not report perception-test.

[4]At least an older version, with the results mostly reported from the recent Claude 3 release. HumanEval looks too low for the Claude 3 release so we referenced the HumanEval leaderboard for this number.

next detail our evaluation protocol and present results for each setting.

3.2.1 Evaluation Setup

For each annotation instance, human raters are given a prompt along with a maximum of 4 generations from different models, and asked to rate the answers according to the guidelines provided. Given that the number of models in our evaluation is higher than 4, we collect multiple such annotations for each prompt, each with a different subset of models. The pairing of models is decided randomly for each prompt, with all combinations being equally likely. We compute ELO scores following Askell et al. (2021), where we only consider pairwise comparisons where annotators express a preference stronger than the weakest available.

We design our evaluation dataset to cover a diverse set of prompts. The following table details the composition of our text-only evaluation set, which comprises 1K+ prompts:

Table 6: Taxonomy of prompts in our text-only human evaluation dataset. The dataset is balanced across subcategories.

Category	**Subcategory**
Knowledge-intensive	Humanities and social sciences Natural sciences Engineering and technology Entertainment Other
Creative writing	Role playing Brainstorming Poetry Literary prose Non-literary prose
Input-based	Data processing Reading comprehension Classification Extraction Summarization Rewriting Translation
Reasoning	Maths Commonsense and logical reasoning Instruction following
Coding	N/A

Similarly, the following table reports the categories covered by our multimodal evaluation set:

Table 7: Distribution of prompts in our multimodal human evaluation dataset.

Category	**Ratio**
Basic image description	23.0%
Advanced image description	20.5%
Coding capability with vision	7.7%
Multilingual multimodal understanding	7.9%
Multimodal knowledge and commonsense	7.7%
Scene and document reasoning	13.0%
Visual referring prompting	5.1%
Creative tasks	2.6%
Other	12.5%

3.2.2 Multimodal Chat Evaluation

We next report the results of our multimodal chat evaluation in comparsion with GPT4-V, Claude 3, Gemini Pro, IDEFICS 80B, Adept Fuyu 8B, and the strongest Llava 1.6B model:

Table 8: ELO scores of all models on our multimodal human evaluation.

Model	ELO	Win rate
GPT-4V	1201	79.4
Reka Core	1130	72.2
Reka Flash	1082	66.8
Claude 3 Opus	1073	66.2
Claude 3 Sonnet	1069	64.1
Llava 1.6 34B	1022	55.9
Gemini Pro	1011	54.2
Reka Edge	986	50.5
IDEFICS 80B	732	18.8
Adept Fuyu 8B	550	6.4

We find that Reka Core outperforms all models except GPT4-V by a substantial margin. Reka Flash ranks next, performing marginally better than Claude 3 Opus. Reka Edge outperforms IDEFICS 80B and Adept Fuyu 8B by a large margin, approaching the performance of Gemini Pro and the largest Llava 1.6 model.

3.2.3 Text-only Chat Evaluation

We compare our models against different versions of GPT, Claude 3, Llama 2 Chat, and Gemini Pro (API version), and report our results next:

Table 9: ELO scores of all models on our text-only human evaluation.

Model	ELO	Win rate
GPT-4 Turbo (1106-preview)	1227	78.6
Claude 3 Opus	1185	73.6
Reka Core	1091	60.6
Claude 3 Sonnet	1074	59.0
GPT-4 (0613)	1062	57.0
Reka Flash	1020	49.1
GPT-3.5 Turbo (0613)	1012	48.9
Llama 2 Chat 70B	984	43.0
Gemini Pro	950	38.3
Reka Edge	903	31.5
Llama 2 Chat 7B	850	24.3

We find that Reka Core ranks competitively on our ELO leaderboard, outperforming Claude 3 Sonnet and GPT-4, and it is only surpassed by GPT-4 Turbo and Claude 3 Opus. Reka Flash obtains strong results for its size, beating GPT-3.5 Turbo, Gemini Pro and the much larger Llama 2 Chat 70B.

3.2.4 Model development and automatic evaluation using Reka Core

We leverage the frontier-class capabilities of Reka Core for model selection and development and show an example of how we use it for multimodal chat. We ask Reka Core to simulate human judgement by rating a response with respect to a prompt and a reference answer. In short, $f(\text{prompt}, \text{model\_output}, \text{reference\_answer}) \in [1, 100]$. We find that Reka Core rankings across models correlate to human judgement despite the gap between pointwise and pairwise (arena style) evaluations. Our general workflow is that we perform lightweight and simple *pointwise* evaluations for continuous sanity checks before sending our models for third party blind human evaluations.

Figure 4: Results using Reka Core as an evaluator. Reka Core evaluator scores align almost perfectly with the final ELO scores we obtain from human raters.

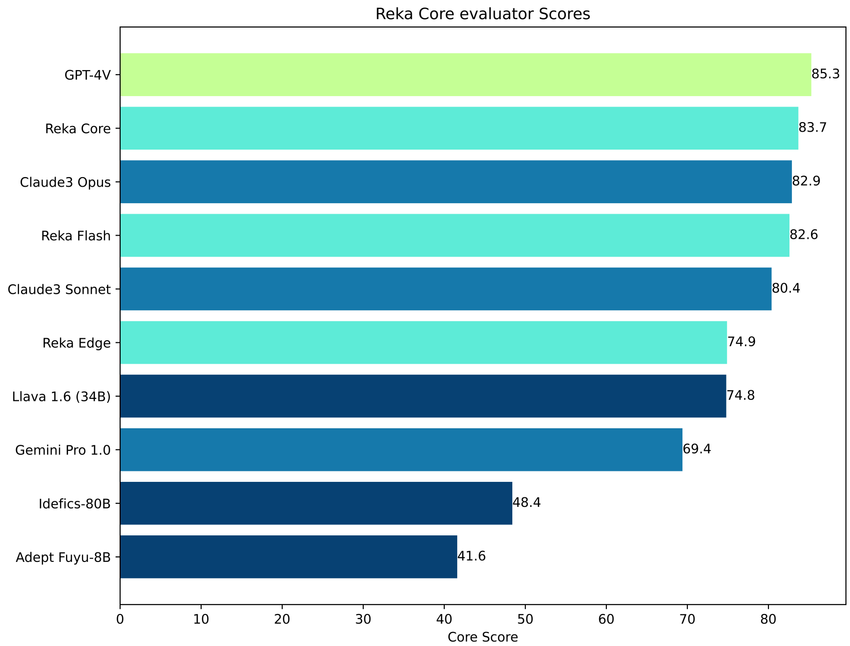


Figure 4 reports the Reka Core scores we obtain right before producing Table 8. Despite Reka Core evaluations being pointwise, we find that it is able to accurately approximate the final rankings. Here, the only key difference is that Reka Flash and Claude Opus have flipped rankings. In practice, these models may be very similar in performance that it could go either way. In Table, 8, we also note that Reka Flash and Claude Opus have very similar win rates and ELO scores, which is well reflected by their Reka Core scores being very close as well. Overall, we find that Reka Core is quite a good approximator of final human evaluation outcomes.

3.3 Cross-lingual Evaluations

We conduct experiments on a suite of general multilingual benchmarks such as multilingual commonsense (XStoryCloze (Lin et al., 2022)), causal reasoning (XCOPA (Ponti et al., 2020)), question answering (Belebele (Bandarkar et al., 2023), XQuAD (Artetxe et al., 2019), TydiQA (Clark et al., 2020)). For all datasets, we report the mean across all languages. We compare our models with Llama 2 70B (Touvron et al., 2023), GPT-3.5 and GPT-4. All evaluations are zero-shot generative except XStoryCloze which uses log-likehood evaluation.

Table 10: Statistics of multilingual datasets.

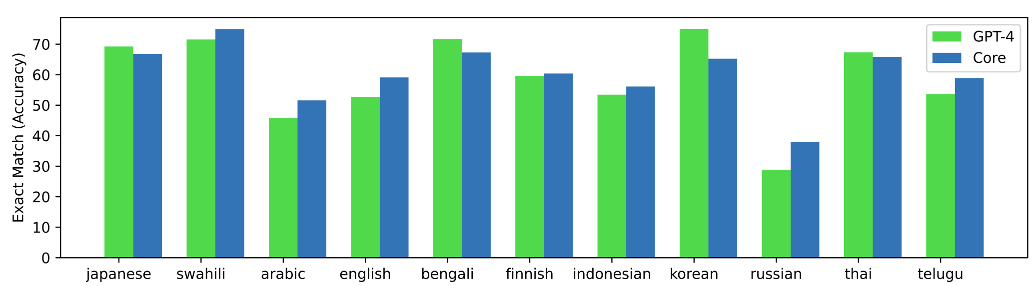
Eval	Languages	Num Langs
XStoryCloze	hi, te, en, zh, ru, my, sw, es, id, eu, ar	11
XCOPA	sw, th, tr, et, vi, qu, id, zh, it, ta, ht	11
XQuAD	ar, de, el, en, es, hi, ro, ru, th,tr, vi, zh	12
XWinograd	fr, en, jp, pt, zh, ru	6
TydiQA	ar, bg, en, fi, id, jp, ko, ru, sw, te, th	11
Belebele	*too many*	150

Table 11: Comparisons of our models on multilingual tasks against GPT-3.5 and GPT-4. All tasks are zero-shot.

Eval / Model	Metric	Reka Core v0.5	Reka Flash v1.5	Llama-2 70B	GPT-3.5	GPT-4
XStoryCloze	acc	72.0	70.1	63.2	N/A	N/A
XCOPA	acc	88.3	68.0	50.6	72.2	86.3
XQuAD	EM	65.7	61.4	25.5	34.6	44.2
XWinograd	acc	86.8	84.0	65.3	72.2	91.5
TydiQA (w context)	EM	60.4	64.8	34.9	53.1	58.9
TydiQA (w/o context)	EM	17.4	15.7	3.9	13.5	21.1
Belebele (all langs)	acc	63.4	57.3	48.0	51.1	N/A

Table 11 reports our evals[5] on multilingual benchmarks. Generally we find that Reka Core outperforms all baselines reliably on most tasks (except GPT-4 where it is mixed). Specifically, Reka Core outperforms GPT-4 on XCOPA, XQuAD, TydiQA but is outperformed on XWinograd and TydiQA (w/o context). Meanwhile, Core outperforms Flash on all benchmarks. Both Flash and Core outperforms Llama-2 70B and GPT-3.5. Finally, Figure 5 shows the language breakdowns of Core vs GPT-4.

Figure 5: Comparison of Reka Core vs GPT-4. Breakdown of languages on 0-shot TydiQA (with context).



[5]We do not run evals for GPT models on XStoryCloze because we use logprobs. As for Belebele, we hit our credit threshold just evaluating on this large evaluation dataset so we stopped.

3.4 Long Context Question Answering

We conduct a series of evaluations on long context question answering. We use internal benchmarks in two domains: (1) movie plots and (2) ToS (terms-of-service) contract with contexts in the ballpark of 100K tokens. Both datasets are question answering tasks where the task is to answer questions given a long document. We compare with Claude 3 (Haiku and Sonnet).

Table 12: Long context question answering evaluation results.

Model	Reka Core	Reka Flash	Claude 3 Haiku	Claude 3 Sonnet
Movie Plots	83.6	79.7	76.6	82.2
ToS Contract	87.5	90.0	85.0	90.0

Table 12 reports results on long context question answering using internal evaluation datasets. Overall we show that Flash and Core are both competitive to the latest Claude 3 models.

3.5 Medical Reasoning

We compare our Reka models against state-of-the-art domain-specific medical models such as Meditron (Chen et al., 2023) and Med-PaLM-2 (Singhal et al., 2023). We also compare with GPT-4 reported from (Singhal et al., 2023). We compare on three benchmarks: MedMCQA, PubMedQA and MMLU (Medical). MMLU medical is a macro-average over *clinical knowledge*, *medical genetics*, *anatomy*, *professional medicine*, *college biology* and *college medicine*.

Table 13: Results on medical reasoning tasks compared to domain specialized models and frontier models.

Benchmark / Model	Reka			Meditron		Med-PaLM-2	GPT-4
	Edge (7B)	Flash (21B)	Core	7B	70B		
MedMCQA	52.6	71.3	80.6	28.7	52.0	71.3	72.4
PubMedQA	71.6	69.0	74.6	69.3	79.8	79.2	80.4
MMLU (Medical)	65.7	79.5	88.3	54.2	72.7	87.8	90.3
Avg	63.3	73.2	81.3	50.7	68.2	79.4	81.0

Table 13 reports results on medical tasks. Meditron and Med-PaLM-2 are specialized models for medicine. Our results show that Reka Core is competitive with some of the best frontier models and specialized models in medicine. Firstly, Reka Flash and Core outperforms the Meditron series. Secondly, Reka Core outperforms both Med-PaLM-2 and GPT-4 on MedMCQA. However, it is outperformed on PubMedQA. Finally, on MMLU (Medical), Reka Core outperforms Med-PaLM-2 and is slightly behind GPT-4. Overall, on average, Reka Core outperforms Med-PaLM-2 and is approximately similar to GPT-4 on medical tasks.

3.6 Detailed comparisons of Edge and Flash

We report detailed results of Reka Edge and Flash against other models of similar compute class. Notably, both Edge and Flash have been improved quite substantially since the initial release in Feb. Hence, numbers have been upgraded since their first appearances.

3.6.1 Reka Edge results

We report results of Reka Edge against other 7B models such as Llama 2 (Touvron et al., 2023), Mistral (Jiang et al., 2023) and Gemma (Gemma et al., 2024).

Table 14: Results comparing Reka Edge with other leading 7B models in the industry. Most benchmarks are reported from other works with the exception of those denoted with †. For multilingual benchmarks, we run them ourselves.

Benchmark	metric	Llama 2 7B	Mistral 7B	Gemma 7B	Reka Edge
MMLU	5-shot	45.3	62.5	64.3	65.7
GSM8K	maj@1	14.6	35.4	46.4	66.2
MATH	4-shot	2.5	12.7	24.3	23.2
HumanEval	0-shot (pass@1)	12.8	26.2	32.3	54.3
XQuAD†	0-shot	16.6	29.7	21.7	54.2
TydiQA†	0-shot	16.4	31.7	35.8	61.5
TydiQA† (w/o context)	0-shot	2.8	5.0	4.7	6.9
Belebele†	0-shot	27.7	32.8	26.8	37.1

Table 14 reports results of Reka Edge against other 7B models (Gemma, Mistral, Llama). We observe that Reka Edge has an edge against all other models (no pun intended). It outperforms Mistral 7B and Llama 7B on all **8** benchmarks. As for Gemma, it outperforms Gemma for all benchmarks except MATH. Overall, Reka Edge is a super strong model at 7B scale.

3.6.2 Reka Flash results

Given that there are not many good models around the same compute class as Reka Flash, we compare Reka Flash with models that are much larger. Specifically, Llama 2 70B (Touvron et al., 2023), Gemini Pro 1.0 (Google, 2023), Mistral Medium (Touvron et al., 2023) and Grok 1.5 (xAI, 2023).

Table 15: Results comparing Reka Flash with other much larger models.

Benchmark	metric	Llama 2 70B	Gemini Pro 1.0	Mistral Medium	Grok 1.5	Reka Flash
MMLU	5-shot	68.9	71.8	75.3	73.1	75.9
GSM8K	maj@8	56.8	86.5	–	56.8	85.8
MATH	4-shot	13.5	32.6	–	23.9	29.6
HumanEval	0-shot	29.9	67.7	38.4	32.3	72.0
MMMU (vision)	0-shot	N/A	47.9	N/A	53.6	53.3
VQAv2	0-shot	N/A	77.2	N/A	N/A	78.4
Perception-test	0-shot	N/A	51.1	N/A	N/A	56.4

Table 15 reports results of Flash (21B) against other models of larger compute class. All competitors are approximately around 70B parameters with the exception of Grok 1.5 which is a sparse model with 30B activated parameters. We see that Flash outperforms (or is competitive to) all competitors on most benchmarks despite being much smaller.

4 Conclusion

We introduce a new series of powerful multimodal models, namely Reka **Core, Flash, Edge**. Reka **Flash** and **Edge** sets a new state-of-the-art on a compute-class basis, often delivering massive outsized value for their scale. Our **Core** model approaches frontier-class models on both human evaluation and automatic benchmarks. Reka **Core** is still improving so we expect to see even more improvements in the medium term. The field of large language models (Radford et al., 2018; Brown et al., 2020; Devlin et al., 2018; Raffel et al., 2019; Chowdhery et al., 2022; Hoffmann et al., 2022) is still nascent but moving very quickly. With that comes the trade-off of significant amount of noise in the landscape. We hope this technical report shows the rigor of what it takes to build frontier-class models *from scratch* given limited resources.

5 Appendix

5.1 MMMU breakdown

In Table 16, we report our category-level scores in MMMU (Yue et al., 2024) for Reka Core.

Table 16: Breakdown of categories from MMMU benchmark (Yue et al., 2024).

Category	Score
Art	86.7
Art Theory	83.3
Design	86.7
Music	46.7
Accounting	46.7
Economics	56.7
Finance	43.3
Manage	40.0
Marketing	50.0
Biology	56.7
Chemistry	46.7
Math	46.7
Physics	36.7
Basic Medical Science	56.7
Clinical Medicine	60.0
Diagnostics and Laboratory Medicine	53.3
Pharmacy	63.3
Public Health	56.7
History	80.0
Literature	90.0
Sociology	73.3
Agriculture	70.0
Architecture and Engineering	40.0
Computer Science	50.0
Electronics	26.7
Energy and Power	43.3
Materials	36.7
Mechanical Engineering	43.3
Overall	56.3

5.2 Historic versioning, changelog and timeline of Reka Chat models

We include the version history of Reka models to easily refer to them across this tech report.

Table 17: Version history of all Reka Edge, Core and Flash models.

Model	Date	Comments
Reka Core v0.5	Q2'24	Apr launch version
Reka Flash v1.5	Q2'24	Apr launch version
Reka Flash v1.0	Q1'24	Feb public launch version
Reka Edge v1.5	Q2'24	Apr launch version
Reka Edge v1.0	Q4'23	Feb public launch version
Reka Prototype v0.5	Q3'23	October private preview version

References

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

Anthropic. The claude 3 model family: Opus, sonnet, haiku. 2024.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. On the cross-lingual transferability of monolingual representations. *CoRR*, abs/1910.11856, 2019.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Jackson Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, and Jared Kaplan. A general language assistant as a laboratory for alignment, 2021.

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

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code. 2021.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, Alexandre Sallinen, Alireza Sakhaeirad, Vinitra Swamy, Igor Krawczuk, Deniz Bayazit, Axel Marmet, Syrielle Montariol, Mary-Anne Hartley, Martin Jaggi, and Antoine Bosselut. Meditron-70b: Scaling medical pretraining for large language models, 2023.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024.

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

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

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

Google, 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, 2023.

Gemini Team Google. Gemini: A family of highly capable multimodal models, 2023.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding, 2021.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

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.

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 generative language models. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.616. URL https://aclanthology.org/2022.emnlp-main.616.

OpenAI. Gpt-4 technical report, 2023.

OpenAI. Gpt-4v(ision) system card. 2024.

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

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. *CoRR*, abs/1912.01703, 2019. URL http://arxiv.org/abs/1912.01703.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. XCOPA: A multilingual dataset for causal commonsense reasoning. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.185. URL https://aclanthology.org/2020.emnlp-main.185.

Viorica Pătrăucean, Lucas Smaira, Ankush Gupta, Adrià Recasens Continente, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Joseph Heyward, Mateusz Malinowski, Yi Yang, Carl Doersch, Tatiana Matejovicova, Yury Sulsky, Antoine Miech, Alex Frechette, Hanna Klimczak, Raphael Koster, Junlin Zhang, Stephanie Winkler, Yusuf Aytar, Simon Osindero, Dima Damen, Andrew Zisserman, and João Carreira. Perception test: A diagnostic benchmark for multimodal video models, 2023.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*, 2019.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. Gpqa: A graduate-level google-proof q&a benchmark, 2023.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347, 2017. URL http://arxiv.org/abs/1707.06347.

Noam Shazeer. Fast transformer decoding: One write-head is all you need. *arXiv preprint arXiv:1911.02150*, 2019.

Noam Shazeer. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaekermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, Nenad Tomasev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. Towards expert-level medical question answering with large language models, 2023.

Jianlin Su, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*, 2021.

Yi Tay. Training great llms entirely from ground up in the wilderness as a startup. 2024.

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, 2023.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. *Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

xAI. Announcing grok. 2023.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of CVPR*, 2024.

Biao Zhang and Rico Sennrich. Root mean square layer normalization, 2019.