Gemini: A Family of Highly Capable Multimodal Models

Gemini Team, Google[1]

This report introduces a new family of multimodal models, Gemini, that exhibit remarkable capabilities across image, audio, video, and text understanding. The Gemini family consists of Ultra, Pro, and Nano sizes, suitable for applications ranging from complex reasoning tasks to on-device memory-constrained use-cases. Evaluation on a broad range of benchmarks shows that our most-capable Gemini Ultra model advances the state of the art in 30 of 32 of these benchmarks — notably being the first model to achieve human-expert performance on the well-studied exam benchmark MMLU, and improving the state of the art in every one of the 20 multimodal benchmarks we examined. We believe that the new capabilities of Gemini models in cross-modal reasoning and language understanding will enable a wide variety of use cases and we discuss our approach toward deploying them responsibly to users.

1. Introduction

We present Gemini, a family of highly capable multimodal models developed at Google. We trained Gemini jointly across image, audio, video, and text data for the purpose of building a model with both strong generalist capabilities across modalities alongside cutting-edge understanding and reasoning performance in each respective domain.

Gemini 1.0, our first version, comes in three sizes: Ultra for highly-complex tasks, Pro for enhanced performance and deployability at scale, and Nano for on-device applications. Each size is specifically tailored to address different computational limitations and application requirements. We evaluate the performance of Gemini models on a comprehensive suite of internal and external benchmarks covering a wide range of language, coding, reasoning, and multimodal tasks.

Gemini advances state-of-the-art in large-scale language modeling (Anil et al., 2023; Brown et al., 2020; Chowdhery et al., 2023; Hoffmann et al., 2022; OpenAI, 2023a; Radford et al., 2019; Rae et al., 2021), image understanding (Alayrac et al., 2022; Chen et al., 2022; Dosovitskiy et al., 2020; OpenAI, 2023b; Reed et al., 2022; Yu et al., 2022a), audio processing (Radford et al., 2023; Zhang et al., 2023), and video understanding(Alayrac et al., 2022; Chen et al., 2023). It also builds on the work on sequence models (Sutskever et al., 2014), a long history of work in deep learning based on neural networks (LeCun et al., 2015), and machine learning distributed systems (Barham et al., 2022; Bradbury et al., 2018; Dean et al., 2012) that enable large-scale training.

Our most capable model, Gemini Ultra, achieves new state-of-the-art results in 30 of 32 benchmarks we report on, including 10 of 12 popular text and reasoning benchmarks, 9 of 9 image understanding benchmarks, 6 of 6 video understanding benchmarks, and 5 of 5 speech recognition and speech translation benchmarks. Gemini Ultra is the first model to achieve human-expert performance on MMLU (Hendrycks et al., 2021a) — a prominent benchmark testing knowledge and reasoning via a suite of exams — with a score above 90%. Beyond text, Gemini Ultra makes notable advances on challenging multimodal reasoning tasks. For example, on the recent MMMU benchmark (Yue et al., 2023), that comprises questions about images on multi-discipline tasks requiring college-level subject

---

[1]See Contributions and Acknowledgments section for full author list. Please send correspondence to gemini-1-report@google.com

knowledge and deliberate reasoning, Gemini Ultra achieves a new state-of-the-art score of 62.4%, outperforming the previous best model by more than 5 percentage points. It provides a uniform performance lift for video question answering and audio understanding benchmarks.

Qualitative evaluation showcases impressive crossmodal reasoning capabilities, enabling the model to understand and reason across an input sequence of audio, images, and text natively (see Figure 5 and Table 13). Consider the educational setting depicted in Figure 1 as an example. A teacher has drawn a physics problem of a skier going down a slope, and a student has worked through a solution to it. Using Gemini's multimodal reasoning capabilities, the model is able to understand the messy handwriting, correctly understand the problem formulation, convert both the problem and solution to mathematical typesetting, identify the specific step of reasoning where the student went wrong in solving the problem, and then give a worked through correct solution to the problem. This opens up exciting educational possibilities, and we believe the new multimodal and reasoning capabilities of Gemini models have dramatic applications across many fields.
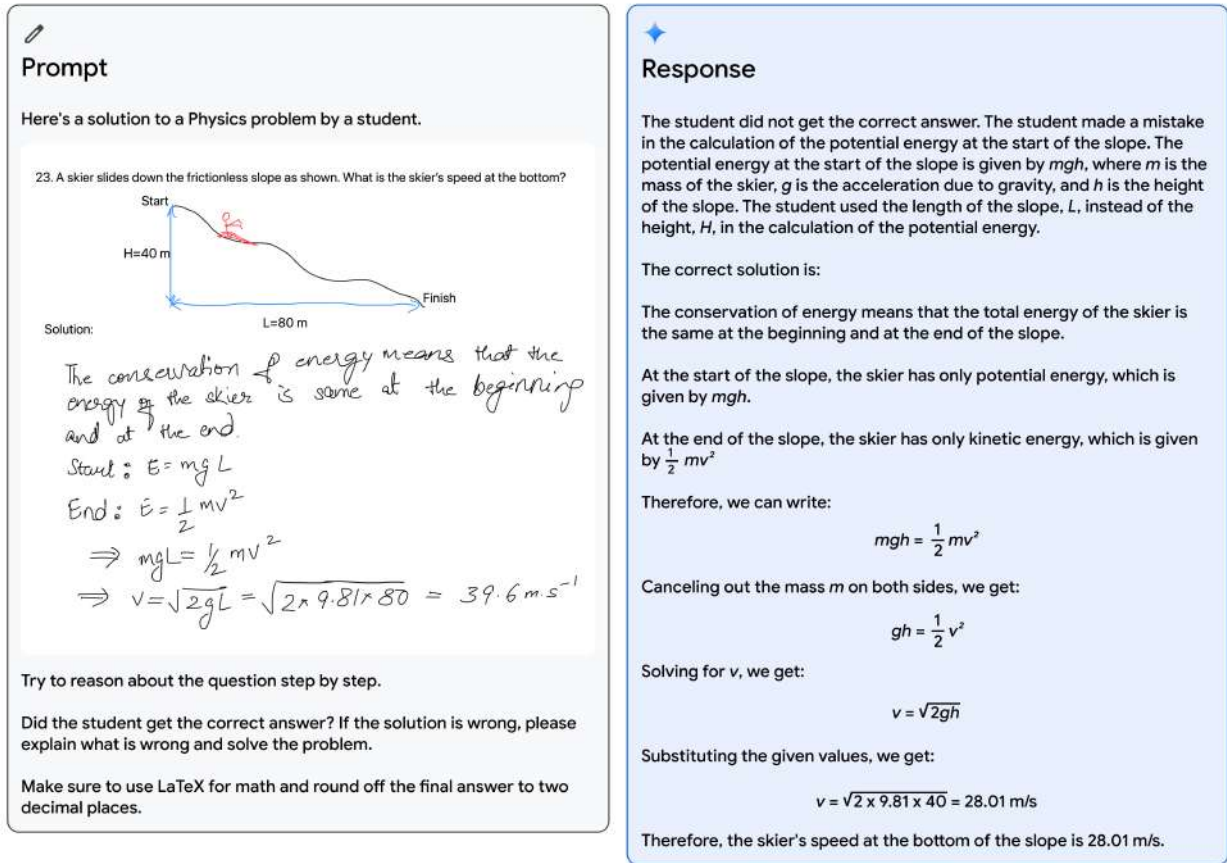

Figure 1 | Verifying a student's solution to a physics problem. The model is able to correctly recognize all of the handwritten content and verify the reasoning. On top of understanding the text in the image, it needs to understand the problem setup and correctly follow instructions to generate LaTeX.

The reasoning capabilities of large language models show promise toward building generalist agents that can tackle more complex multi-step problems. The AlphaCode team built AlphaCode 2 (Leblond et al, 2023), a new Gemini-powered agent, that combines Gemini's reasoning capabilities with search and tool-use to excel at solving competitive programming problems. AlphaCode 2 ranks within the top 15% of entrants on the Codeforces competitive programming platform, a large improvement over its state-of-the-art predecessor in the top 50% (Li et al., 2022).

In tandem, we advance the frontier of efficiency with Gemini Nano, a series of small models targeting on-device deployment. These models excel in on-device tasks, such as summarization, reading comprehension, text completion tasks, and exhibit impressive capabilities in reasoning, STEM, coding, multimodal, and multilingual tasks relative to their sizes.

In the following sections, we first provide an overview of the model architecture, training infrastructure, and training dataset. We then present detailed *evaluations* of the Gemini model family, covering well-studied benchmarks and human-preference evaluations across text, code, image, audio and video — which include both English performance and multilingual capabilities. We also discuss our approach to *responsible deployment*,[2] including our process for impact assessments, developing model policies, evaluations, and mitigations of harm before deployment decisions. Finally, we discuss the broader implications of Gemini, its limitations alongside its potential applications — paving the way for a new era of research and innovation in AI.

2. Model Architecture

Gemini models build on top of Transformer decoders (Vaswani et al., 2017) that are enhanced with improvements in architecture and model optimization to enable stable training at scale and optimized inference on Google's Tensor Processing Units. They are trained to support 32k context length, employing efficient attention mechanisms (for e.g. multi-query attention (Shazeer, 2019)). Our first version, Gemini 1.0, comprises three main sizes to support a wide range of applications as discussed in Table 1.

Model size	Model description
Ultra	Our most capable model that delivers state-of-the-art performance across a wide range of highly complex tasks, including reasoning and multimodal tasks. It is efficiently serveable at scale on TPU accelerators due to the Gemini architecture.
Pro	A performance-optimized model in terms of cost as well as latency that delivers significant performance across a wide range of tasks. This model exhibits strong reasoning performance and broad multimodal capabilities.
Nano	Our most efficient model, designed to run on-device. We trained two versions of Nano, with 1.8B (Nano-1) and 3.25B (Nano-2) parameters, targeting low and high memory devices respectively. It is trained by distilling from larger Gemini models. It is 4-bit quantized for deployment and provides best-in-class performance.

Table 1 | An overview of the Gemini 1.0 model family.

Gemini models are trained to accommodate textual input interleaved with a wide variety of audio and visual inputs, such as natural images, charts, screenshots, PDFs, and videos, and they can produce text and image outputs (see Figure 2). The visual encoding of Gemini models is inspired by our own foundational work on Flamingo (Alayrac et al., 2022), CoCa (Yu et al., 2022a), and PaLI (Chen et al., 2022), with the important distinction that the models are multimodal from the beginning and can natively output images using discrete image tokens (Ramesh et al., 2021; Yu et al., 2022b).

Video understanding is accomplished by encoding the video as a sequence of frames in the large context window. Video frames or images can be interleaved naturally with text or audio as part of the model input. The models can handle variable input resolution in order to spend more compute on

---

[2]We plan to update this report with more details ahead of the general availability of the Gemini Ultra model.
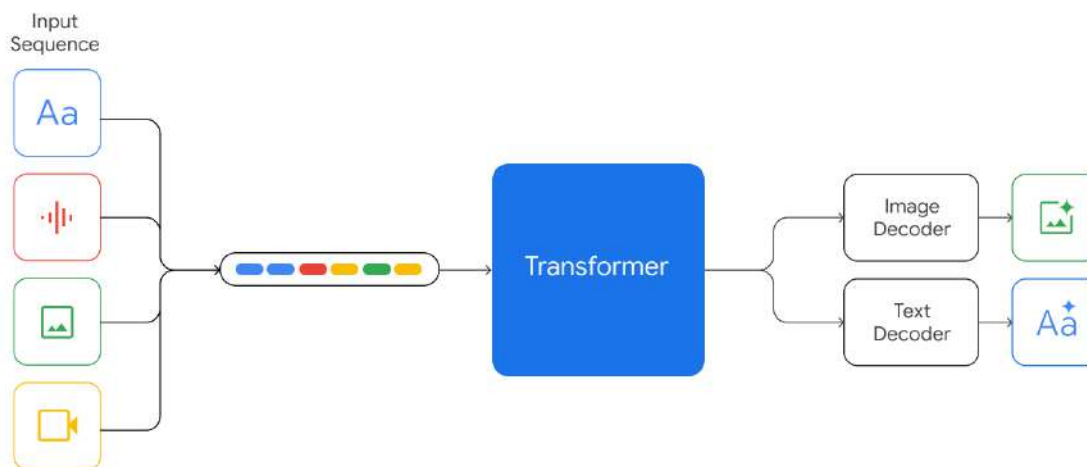
Figure 2 | Gemini supports interleaved sequences of text, image, audio, and video as inputs (illustrated by tokens of different colors in the input sequence). It can output responses with interleaved image and text.

tasks that require fine-grained understanding. In addition, Gemini can directly ingest audio signals at 16kHz from Universal Speech Model (USM) (Zhang et al., 2023) features. This enables the model to capture nuances that are typically lost when the audio is naively mapped to a text input (for example, see audio understanding demo on the website).

Training the Gemini family of models required innovations in training algorithms, dataset, and infrastructure. For the Pro model, the inherent scalability of our infrastructure and learning algorithms enable us to complete pretraining in a matter of weeks, leveraging a fraction of the Ultra's resources. The Nano series of models leverage additional advancements in distillation and training algorithms to produce the best-in-class small language models for a wide variety of tasks, such as summarization and reading comprehension, which power our next generation on-device experiences.

3. Training Infrastructure

We trained Gemini models using TPUv5e and TPUv4 (Jouppi et al., 2023), depending on their sizes and configuration. Training Gemini Ultra used a large fleet of TPUv4 accelerators across multiple datacenters. This represents a significant increase in scale over our prior flagship model PaLM-2 which presented new infrastructure challenges. Scaling up the number of accelerators results in a proportionate decrease in the mean time between failure of hardware in the overall system. We minimized the rate of planned reschedules and preemptions, but genuine machine failures are commonplace across all hardware accelerators at such large scales, due to external factors such as cosmic rays (Michalak et al., 2012).

TPUv4 accelerators are deployed in "SuperPods" of 4096 chips, each connected to a dedicated optical switch, which can dynamically reconfigure 4x4x4 chip cubes into arbitrary 3D torus topologies in around 10 seconds (Jouppi et al., 2023). For Gemini Ultra, we decided to retain a small number of cubes per superpod to allow for hot standbys and rolling maintenance.

TPU accelerators primarily communicate over the high speed inter-chip-interconnect, but at Gemini Ultra scale, we combine SuperPods in multiple datacenters using Google's intra-cluster and inter-cluster network (Poutievski et al., 2022; Wetherall et al., 2023; yao Hong et al., 2018). Google's

network latencies and bandwidths are sufficient to support the commonly used synchronous training paradigm, exploiting model parallelism within superpods and data-parallelism across superpods.

The 'single controller' programming model of Jax (Bradbury et al., 2018) and Pathways (Barham et al., 2022) allows a single Python process to orchestrate the entire training run, dramatically simplifying the development workflow. The GSPMD partitioner (Xu et al., 2021) in the XLA compiler partitions the training step computation, and the MegaScale XLA compiler (XLA, 2019) pass statically schedules appropriate collectives so that they maximally overlap with the computation with very little variation in step time.

Maintaining a high goodput[3] at this scale would have been impossible using the conventional approach of periodic checkpointing of weights to persistent cluster storage. For Gemini, we instead made use of redundant in-memory copies of the model state, and on any unplanned hardware failures, we rapidly recover directly from an intact model replica. Compared to both PaLM and PaLM-2 (Anil et al., 2023), this provided a substantial speedup in recovery time, despite the significantly larger training resources being used. As a result, the overall goodput for the largest-scale training job increased from 85% to 97%.

Training at unprecedented scale invariably surfaces new and interesting systems failure modes - and in this instance one of the problems that we needed to address was that of "Silent Data Corruption (SDC)" (Dixit et al., 2021; Hochschild et al., 2021; Vishwanathan et al., 2015). Although these are extremely rare, the scale of Gemini means that we can expect SDC events to impact training every week or two. Rapidly detecting and removing faulty hardware required several new techniques that exploit deterministic replay to isolate incorrect computations, combined with proactive SDC scanners on idle machines and hot standbys. Our fully deterministic infrastructure allowed us to quickly identify root causes (including hardware failures) during the development leading up to the Ultra model, and this was a crucial ingredient towards stable training.

4. Training Dataset

Gemini models are trained on a dataset that is both multimodal and multilingual. Our pretraining dataset uses data from web documents, books, and code, and includes image, audio, and video data.

We use the SentencePiece tokenizer (Kudo and Richardson, 2018) and find that training the tokenizer on a large sample of the entire training corpus improves the inferred vocabulary and subsequently improves model performance. For example, we find Gemini models can efficiently tokenize non-Latin scripts which can, in turn, benefit model quality as well as training and inference speed.

The number of tokens used to train the largest models were determined following the approach in Hoffmann et al. (2022). The smaller models are trained for significantly more tokens to improve performance for a given inference budget, similar to the approach advocated in Touvron et al. (2023a).

We apply quality filters to all datasets, using both heuristic rules and model-based classifiers. We also perform safety filtering to remove harmful content. We filter our evaluation sets from our training corpus. The final data mixtures and weights were determined through ablations on smaller models. We stage training to alter the mixture composition during training – increasing the weight of domain-relevant data towards the end of training. We find that data quality is critical to a highly-performing model, and believe that many interesting questions remain around finding the optimal dataset distribution for pretraining.

---

[3]We define goodput as the time spent computing useful new steps over the elapsed time of the training job.

5. Evaluation

The Gemini models are natively multimodal, as they are trained jointly across text, image, audio, and video. One open question is whether this joint training can result in a model which has strong capabilities in each domain – even when compared to models and approaches that are narrowly tailored to single domains. We find this to be the case: Gemini sets a new state of the art across a wide range of text, image, audio, and video benchmarks.

5.1. Text

5.1.1. Academic Benchmarks

We compare Gemini Pro and Ultra to a suite of external LLMs and our previous best model PaLM 2 across a series of text-based academic benchmarks covering reasoning, reading comprehension, STEM, and coding. We report these results in Table 2. Broadly, we find that the performance of Gemini Pro outperforms inference-optimized models such as GPT-3.5 and performs comparably with several of the most capable models available, and Gemini Ultra outperforms all current models. In this section, we examine some of these findings.

On MMLU (Hendrycks et al., 2021a), Gemini Ultra can outperform all existing models, achieving an accuracy of 90.04%. MMLU is a holistic exam benchmark, which measures knowledge across a set of 57 subjects. Human expert performance is gauged at 89.8% by the benchmark authors, and Gemini Ultra is the first model to exceed this threshold, with the prior state-of-the-art result at 86.4%. Achieving high performance requires specialist knowledge across many domains (e.g. law, biology, history, etc.), alongside reading comprehension and reasoning. We find Gemini Ultra achieves highest accuracy when used in combination with a chain-of-thought prompting approach (Wei et al., 2022) that accounts for model uncertainty. The model produces a chain of thought with k samples, for example 8 or 32. If there is a consensus above a preset threshold (selected based on the validation split), it selects this answer, otherwise it reverts to a greedy sample based on maximum likelihood choice without chain of thought. We refer the reader to appendix for a detailed breakdown of how this approach compares with only chain-of-thought prompting or only greedy sampling.

In mathematics, a field commonly used to benchmark the analytical capabilities of models, Gemini Ultra shows strong performance on both elementary exams and competition-grade problem sets. For the grade-school math benchmark, GSM8K (Cobbe et al., 2021), we find Gemini Ultra reaches 94.4% accuracy with chain-of-thought prompting and self-consistency (Wang et al., 2022) compared to the previous best accuracy of 92% with the same prompting technique. Similar positive trends are observed in increased difficulty math problems drawn from middle- and high-school math competitions (MATH benchmark), with the Gemini Ultra model outperforming all competitor models, reaching 53.2% using 4-shot prompting. The model also outperforms the state of the art on even harder tasks derived from American Mathematical Competitions (150 questions from 2022 and 2023). Smaller models perform poorly on this challenging task scoring close to random, but Gemini Ultra can solve 32% of the questions, compared to the 30% solve rate for GPT-4.

Gemini Ultra also excels in coding, a popular use case of current LLMs. We evaluate the model on many conventional and internal benchmarks and also measure its performance as part of more complex reasoning systems such as AlphaCode 2 (see section 5.1.7 on complex reasoning systems). For example, on HumanEval, a standard code-completion benchmark (Chen et al., 2021) mapping function descriptions to Python implementations, instruction-tuned Gemini Ultra correctly implements 74.4% of problems. On a new held-out evaluation benchmark for python code generation tasks, Natural2Code, where we ensure no web leakage, Gemini Ultra achieves the highest score of 74.9%.

	Gemini Ultra	Gemini Pro	GPT-4	GPT-3.5	PaLM 2-L	Claude 2	Inflection-2	Grok 1	LLAMA-2
MMLU Multiple-choice questions in 57 subjects (professional & academic) (Hendrycks et al., 2021a)	**90.04%** CoT@32* 83.7% 5-shot	79.13% CoT@8* 71.8% 5-shot	87.29% CoT@32 (via API**) 86.4% 5-shot (reported)	70% 5-shot	78.4% 5-shot	78.5% 5-shot CoT	79.6% 5-shot	73.0% 5-shot	68.0%*** 5-shot
GSM8K Grade-school math (Cobbe et al., 2021)	**94.4%** Maj1@32	86.5% Maj1@32	92.0% SFT & 5-shot CoT	57.1% 5-shot	80.0% 5-shot	88.0% 0-shot	81.4% 8-shot	62.9% 8-shot	56.8% 5-shot
MATH Math problems across 5 difficulty levels & 7 subdisciplines (Hendrycks et al., 2021b)	**53.2%** 4-shot	32.6% 4-shot	52.9% 4-shot (via API**) 50.3% (Zheng et al., 2023)	34.1% 4-shot (via API**)	34.4% 4-shot	—	34.8%	23.9% 4-shot	13.5% 4-shot
BIG-Bench-Hard Subset of hard BIG-bench tasks written as CoT problems (Srivastava et al., 2022)	**83.6%** 3-shot	75.0% 3-shot	83.1% 3-shot (via API**)	66.6% 3-shot (via API**)	77.7% 3-shot	—	—	—	51.2% 3-shot
HumanEval Python coding tasks (Chen et al., 2021)	**74.4%** 0-shot (IT)	67.7% 0-shot (IT)	67.0% 0-shot (reported)	48.1% 0-shot	—	70.0% 0-shot	44.5% 0-shot	63.2% 0-shot	29.9% 0-shot
Natural2Code Python code generation. (New held-out set with no leakage on web)	**74.9%** 0-shot	69.6% 0-shot	73.9% 0-shot (via API**)	62.3% 0-shot (via API**)	—	—	—	—	—
DROP Reading comprehension & arithmetic. (metric: F1-score) (Dua et al., 2019)	**82.4** Variable shots	74.1 Variable shots	80.9 3-shot (reported)	64.1 3-shot	82.0 Variable shots	—	—	—	—
HellaSwag (validation set) Common-sense multiple choice questions (Zellers et al., 2019)	87.8% 10-shot	84.7% 10-shot	**95.3%** 10-shot (reported)	85.5% 10-shot	86.8% 10-shot	—	89.0% 10-shot	—	80.0%***
WMT23 Machine translation (metric: BLEURT) (Tom et al., 2023)	**74.4** 1-shot (IT)	71.7 1-shot	73.8 1-shot (via API**)	—	72.7 1-shot	—	—	—	—

Table 2 | Gemini performance on text benchmarks with external comparisons and PaLM 2-L.

* The model produces a chain of thought with k = 8 or 32 samples, if there is a consensus above a threshold (chosen based on the validation split), it selects this answer, otherwise it reverts to a greedy sample. Further analysis in Appendix 9.1.

** Results self-collected via the API in Nov, 2023.

*** Results shown use the decontaminated numbers from Touvron et al. (2023b) report as the most relevant comparison to Gemini models which have been decontaminated as well.

Evaluation on these benchmarks is challenging and may be affected by data contamination. We performed an extensive leaked data analysis after training to ensure the results we report here are as scientifically sound as possible, but still found some minor issues and decided not to report results on e.g. LAMBADA (Paperno et al., 2016). As part of the evaluation process, on a popular benchmark, HellaSwag (Zellers et al., 2019), we find that an additional hundred finetuning steps on specific website extracts corresponding to the HellaSwag training set (which were not included in Gemini pretraining set) improve the validation accuracy of Gemini Pro to 89.6% and Gemini Ultra to 96.0%, when measured with 1-shot prompting (we measured GPT-4 obtained 92.3% when evaluated 1-shot via the API). This suggests that the benchmark results are susceptible to the pretraining dataset composition. We choose to report HellaSwag decontaminated results only in a 10-shot evaluation setting. We believe there is a need for more robust and nuanced standardized evaluation benchmarks with no leaked data. So, we evaluate Gemini models on several new held-out evaluation datasets that were recently released, such as WMT23 and Math-AMC 2022-2023 problems, or internally generated from non-web sources, such as Natural2Code. We refer the reader to the appendix for a comprehensive list of our evaluation benchmarks.

Even so, model performance on these benchmarks gives us an indication of the model capabilities and where they may provide impact on real-world tasks. For example, Gemini Ultra's impressive reasoning and STEM competencies pave the way for advancements in LLMs within the educational domain[4]. The ability to tackle complex mathematical and scientific concepts opens up exciting possibilities for personalized learning and intelligent tutoring systems.

5.1.2. Trends in Capabilities

We investigate the trends in capabilities across the Gemini model family by evaluating them on a holistic harness of more than 50 benchmarks in six different capabilities, noting that some of the most notable benchmarks were discussed in the last section. These capabilities are: "Factuality" covering open/closed-book retrieval and question answering tasks; "Long-Context" covering long-form summarization, retrieval and question answering tasks; "Math/Science" including tasks for mathematical problem solving, theorem proving, and scientific exams; "Reasoning" tasks that require arithmetic, scientific, and commonsense reasoning; "Multilingual" tasks for translation, summarization, and reasoning in multiple languages. Please see appendix for a detailed list of tasks included for each capability.
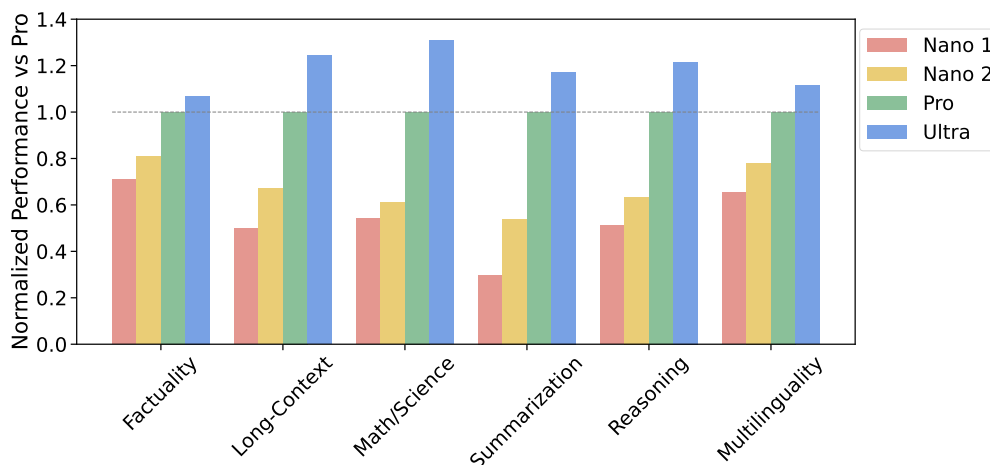
Figure 3 | Language understanding and generation performance of Gemini model family across different capabilities (normalized by the Gemini Pro model).

We observe consistent quality gains with increased model size in Figure 3, especially in reasoning, math/science, summarization and long-context. Gemini Ultra is the best model across the board for all six capabilities. Gemini Pro, the second-largest model in the Gemini family of models, is also quite competitive while being a lot more efficient to serve.

5.1.3. Nano

Bringing AI closer to the user, we discuss the Gemini Nano 1 and Nano 2 models engineered for on-device deployments. These models excel in summarization and reading comprehension tasks with per-task finetuning. Figure 3 shows the performance of these pretrained models in comparison to the much larger Gemini Pro model, while Table 3 dives deeper into specific factuality, coding, Math/Science, and reasoning tasks. Nano-1 and Nano-2 model sizes are only 1.8B and 3.25B parameters respectively. Despite their size, they show exceptionally strong performance on factuality, i.e. retrieval-related tasks, and significant performance on reasoning, STEM, coding, multimodal and

---

[4]See demos on website `https://deepmind.google/gemini`.

multilingual tasks. With new capabilities accessible to a broader set of platforms and devices, the Gemini models expand accessibility to everyone.

	Gemini Nano 1		Gemini Nano 2	
	accuracy	normalized by Pro	accuracy	normalized by Pro
BoolQ	71.6	0.81	79.3	0.90
TydiQA (GoldP)	68.9	0.85	74.2	0.91
NaturalQuestions (Retrieved)	38.6	0.69	46.5	0.83
NaturalQuestions (Closed-book)	18.8	0.43	24.8	0.56
BIG-Bench-Hard (3-shot)	34.8	0.47	42.4	0.58
MBPP	20.0	0.33	27.2	0.45
MATH (4-shot)	13.5	0.41	22.8	0.70
MMLU (5-shot)	45.9	0.64	55.8	0.78

Table 3 | Performance of Gemini Nano series on factuality, summarization, reasoning, coding and STEM tasks compared to significantly larger Gemini Pro model.

5.1.4. Multilinguality

The multilingual capabilities of the Gemini models are evaluated using a diverse set of tasks requiring multilingual understanding, cross-lingual generalization, and the generation of text in multiple languages. These tasks include machine translation benchmarks (WMT 23 for high-medium-low resource translation; Flores, NTREX for low and very low resource languages), summarization benchmarks (XLSum, Wikilingua), and translated versions of common benchmarks (MGSM: professionally translated into 11 languages).

Machine Translation Translation is a canonical benchmark in machine learning with a rich history. We evaluated Gemini Ultra with instruction-tuning applied (see section 6.4.2) on the entire set of language pairs in the WMT 23 translation benchmark in a few-shot setting. Overall, we found that Gemini Ultra (and other Gemini models) performed remarkably well at translating from English to any other language, and surpassed the LLM-based translation methods when translating out-of-English, on high-resource, mid-resource and low-resource languages. In the WMT 23 out-of-English translation tasks, Gemini Ultra achieved the highest LLM-based translation quality, with an average BLEURT (Sellam et al., 2020) score of 74.8, compared to GPT-4's score of 73.6, and PaLM 2's score of 72.2. When averaged across all language pairs and directions for WMT 23, we see a similar trend with Gemini Ultra 74.4, GPT-4 73.8 and PaLM 2-L 72.7 average BLEURT scores on this benchmark.

WMT 23 (Avg BLEURT)	Gemini Ultra	Gemini Pro	Gemini Nano 2	Gemini Nano 1	GPT-4	PaLM 2-L
High Resource	**74.2**	71.7	67.7	64.1	74.0	72.6
Mid Resource	**74.7**	71.8	67.0	64.8	73.6	72.7
Out-of-English	**74.8**	71.5	66.2	65.2	73.6	72.2
Into-English	73.9	72.0	69.0	63.5	**74.1**	73.4
All languages	**74.4**	71.7	67.4	64.8	73.8	72.7

Table 4 | Performance of Gemini models on WMT 23 translation benchmark. All numbers with 1-shot.

In addition to the languages and translation tasks above, we also evaluate Gemini Ultra on very low-resource languages. These languages were sampled from the tail of the following language sets: Flores-200 (Tamazight and Kanure), NTREX (North Ndebele), and an internal benchmark (Quechua).

For these languages, both from and into English, Gemini Ultra achieved an average chrF score of 27.0 in 1-shot setup, while the next-best model, PaLM 2-L, achieved a score of 25.3.

Multilingual Math and Summarization   Beyond translation, we evaluated how well Gemini performs in challenging tasks across a range of languages. We specifically investigated the math benchmark MGSM (Shi et al., 2023), which is a translated variant of the math benchmark GSM8K (Cobbe et al., 2021). We find Gemini Ultra achieves an accuracy of 79.0%, an advance over PaLM 2-L which scores 74.7%, when averaged across all languages in an 8-shot setup. We also benchmark Gemini on the multilingual summarization benchmarks – XLSum (Hasan et al., 2021) and WikiLingua (Ladhak et al., 2020). In XLSum, Gemini Ultra reached an average of 17.6 rougeL score compared to 15.4 for PaLM 2. For Wikilingua, Gemini Ultra (5-shot) trails behind PaLM 2 (3-shot) measured in BLEURT score. See Table 5 for the full results. Overall the diverse set of multilingual benchmarks show that Gemini family models have a broad language coverage, enabling them to also reach locales and regions with low-resource languages.

		Gemini Ultra	Gemini Pro	GPT-4	PaLM 2-L
MGSM	(8-shot)	**79.0**	63.5	74.5	74.7
XLsum	(3-shot)	**17.6**	16.2	—	15.4
Wikilingua		48.9	47.8	—	**50.4**

Table 5 | Performance of Gemini models on multilingual math and summarization.

5.1.5. Long Context

Gemini models are trained with a sequence length of 32,768 tokens and we find that they make use of their context length effectively. We first verify this by running a synthetic retrieval test: we place key-value pairs at the beginning of the context, then add long filler text, and ask for value associated with a particular key. We find that the Ultra model retrieves the correct value with 98% accuracy when queried across the full context length. We further investigate this by plotting the negative log likelihood (NLL) versus the token index across a held-out set of long documents in Figure 4. We find that the NLL decreases with sequence position up to the full 32K context length. The longer context length of Gemini models enable new use cases such as retrieval over documents and video understanding discussed in section 5.2.2.
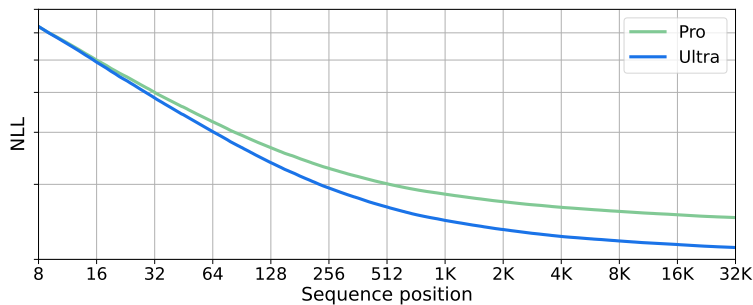

Figure 4 | Negative log likelihood as a function of token index across 32K context length on a held-out set of long documents.

5.1.6. Human Preference Evaluations

Human preference of the model outputs provides an important indication of quality that complements automated evaluations. We have evaluated the Gemini models in side-by-side blind evaluations where human raters judge responses of two models to the same prompt. We instruction tune (Ouyang et al., 2022) the pretrained model using techniques discussed in the section 6.4.2. The instruction-tuned version of the model is evaluated on a range of specific capabilities, such as following instructions, creative writing, multimodal understanding, long-context understanding, and safety. These capabilities encompass a range of use cases inspired by current user needs and research-inspired potential future use cases.

Instruction-tuned Gemini Pro models provide a large improvement on a range of capabilities, including preference for the Gemini Pro model over the PaLM 2 model API, 65.0% time in creative writing, 59.2% in following instructions, and 68.5% time for safer responses as shown in Table 6. These improvements directly translate into a more helpful and safer user experience.

	Creativity	Instruction Following	Safety
Win-rate	65.0%	59.2%	68.5%
95% Conf. Interval	[62.9%, 67.1%]	[57.6%, 60.8%]	[66.0%, 70.8%]

Table 6 | Win rate of Gemini Pro over PaLM 2 (text-bison@001) with 95% confidence intervals.

5.1.7. Complex Reasoning Systems

Gemini can also be combined with additional techniques such as search and tool-use to create powerful reasoning systems that can tackle more complex multi-step problems. One example of such a system is AlphaCode 2, a new state-of-the-art agent that excels at solving competitive programming problems (Leblond et al, 2023). AlphaCode 2 uses a specialized version of Gemini Pro – tuned on competitive programming data similar to the data used in Li et al. (2022) – to conduct a massive search over the space of possible programs. This is followed by a tailored filtering, clustering and reranking mechanism. Gemini Pro is fine-tuned both to be a coding model to generate proposal solution candidates, and to be a reward model that is leveraged to recognize and extract the most promising code candidates.

AlphaCode 2 is evaluated on Codeforces,[5] the same platform as AlphaCode, on 12 contests from division 1 and 2, for a total of 77 problems. AlphaCode 2 solved 43% of these competition problems, a 1.7x improvement over the prior record-setting AlphaCode system which solved 25%. Mapping this to competition rankings, AlphaCode 2 built on top of Gemini Pro sits at an estimated 85th percentile on average – i.e. it performs better than 85% of entrants. This is a significant advance over AlphaCode, which only outperformed 50% of competitors.

The composition of powerful pretrained models with search and reasoning mechanisms is an exciting direction towards more general agents; another key ingredient is deep understanding across a range of modalities which we discuss in the next section.

---

[5]http://codeforces.com/

5.2. Multimodal

Gemini models are natively multimodal. These models exhibit the unique ability to seamlessly combine their capabilities across modalities (e.g. extracting information and spatial layout out of a table, a chart, or a figure) with the strong reasoning capabilities of a language model (e.g. its state-of-art-performance in math and coding) as seen in examples in Figures 5 and 12. The models also show strong performance in discerning fine-grained details in inputs, aggregating context across space and time, and applying these capabilities over a temporally-related sequence of video frames and/or audio inputs.

The sections below provide more detailed evaluation of the model across different modalities (image, video, and audio), together with qualitative examples of the model's capabilities for image generation and the ability to combine information across different modalities.

5.2.1. Image Understanding

We evaluate the model on four different capabilities: high-level object recognition using captioning or question-answering tasks such as VQAv2; fine-grained transcription using tasks such as TextVQA and DocVQA requiring the model to recognize low-level details; chart understanding requiring spatial understanding of input layout using ChartQA and InfographicVQA tasks; and multimodal reasoning using tasks such as Ai2D, MathVista and MMMU. For zero-shot QA evaluation, the model is instructed to provide short answers aligned with the specific benchmark. All numbers are obtained using greedy sampling and without any use of external OCR tools.

	Gemini Ultra (pixel only)	Gemini Pro (pixel only)	Gemini Nano 2 (pixel only)	Gemini Nano 1 (pixel only)	GPT-4V	Prior SOTA
MMMU (val) Multi-discipline college-level problems (Yue et al., 2023)	**59.4%** pass@1 **62.4%** Maj1@32	47.9%	32.6%	26.3%	56.8%	56.8% GPT-4V, 0-shot
TextVQA (val) Text reading on natural images (Singh et al., 2019)	**82.3%**	74.6%	65.9%	62.5%	78.0%	79.5% Google PaLI-3, fine-tuned
DocVQA (test) Document understanding (Mathew et al., 2021)	**90.9%**	88.1%	74.3%	72.2%	88.4% (pixel only)	88.4% GPT-4V, 0-shot
ChartQA (test) Chart understanding (Masry et al., 2022)	**80.8%**	74.1%	51.9%	53.6%	78.5% (4-shot CoT)	79.3% Google DePlot, 1-shot PoT
InfographicVQA (test) Infographic understanding (Mathew et al., 2022)	**80.3%**	75.2%	54.5%	51.1%	75.1% (pixel only)	75.1% GPT-4V, 0-shot
MathVista (testmini) Mathematical reasoning (Lu et al., 2023)	**53.0%**	45.2%	30.6%	27.3%	49.9%	49.9% GPT-4V, 0-shot
AI2D (test) Science diagrams (Kembhavi et al., 2016)	**79.5%**	73.9%	51.0%	37.9%	78.2%	81.4% Google PaLI-X, fine-tuned
VQAv2 (test-dev) Natural image understanding (Goyal et al., 2017)	**77.8%**	71.2%	67.5%	62.7%	77.2%	86.1% Google PaLI-X, fine-tuned

Table 7 | **Image understanding** Gemini Ultra consistently outperforms existing approaches even in zero-shot, especially for OCR-related image understanding tasks for natural images, text, documents, and figures without using any external OCR engine ('pixel only'). Many existing approaches fine-tune on the respective tasks, highlighted in gray, which makes the comparison with 0-shot not apples-to-apples.

We find that Gemini Ultra is state of the art across a wide range of image-understanding benchmarks in Table 7. It achieves strong performance across a diverse set of tasks such as answering questions on natural images and scanned documents as well as understanding infographics, charts and science diagrams. When compared against publicly reported results from other models (most notably GPT-4V), Gemini is better in zero-shot evaluation by a significant margin. It also exceeds several existing models that are specifically fine-tuned on the benchmark's training sets for the majority of tasks. The capabilities of the Gemini models lead to significant improvements in the state of the art on academic benchmarks like MathVista (+3.1%)[6] or InfographicVQA (+5.2%).

MMMU (Yue et al., 2023) is a recently released evaluation benchmark, which consists of questions about images across 6 disciplines with multiple subjects within each discipline that require college-level knowledge to solve these questions. Gemini Ultra achieves the best score on this benchmark advancing the state-of-the-art result by more than 5 percentage points and outperforms the previous best result in 5 of 6 disciplines (see Table 8), thus showcasing its multimodal reasoning capabilities.

MMMU (val)	Gemini Ultra (0-shot)		GPT-4V (0-shot)
	Maj@32	pass@1	pass@1
Art & Design	**74.2**	70.0	65.8
Business	**62.7**	56.7	59.3
Science	49.3	48.0	**54.7**
Health & Medicine	**71.3**	67.3	64.7
Humanities & Social Science	**78.3**	78.3	72.5
Technology & Engineering	**53.0**	47.1	36.7
Overall	**62.4**	59.4	56.8

Table 8 | **Gemini Ultra performance on the MMMU** benchmark (Yue et al., 2023) per discipline. Each discipline covers multiple subjects, requiring college-level knowledge and complex reasoning.

Gemini models are also capable of operating across modalities and a diverse set of global languages simultaneously, both for image understanding tasks (e.g., images containing text in Icelandic) and for generation tasks (e.g., generating image descriptions for a wide range of languages). We evaluate the performance of generating image descriptions on a selected subset of languages in the Crossmodal-3600 (XM-3600) benchmark in a 4-shot setting, using the Flamingo evaluation protocol (Alayrac et al., 2022), without any fine-tuning for all models. As shown in Table 9, Gemini models achieve a significant improvement over the existing best model, Google PaLI-X.

XM-3600 (CIDER)	Gemini Ultra 4-shot	Gemini Pro 4-shot	Google PaLI-X 4-shot
English	86.4	**87.1**	77.8
French	**77.9**	76.7	62.5
Hindi	**31.1**	29.8	22.2
Modern Hebrew	**54.5**	52.6	38.7
Romanian	**39.0**	37.7	30.2
Thai	**86.7**	77.0	56.0
Chinese	**33.3**	30.2	27.7
Average (of 7)	**58.4**	55.9	45.0

Table 9 | **Multilingual image understanding** Gemini models outperform existing models in captioning images in many languages when benchmarked on a subset of languages in XM-3600 dataset (Thapliyal et al., 2022).

---

[6]MathVista is a comprehensive mathematical reasoning benchmark consisting of 28 previously published multimodal datasets and three newly created datasets. Our MathVista results were obtained by running the MathVista authors' evaluation script.
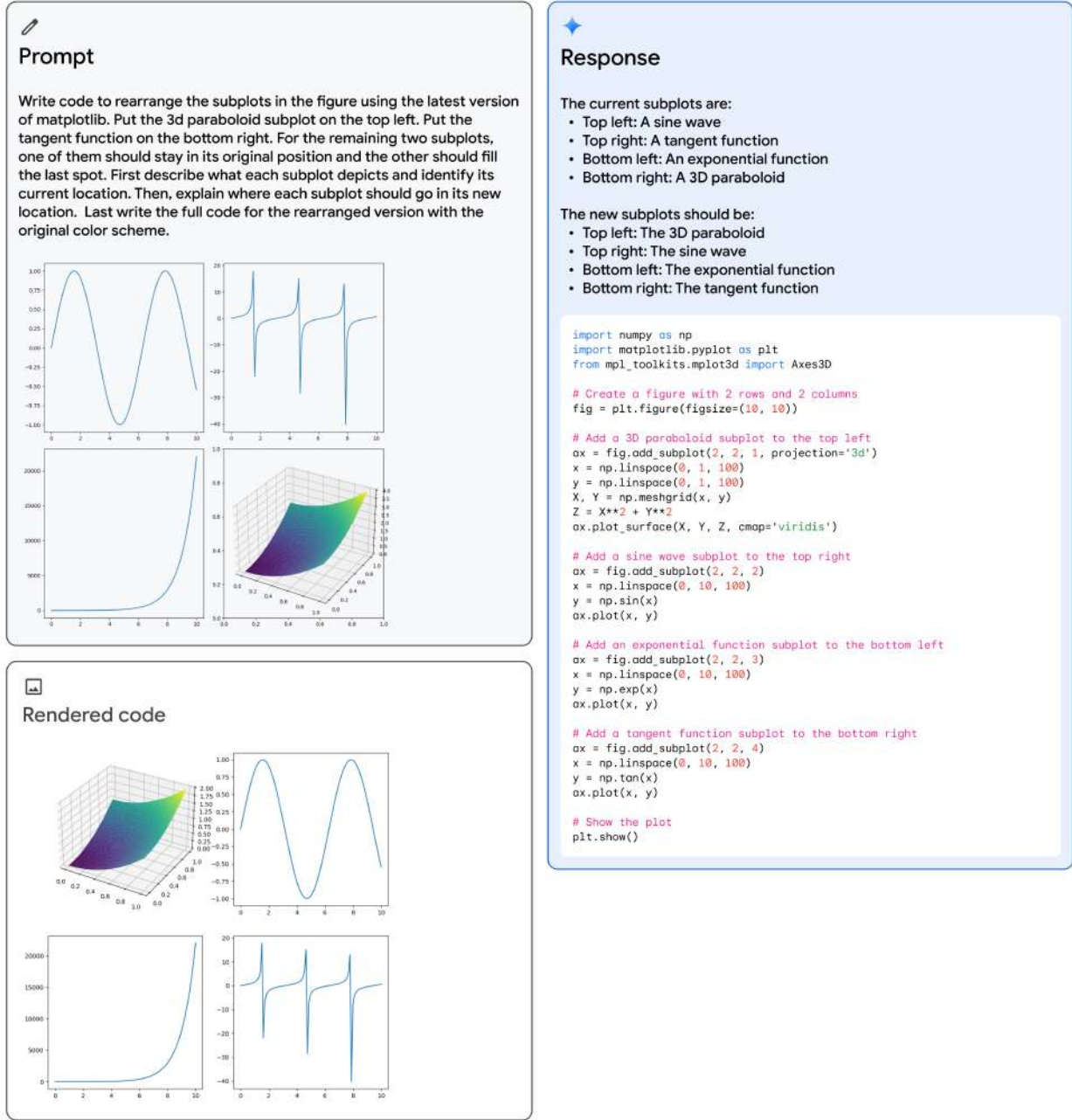
Figure 5 | **Gemini's multimodal reasoning capabilities** to generate matplotlib code for rearranging the subplots. The multimodal prompt is shown at the top-left in gray. Gemini Ultra's response, including its generated code, is shown in the right column in blue. The bottom left figure shows rendered version of the generated code. Successfully solving this task shows the model's capability to combine several capabilities: (1) **recognition** of the functions depicted in the plots; (2) **inverse graphics** to infer the code that would have generated the subplots; (3) **instruction-following** to put subplots in their desired positions; and (4) **abstract reasoning** to infer that the exponential plot must stay in its original place, because the sine plot must move out of the way for the 3-dimensional plot.

Qualitative evaluation in Figure 5 illustrates an example of Gemini Ultra's multimodal reasoning capabilities. The model is required to solve the task of generating matplotlib code that would rearrange a set of subplots provided by the user. The model output shows that it successfully solves this task

combining multiple capabilities of understanding the user plot, inferring the code required to generate it, following user instructions to put subplots in their desired positions, and abstract reasoning about the output plot. This highlights Gemini Ultra's native multimodality and eludes to its more complex reasoning abilities across interleaved sequences of image and text. We refer the reader to the appendix for more qualitative examples.

5.2.2. Video Understanding

Understanding video input is an important step towards a useful generalist agent. We measure the video understanding capability across several established benchmarks that are held-out from training. These tasks measure whether the model is able to understand and reason over a temporally-related sequence of frames. For each video task, we sample 16 equally-spaced frames from each video clip and feed them to the Gemini models. For the YouTube video datasets (all datasets except NextQA and the Perception test), we evaluate the Gemini models on videos that were still publicly available in the month of November, 2023.

Gemini Ultra achieves state-of-the-art results on various few-shot video captioning tasks as well as zero-shot video question answering tasks as shown in Table 10. This demonstrates its capability of strong temporal reasoning across several frames. Figure 21 in the appendix provides a qualitative example of understanding the video of the ball-striking mechanics of a soccer player and reasoning about the player can improve their game.

Task	Gemini Ultra	Gemini Pro	Few-shot SoTA
VATEX (test)	**62.7**	57.4	56.0
English video captioning (Wang et al., 2019)	4-shots	4-shots	DeepMind Flamingo, 4-shots
VATEX ZH (test)	**51.3**	50.0	–
Chinese video captioning (Wang et al., 2019)	4-shots	4-shots	
YouCook2 (val)	**135.4**	123.2	74.5
English cooking video captioning (Zhou et al., 2018)	4-shots	4-shots	DeepMind Flamingo, 4-shots
NextQA (test)	**29.9**	28.0	26.7
Video question answering (Xiao et al., 2021)	0-shot	0-shot	DeepMind Flamingo, 0-shot
ActivityNet-QA (test)	**52.2**	49.8	45.3
Video question answering (Yu et al., 2019)	0-shot	0-shot	Video-LLAVA, 0-shot
Perception Test MCQA (test)	**54.7**	51.1	46.3
Video question answering (Pătrăucean et al., 2023)	0-shot	0-shot	SeViLA (Yu et al., 2023), 0-shot

Table 10 | **Few-shot video understanding across tasks and languages** on selected academic benchmarks. The reported metric is CIDER for video captioning, WUPS for NextQA, and top-1 accuracy for the Perception Test and ActivityNet-QA. For ActivityNet-QA, we use the Video-LLAVA (Lin et al., 2023) evaluation protocol.

5.2.3. Image Generation

Gemini is able to output images natively, without having to rely on an intermediate natural language description that can bottleneck the model's ability to express images. This uniquely enables the model to generate images with prompts using interleaved sequences of image and text in a few-shot setting. For example, the user might prompt the model to design suggestions of images and text for a blog post or a website (see Figure 10 in the appendix).

Figure 6 shows an example of image generation in 1-shot setting. Gemini Ultra model is prompted with one example of interleaved image and text where the user provides two colors (blue and yellow) and image suggestions of creating a cute blue cat or a blue dog with yellow ear from yarn. The model is then given two new colors (pink and green) and asked for two ideas about what to create using these colors. The model successfully generates an interleaved sequence of images and text with suggestions to create a cute green avocado with pink seed or a green bunny with pink ears from yarn.
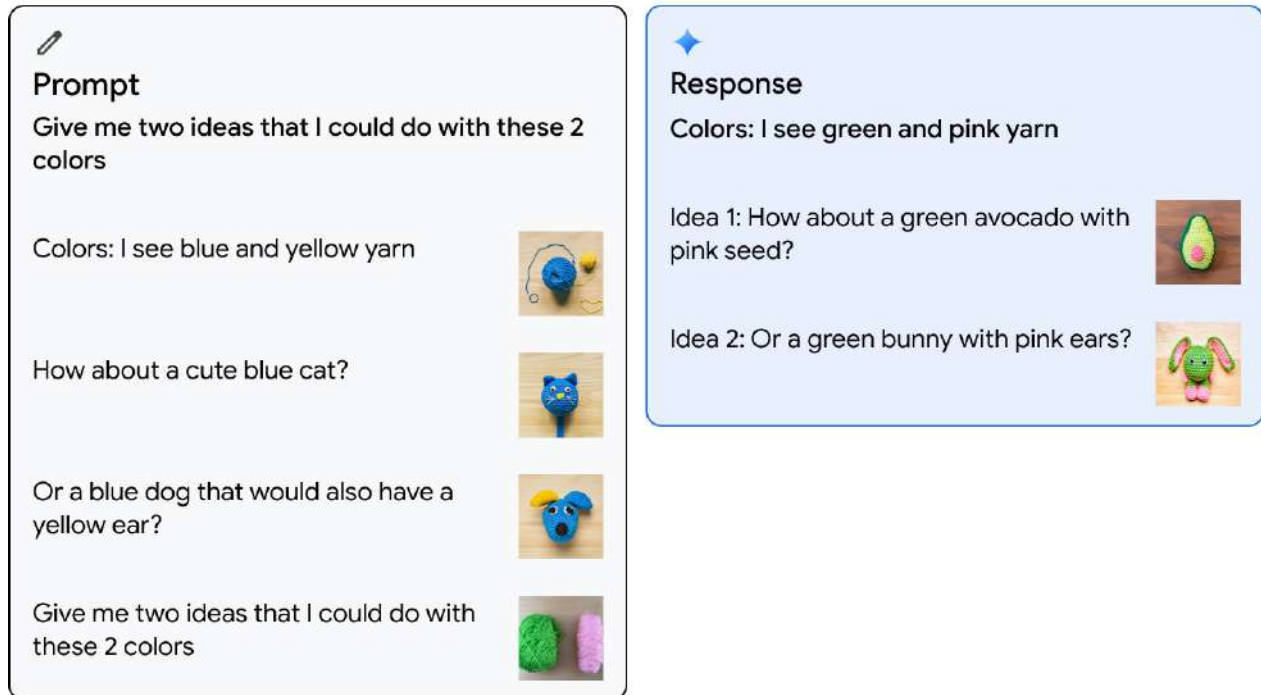
Figure 6 | **Image Generation.** Gemini can output multiple images interleaved with text given a prompt composed of image and text. In the left figure, Gemini Ultra is prompted in a 1-shot setting with a user example of generating suggestions of creating cat and dog from yarn when given two colors, blue and yellow. Then, the model is prompted to generate creative suggestions with two new colors, pink and green, and it generates images of creative suggestions to make a cute green avocado with pink seed or a green bunny with pink ears from yarn as shown in the right figure.

5.2.4. Audio Understanding

We evaluate the Gemini Nano-1 and Gemini Pro models on a variety of public benchmarks and compare it with Universal Speech Model (USM) (Zhang et al., 2023) and Whisper (large-v2 (Radford et al., 2023) or large-v3 (OpenAI, 2023) as indicated). These benchmarks include automatic speech recognition (ASR) tasks such as FLEURS (Conneau et al., 2023), VoxPopuli, (Wang et al., 2021), Multi-lingual Librispeech (Panayotov et al., 2015), as well as the speech translation task CoVoST 2, translating different languages into English (Wang et al., 2020). We also report on an internal benchmark YouTube test set. ASR tasks report a word error rate (WER) metric, where a lower number is better. Translation tasks report a BiLingual Evaluation Understudy (BLEU) score, where a higher number is better. FLEURS is reported on 62 languages that have language overlap with the training data. Four segmented languages (Mandarin, Japanese, Korean and Thai) report character error rate (CER), instead of WER, similar to Whisper (Radford et al., 2023).

Table 11 indicates that our Gemini Pro model significantly outperforms the USM and Whisper models across all ASR and AST tasks, both for English and multilingual test sets. Note that there is a large gain in FLEURS, compared to USM and Whisper, as our model is also trained with the FLEURS training dataset. However, training the same model without FLEURS dataset results in a WER of 15.8, which still outperforms Whisper. Gemini Nano-1 model also outperforms both USM and Whisper on all datasets except FLEURS. Note that we did not evaluate Gemini Ultra on audio yet, though we expect better performance from increased model scale.

	Task	Metric	Gemini Pro	Gemini Nano-1	Whisper (OpenAI, 2023; Radford et al., 2023)	USM (Zhang et al., 2023)
Automatic Speech Recognition	**YouTube** (en-us)	WER (↓)	**4.9%**	5.5%	6.5% (v3)	6.2%
	Multilingual Librispeech (en-us) (Panayotov et al., 2015)	WER (↓)	**4.8%**	5.9%	6.2% (v2)	7.0 %
	FLEURS (62 lang) (Conneau et al., 2023)	WER (↓)	**7.6%**	14.2%	17.6% (v3)	11.8%
	VoxPopuli (14 lang) (Wang et al., 2021)	WER (↓)	**9.1%**	9.5%	15.9% (v2)	13.4%
Automatic Speech Translation	**CoVoST 2** (21 lang) (Wang et al., 2020)	BLEU (↑)	**40.1**	35.4	29.1 (v2)	30.7

Table 11 | Speech evaluation results on selected benchmarks for ASR and AST. For ASR, the reported metric is WER where lower is better. For AST, the reported metric is BLEU where higher is better.

Table 12 shows further error analysis with USM and Gemini Pro. We find that Gemini Pro produces more understandable responses, particularly on rare words and proper nouns.

Domain	Truth	USM	Gemini Pro	Wav
Fleurs	Scotturb bus 403 travels regularly to Sintra, stopping at Cabo da Roca.	Scotboard bus four3 traversed regularly to Centra stopping at Cabo de Roga.	Scotturb bus 403 travels regularly to Sintra, stopping at Cabo da Roca.	🔊
Fleurs	The archipelago lies 120 km north of the Peninsula. The largest is King George Island, with the settlement of Villa Las Estrellas.	The archipelago lines 120 km north of peninsula. The largest is Kingurch island with the settlement of Cua Losas.	The archipelago lies 120 km north of the Peninsula. The largest is King George Island, with the settlement of Villa Las Estrellas.	🔊

Table 12 | Qualitative examples for the ASR task in the benchmark. Incorrect transcriptions are highlighted in red.

5.2.5. Modality Combination

Multimodal demonstrations often include a combination of text interleaved with a single modality, usually images. We demonstrate the ability to process a sequence of audio and images natively.

Consider a cooking scenario about making an omelet where we prompt the model with a sequence of audio and images. Table 13 indicates a turn-by-turn interaction with the model, providing pictures and verbally asking questions about the next steps for cooking an omelet. We note that the model response text is reasonably accurate, and shows that model processes fine-grained image details to evaluate when the omelet is fully cooked. See demo on the website.

Input Image	Input Audio (transcribed)	Model Response: Text
	🔊 What's the first step to make a veggie omelet with these ingredients?	Crack the eggs into a bowl and whisk them.
	🔊 Thank you for the instructions. I started making my omelet, does it look ready now?	It looks like it's almost ready. You can flip it over to cook the other side.
(No image - it's a follow up on the previous question)	🔊 Why is it not ready?	It's not ready because the eggs are still runny.
	🔊 What about now?	It looks ready now. You can take it off the heat and serve it.

Table 13 | Audio-visual qualitative example showcasing the ability of Gemini models to process interleaved sequences of text, vision, and audio, as well as reason across modalities. This example inputs interleaved images and audio from the user in a cooking scenario. The user prompts the model for instructions to make an omelet and to inspect whether it is fully cooked.

6. Responsible Deployment

During the development of the Gemini models, we follow a structured approach to responsible deployment in order to identify, measure, and manage foreseeable downstream societal impacts of our models, in line with previous releases of Google's AI technology (Kavukcuoglu et al., 2022). Throughout the lifecycle of the project, we follow the structure below. This section outlines our broad approach and key findings through this process. We will share more details on this in an upcoming report.
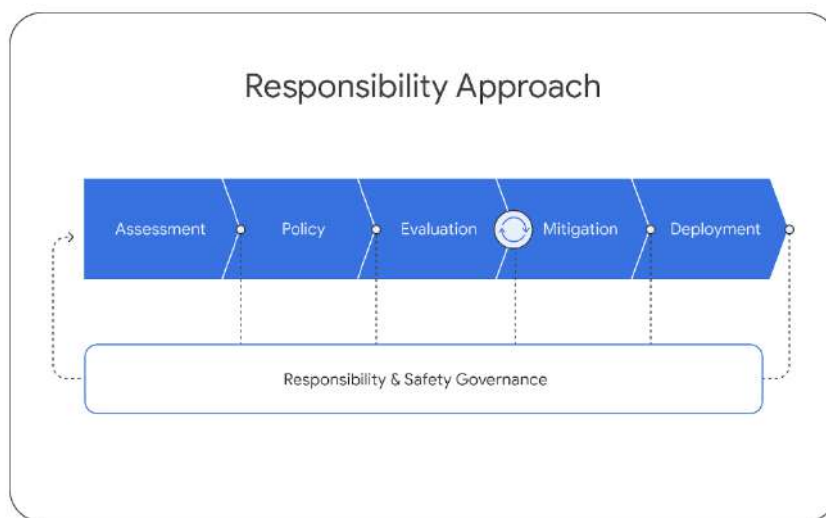
6.1. Impact Assessment

We develop model impact assessments to identify, assess, and document key downstream societal benefits and harms associated with the development of advanced Gemini models. These are informed by prior academic literature on language model risks (Weidinger et al., 2021), findings from similar prior exercises conducted across the industry (Anil et al., 2023; Anthropic, 2023; OpenAI, 2023a), ongoing engagement with experts internally and externally, and unstructured attempts to discover new model vulnerabilities. Areas of focus include: factuality, child safety, harmful content, cybersecurity, biorisk, representation and inclusivity. These assessments are updated in tandem with model development.

Impact assessments are used to guide mitigation and product delivery efforts, and inform deployment decisions. Gemini impact assessments spanned across different capabilities of Gemini models, assessing the potential consequences of these capabilities with Google's AI Principles (Google, 2023).

6.2. Model Policy

Building upon this understanding of known and anticipated effects, we developed a set of "model policies" to steer model development and evaluations. Model policy definitions act as a standardized criteria and prioritization schema for responsible development and as an indication of launch-readiness. Gemini model policies cover a number of domains including: child safety, hate speech, factual accuracy, fairness and inclusion, and harassment.

6.3. Evaluations

To assess the Gemini models against policy areas and other key risk areas identified within impact assessments, we developed a suite of evaluations across the lifecycle of model development.

Development evaluations are conducted for the purpose of 'hill-climbing' throughout training and fine-tuning Gemini models. These evaluations are designed by the Gemini team, or are assessments against external academic benchmarks. Evaluations consider issues such as helpfulness (instruction following and creativity), safety and factuality. See section 5.1.6 and the next section on mitigations for a sample of results.

Assurance evaluations are conducted for the purpose of governance and review, usually at the end of key milestones or training runs by a group outside of the model development team. Assurance evaluations are standardized by modality and datasets are strictly held-out. Only high-level insights are fed back into the training process to assist with mitigation efforts. Assurance evaluations include testing across Gemini policies, and include ongoing testing for dangerous capabilities such as potential biohazards, persuasion, and cybersecurity (Shevlane et al., 2023).

External evaluations are conducted by partners outside of Google to identify blindspots. External groups stress-test our models across a range of issues, including across areas listed in the White House Commitments,[7] and tests are conducted through a mixture of structured evaluations and unstructured red teaming. The design of these evaluations are independent and results are reported periodically to the Google DeepMind team.

In addition to this suite of external evaluations, specialist internal teams conduct ongoing red teaming of our models across areas such as the Gemini policies and security. These activities include less structured processes involving sophisticated adversarial attacks to identify new vulnerabilities. Discovery of potential weaknesses can then be used to mitigate risks and improve evaluation approaches internally. We are committed to ongoing model transparency and plan to share additional results from across our evaluation suite over time.

6.4. Mitigations

Mitigations are developed in response to the outcomes of the assessment, policy, and evaluation approaches described above. Evaluations and mitigations are used in an iterative way, with evaluations being re-run following mitigation efforts. We discuss our efforts on mitigating model harms across data, instruction-tuning, and factuality below.

6.4.1. Data

Prior to training, we take various steps to mitigate potential downstream harms at the data curation and data collection stage. As discussed in the section on "Training Data", we filter training data for high-risk content and to ensure all training data is sufficiently high quality. Beyond filtering, we also take steps to ensure all data collected meets Google DeepMind's best practices on data enrichment,[8] developed based on the Partnership on AI's "Responsible Sourcing of Data Enrichment Services"[9]. This includes ensuring all data enrichment workers are paid at least a local living wage.

---

[7] https://whitehouse.gov/wp-content/uploads/2023/07/Ensuring-Safe-Secure-and-Trustworthy-AI.pdf
[8] https://deepmind.google/discover/blog/best-practices-for-data-enrichment/
[9] https://partnershiponai.org/responsible-sourcing-considerations/

6.4.2. Instruction Tuning

Instruction tuning encompasses supervised fine tuning (SFT) and reinforcement learning through human feedback (RLHF) using a reward model. We apply instruction tuning in both text and multimodal settings. Instruction tuning recipes are carefully designed to balance the increase in helpfulness with decrease in model harms related to safety and hallucinations (Bai et al., 2022a).

Curation of "quality" data is critical for SFT, reward model training, and RLHF. The data mixture ratios are ablated with smaller models to balance the metrics on helpfulness (such as instruction following, creativity) and reduction of model harms, and these results generalize well to larger models. We have also observed that data quality is more important than quantity (Touvron et al., 2023b; Zhou et al., 2023), especially for larger models. Similarly, for reward model training, we find it critical to balance the dataset with examples where the model prefers to say, "I cannot help with that," for safety reasons and examples where the model outputs helpful responses. We use multi-objective optimization with a weighted sum of reward scores from helpfulness, factuality, and safety, to train a multi-headed reward model.

We further elaborate our approach to mitigate risks of harmful text generation. We enumerate approximately 20 harm types (e.g. hate speech, providing medical advice, suggesting dangerous behavior) across a wide variety of use cases. We generate a dataset of potential harm-inducing queries in these categories, either manually by policy experts and ML engineers, or via prompting high capability language models with topical keywords as seeds.

Given the harm-inducing queries, we probe our Gemini models and analyze the model responses via side-by-side evaluation. As discussed above, we balance the objective of model output response being harmless versus being helpful. From the detected risk areas, we create additional supervised fine-tuning data to demonstrate the desirable responses. To generate such responses at scale, we heavily rely on a custom data generation recipe loosely inspired from Constitutional AI (Bai et al., 2022b), where we inject variants of Google's content policy language as "constitutions", and utilize language model's strong zero-shot reasoning abilities (Kojima et al., 2022) to revise responses and choose between multiple response candidates. We have found this recipe to be effective – for example in Gemini Pro, this overall recipe was able to mitigate a majority of our identified text harm cases, without any perceptible decrease on response helpfulness.

6.4.3. Factuality

It is important that our models generate responses that are factual in a variety of scenarios, and to reduce the frequency of hallucinations. We focused instruction tuning efforts on three key desired behaviors, reflecting real-world scenarios:

1. **Attribution**: If instructed to generate a response that should be fully attributed to a given context in the prompt, Gemini should produce a response with the highest degree of faithfulness to the context (Rashkin et al., 2023). This includes the summarization of a user-provided source, generating fine-grained citations given a question and provided snippets akin to Menick et al. (2022); Peng et al. (2023), answering questions from a long-form source such as a book (Mihaylov et al., 2018), and transforming a given source to a desired output (e.g. an email from a portion of a meeting transcript).
2. **Closed-Book Response Generation**: If provided with a fact-seeking prompt without any given source, Gemini should not hallucinate incorrect information (see Section 2 of Roberts et al. (2020) for a definition). These prompts can range from information-seeking prompts (e.g. "Who is the prime minister of India?") to semi-creative prompts that may request factual information (e.g. "Write a 500-word speech in favor of the adoption of renewable energy").

3. **Hedging**: If prompted with an input such that it is "unanswerable", Gemini should not hallucinate. Rather, it should acknowledge that it cannot provide a response by hedging. These include scenarios where the input prompt contains false-premise questions (see examples in Hu et al. (2023)), the input prompt instructs the model to perform open-book QA, but the answer is not derivable from the given context, and so forth.

We elicited these desired behaviors from Gemini models by curating targeted supervised-fine tuning datasets and performing RLHF. Note that the results produced here do not include endowing Gemini with tools or retrieval that purportedly could boost factuality (Menick et al., 2022; Peng et al., 2023). We provide three key results on respective challenge sets below.

1. **Factuality Set:** An evaluation set containing fact-seeking prompts (primarily closed-book). This is evaluated via human annotators who fact-check each response manually; we report the percentage of factually-inaccurate responses as judged by annotators.
2. **Attribution Set:** An evaluation set containing a variety of prompts that require attribution to sources in the prompt. This is evaluated via human annotators who check for attribution to sources in the prompt for each response manually; the reported metric is AIS (Rashkin et al., 2023).
3. **Hedging Set:** An automatic evaluation setup where we measure whether Gemini models hedge accurately.

We compare Gemini Pro with a version of instruction-tuned Gemini Pro model without any factuality-focused adaptation in Table 14. We observe that the rate of inaccuracy is halved in the factuality set, the accuracy of attribution is increased by 50% from the attribution set, and the model successfully hedges 70% (up from 0%) in the provided hedging set task.

	Factuality Set (Inaccurate Rate)	Attribution Set (AIS)	Hedging Set (Accuracy)
Gemini Pro No factuality-focused adaptation	7.9% [7%, 9%]	40.2% [37.9%, 42.4%]	0%
Gemini Pro Final stage of instruction tuning	3.4% [2.8%, 4.1%]	59.7% [57.2%, 61.9%]	69.30%

Table 14 | Factuality mitigations: Impact of instruction-tuning on the rate of inaccuracy, presence of attribution and the rate of accurate hedging (with corresponding 95% confidence intervals).

6.5. Deployment

Following the completion of reviews, model cards for each approved Gemini model are created for structured and consistent internal documentation of critical performance and responsibility metrics as well as to inform appropriate external communication of these metrics over time.

6.6. Responsible Governance

Across the responsible development process, we undertake ethics and safety reviews with the Google DeepMind's Responsibility and Safety Council (RSC),[10] an interdisciplinary group which evaluates Google DeepMind's projects, papers and collaborations against Google's AI Principles. The RSC provides input and feedback on impact assessments, policies, evaluations and mitigation efforts. During the Gemini project, the RSC set specific evaluation targets across key policy domains (e.g. child safety).

---

[10] https://deepmind.google/about/responsibility-safety/

7. Discussion and Conclusion

We have presented Gemini, a new family of models that advance multimodal model capabilities in text, code, image, audio, and video. This technical report evaluates the capabilities of Gemini on a diverse set of widely-studied benchmarks, and our most capable model Gemini Ultra makes significant advances across the board. In the natural language domain, the performance gains from careful developments in data and model training at scale continue to deliver quality improvements, setting new state of the art in several benchmarks. In particular, Gemini Ultra surpasses human-expert performance on the exam benchmark MMLU, scoring 90.0%, which has been a defacto measure of progress for LLMs ever since it was first released in 2020. In the multimodal domain, Gemini Ultra sets new state of the art on most of the image understanding, video understanding, and audio understanding benchmarks without task-specific modifications or tuning. In particular, Gemini Ultra's multimodal reasoning capabilities are evident from its state-of-the-art performance on the recent MMMU benchmark (Yue et al., 2023), that comprises questions about images requiring college-level subject knowledge and deliberate reasoning.

Beyond the state-of-art results on benchmarks, what we are most excited about is the new use cases enabled by Gemini models. The new capabilities of Gemini models to parse complex images, such as charts or infographics, reason over interleaved sequences of images, audio, and text, and generate interleaved text and images as responses open a wide variety of new applications. As shown in figures throughout the report and appendix, Gemini can enable new approaches in areas like education, everyday problem solving, multilingual communication, information summarization, extraction, and creativity. We expect that the users of these models will find all kinds of beneficial new uses that we have only scratched the surface of in our own investigations.

Despite their impressive capabilities, we should note that there are limitations to the use of LLMs. There is a continued need for ongoing research and development on "hallucinations" generated by LLMs to ensure that model outputs are more reliable and verifiable. LLMs also struggle with tasks requiring high-level reasoning abilities like causal understanding, logical deduction, and counterfactual reasoning even though they achieve impressive performance on exam benchmarks. This underscores the need for more challenging and robust evaluations to measure their true understanding as the current state-of-the-art LLMs saturate many benchmarks.

Gemini is a further step towards our mission to solve intelligence, advance science and benefit humanity, and we are enthusiastic to see how these models are used by our colleagues at Google and beyond. We build on many innovations in machine learning, data, infrastructure, and responsible development – areas that we have been pursuing at Google for over a decade. The models we present in this report provide a strong foundation towards our broader future goal to develop a large-scale, modularized system that will have broad generalization capabilities across many modalities.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.

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.

Anthropic. Model Card and Evaluations for Claude Models, 2023.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback. April 2022a. URL https://arxiv.org/abs/2204.05862.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022b.

Paul Barham, Aakanksha Chowdhery, Jeff Dean, Sanjay Ghemawat, Steven Hand, Daniel Hurt, Michael Isard, Hyeontaek Lim, Ruoming Pang, Sudip Roy, et al. Pathways: Asynchronous distributed dataflow for ml. *Proceedings of Machine Learning and Systems*, 4:430–449, 2022.

James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018. URL http://github.com/google/jax.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey

Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

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. *arXiv preprint arXiv:2107.03374*, 2021. URL https://arxiv.org/abs/2107.03374.

Xi Chen, Xiao Wang, Soravit Changpinyo, A J Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme, Andreas Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. PaLI: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*, 2022. URL https://arxiv.org/abs/2209.06794.

Xi Chen, Josip Djolonga, Piotr Padlewski, Basil Mustafa, Soravit Changpinyo, Jialin Wu, Carlos Riquelme Ruiz, Sebastian Goodman, Xiao Wang, Yi Tay, Siamak Shakeri, Mostafa Dehghani, Daniel Salz, Mario Lucic, Michael Tschannen, Arsha Nagrani, Hexiang Hu, Mandar Joshi, Bo Pang, Ceslee Montgomery, Paulina Pietrzyk, Marvin Ritter, AJ Piergiovanni, Matthias Minderer, Filip Pavetic, Austin Waters, Gang Li, Ibrahim Alabdulmohsin, Lucas Beyer, Julien Amelot, Kenton Lee, Andreas Peter Steiner, Yang Li, Daniel Keysers, Anurag Arnab, Yuanzhong Xu, Keran Rong, Alexander Kolesnikov, Mojtaba Seyedhosseini, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. PaLI-X: On Scaling up a Multilingual Vision and Language Model. *arXiv preprint arXiv:2305.18565*, 2023.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240): 1–113, 2023. URL http://jmlr.org/papers/v24/22-1144.html.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936, 2019. URL https://aclanthology.org/N19-1300.

Jon Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. TydiQA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 2020. URL https://storage.googleapis.com/tydiqa/tydiqa.pdf.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. URL https://arxiv.org/abs/2110.14168.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. Fleurs: Few-shot learning evaluation of universal representations of speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 798–805. IEEE, 2023.

Jeffrey Dean, Greg Corrado, Rajat Monga, Kai Chen, Matthieu Devin, Mark Mao, Marc'aurelio Ranzato, Andrew Senior, Paul Tucker, Ke Yang, et al. Large scale distributed deep networks. *Advances in neural information processing systems*, 25, 2012.

Harish Dattatraya Dixit, Sneha Pendharkar, Matt Beadon, Chris Mason, Tejasvi Chakravarthy, Bharath Muthiah, and Sriram Sankar. Silent data corruptions at scale. *arXiv preprint arXiv:2102.11245*, 2021.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2020.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2368–2378, 2019. URL https://aclanthology.org/N19-1246.

Christian Federmann, Tom Kocmi, and Ying Xin. NTREX-128 – news test references for MT evaluation of 128 languages. In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pages 21–24, Online, nov 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.sumeval-1.4.

Google. Google's AI Principles. 2023. URL https://ai.google/responsibility/principles/.

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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017.

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

pages 4693–4703, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.413. URL https://aclanthology.org/2021.findings-acl.413.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021a.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. *arXiv preprint arXiv:2103.03874*, 2021b. URL https://arxiv.org/abs/2103.03874.

Peter H Hochschild, Paul Turner, Jeffrey C Mogul, Rama Govindaraju, Parthasarathy Ranganathan, David E Culler, and Amin Vahdat. Cores that don't count. In *Proceedings of the Workshop on Hot Topics in Operating Systems*, pages 9–16, 2021.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Shengding Hu, Yifan Luo, Huadong Wang, Xingyi Cheng, Zhiyuan Liu, and Maosong Sun. Won't get fooled again: Answering questions with false premises. *arXiv preprint arXiv:2307.02394*, 2023.

EunJeong Hwang and Vered Shwartz. Memecap: A dataset for captioning and interpreting memes, 2023.

Norm Jouppi, George Kurian, Sheng Li, Peter Ma, Rahul Nagarajan, Lifeng Nai, Nishant Patil, Suvinay Subramanian, Andy Swing, Brian Towles, et al. Tpu v4: An optically reconfigurable supercomputer for machine learning with hardware support for embeddings. In *Proceedings of the 50th Annual International Symposium on Computer Architecture*, pages 1–14, 2023.

Ashwin Kalyan, Abhinav Kumar, Arjun Chandrasekaran, Ashish Sabharwal, and Peter Clark. How much coffee was consumed during emnlp 2019? fermi problems: A new reasoning challenge for ai, 2021.

Jungo Kasai, Keisuke Sakaguchi, Yoichi Takahashi, Ronan Le Bras, Akari Asai, Xinyan Yu, Dragomir Radev, Noah A. Smith, Yejin Choi, and Kentaro Inui. RealTime QA: What's the answer right now?, 2022. URL https://arxiv.org/abs/2207.13332.

K Kavukcuoglu, P Kohli, L Ibrahim, D Bloxwich, and S Brown. How our principles helped define alphafold's release. google deepmind, 2022.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *ECCV*, 2016.

Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. The NarrativeQA reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328, 2018. doi: 10.1162/tacl_a_00023. URL https://aclanthology.org/Q18-1023.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák,

Martin Popel, and Maja Popović. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, December 2022. URL https://aclanthology.org/2022.wmt-1.1.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *NeurIPS*, 2022. URL https://arxiv.org/abs/2205.11916.

Taku Kudo and John Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *EMNLP (System Demonstrations)*, 2018. doi: 10.18653/v1/D18-2012. URL https://aclanthology.org/D18-2012.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019. doi: 10.1162/tacl_a_00276. URL https://aclanthology.org/Q19-1026.

Faisal Ladhak, Esin Durmus, Claire Cardie, and Kathleen McKeown. WikiLingua: A new benchmark dataset for cross-lingual abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4034–4048, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.360. URL https://www.aclweb.org/anthology/2020.findings-emnlp.360.

Leblond et al. AlphaCode 2 Technical Report. 2023. URL https://storage.googleapis.com/deepmind-media/AlphaCode2/AlphaCode2_Tech_Report.pdf.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Rémi Leblond, Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, et al. Competition-level code generation with alphacode. *Science*, 378(6624):1092–1097, 2022.

Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023.

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. In *The Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021)*, 2021.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023.

Ahmed Masry, Do Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of ACL*, 2022.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209, 2021.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706, 2022.

Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, Francis Song, Martin Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, et al. Teaching language models to support answers with verified quotes. *arXiv preprint arXiv:2203.11147*, 2022.

Sarah E. Michalak, Andrew J. DuBois, Curtis B. Storlie, Heather M. Quinn, William N. Rust, David H. DuBois, David G. Modl, Andrea Manuzzato, and Sean P. Blanchard. Assessment of the impact of cosmic-ray-induced neutrons on hardware in the roadrunner supercomputer. *IEEE Transactions on Device and Materials Reliability*, 12(2):445–454, 2012. doi: 10.1109/TDMR.2012.2192736.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1260. URL https://aclanthology.org/D18-1260.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1206. URL https://aclanthology.org/D18-1206.

Oktatási Hivatal. Matematika írásbéli vizsga. Középszintű Írásbéli Vizsga, May 2023. URL https://dload-oktatas.educatio.hu/erettsegi/feladatok_2023tavasz_kozep/k_matang_23maj_fl.pdf. Angol Nyelven.

OpenAI. GPT-4 Technical Report. 2023a.

OpenAI. GPT-4V(ision) System Card, 2023b.

OpenAI. Whisper, 2023. URL https://github.com/openai/whisper.

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. *Preprint*, 2022. URL https://cdn.openai.com/papers/Training_language_models_to_follow_instructions_with_human_feedback.pdf.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE, 2015.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Quan Ngoc Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. The LAMBADA dataset: Word prediction requiring a broad discourse context. *arXiv preprint arXiv:1606.06031*, 2016.

Viorica Pătrăucean, Lucas Smaira, Ankush Gupta, Adrià Recasens Continente, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Joseph Heyward, Mateusz Malinowski, Yi Yang, et al. Perception test: A diagnostic benchmark for multimodal video models. *arXiv preprint arXiv:2305.13786*, 2023.

Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, et al. Check your facts and try again: Improving large language models with external knowledge and automated feedback. *arXiv preprint arXiv:2302.12813*, 2023.

Leon Poutievski, Omid Mashayekhi, Joon Ong, Arjun Singh, Mukarram Tariq, Rui Wang, Jianan Zhang, Virginia Beauregard, Patrick Conner, Steve Gribble, et al. Jupiter evolving: transforming google's datacenter network via optical circuit switches and software-defined networking. In *Proceedings of the ACM SIGCOMM 2022 Conference,* pages 66–85, 2022.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI blog,* 1(8):9, 2019. URL https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning,* pages 28492–28518. PMLR, 2023.

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, H. Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant M. Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew Johnson, Blake A. Hechtman, Laura Weidinger, Iason Gabriel, William S. Isaac, Edward Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. Scaling language models: Methods, analysis & insights from training Gopher. *CoRR,* abs/2112.11446, 2021.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning,* pages 8821–8831. PMLR, 2021.

Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. Measuring attribution in natural language generation models. *Computational Linguistics*, pages 1–64, 2023.

Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gomez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, Mai Gimenez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, et al. A generalist agent. *arXiv preprint arXiv:2205.06175,* 2022.

Parker Riley, Timothy Dozat, Jan A Botha, Xavier Garcia, Dan Garrette, Jason Riesa, Orhan Firat, and Noah Constant. Frmt: A benchmark for few-shot region-aware machine translation. *Transactions of the Association for Computational Linguistics,* 2023.

Hannah Ritchie, Veronika Samborska, and Max Roser. Plastic pollution. *Our World in Data*, 2023. https://ourworldindata.org/plastic-pollution.

Adam Roberts, Colin Raffel, and Noam Shazeer. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.437. URL https://aclanthology.org/2020.emnlp-main.437.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.704. URL https://aclanthology.org/2020.acl-main.704.

Uri Shaham, Elad Segal, Maor Ivgi, Avia Efrat, Ori Yoran, Adi Haviv, Ankit Gupta, Wenhan Xiong, Mor Geva, Jonathan Berant, and Omer Levy. SCROLLS: Standardized CompaRison over long language sequences. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 12007–12021, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.emnlp-main.823.

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

Toby Shevlane, Sebastian Farquhar, Ben Garfinkel, Mary Phuong, Jess Whittlestone, Jade Leung, Daniel Kokotajlo, Nahema Marchal, Markus Anderljung, Noam Kolt, et al. Model evaluation for extreme risks. *arXiv preprint arXiv:2305.15324*, 2023.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, et al. Language models are multilingual chain-of-thought reasoners. *ICLR*, 2023.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards VQA models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*, 2022. URL https://arxiv.org/abs/2206.04615.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27, 2014.

Oyvind Tafjord, Bhavana Dalvi, and Peter Clark. ProofWriter: Generating implications, proofs, and abductive statements over natural language. In *Findings*, 2020. URL https://api.semanticscholar.org/CorpusID:229371222.

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

Ashish V. Thapliyal, Jordi Pont-Tuset, Xi Chen, and Radu Soricut. Crossmodal-3600: A massively multilingual multimodal evaluation dataset. In *EMNLP*, 2022.

Kocmi Tom, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, et al. Findings of the 2023 conference on machine translation (wmt23): Llms are here but not quite there yet. In *WMT23-Eighth Conference on Machine Translation*, pages 198–216, 2023.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023a.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017. URL http://arxiv.org/abs/1706.03762.

Petar Veličković, Adrià Puigdomènech Badia, David Budden, Razvan Pascanu, Andrea Banino, Misha Dashevskiy, Raia Hadsell, and Charles Blundell. The clrs algorithmic reasoning benchmark. *arXiv preprint arXiv:2205.15659*, 2022.

Manoj Vishwanathan, Ronak Shah, Kyung Ki Kim, and Minsu Choi. Silent data corruption (sdc) vulnerability of gpu on various gpgpu workloads. In *2015 International SoC Design Conference (ISOCC)*, pages 11–12, 2015. doi: 10.1109/ISOCC.2015.7401681.

Changhan Wang, Anne Wu, and Juan Pino. Covost 2 and massively multilingual speech-to-text translation. *arXiv preprint arXiv:2007.10310*, 2020.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. *arXiv preprint arXiv:2101.00390*, 2021.

Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. VATEX: A large-scale, high-quality multilingual dataset for video-and-language research. In *ICCV*, 2019.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

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. *NeurIPS*, 2022. URL https://arxiv.org/abs/2201.11903.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William S. Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. Ethical and social risks of harm from language models. *CoRR*, abs/2112.04359, 2021. URL https://arxiv.org/abs/2112.04359.

David Wetherall, Abdul Kabbani, Van Jacobson, Jim Winget, Yuchung Cheng, Brad Morrey, Uma Parthavi Moravapalle, Phillipa Gill, Steven Knight, and Amin Vahdat. Improving network

availability with protective reroute. In *SIGCOMM 2023*, 2023. URL https://dl.acm.org/doi/10.1145/3603269.3604867.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. NExT-QA: Next phase of question-answering to explaining temporal actions. In *CVPR*, 2021.

XLA. XLA: Optimizing compiler for TensorFlow. https://www.tensorflow.org/xla, 2019. [Online; accessed December-2023].

Yuanzhong Xu, HyoukJoong Lee, Dehao Chen, Blake Hechtman, Yanping Huang, Rahul Joshi, Maxim Krikun, Dmitry Lepikhin, Andy Ly, Marcello Maggioni, et al. Gspmd: general and scalable parallelization for ml computation graphs. *arXiv preprint arXiv:2105.04663*, 2021.

Chi yao Hong, Subhasree Mandal, Mohammad A. Alfares, Min Zhu, Rich Alimi, Kondapa Naidu Bollineni, Chandan Bhagat, Sourabh Jain, Jay Kaimal, Jeffrey Liang, Kirill Mendelev, Steve Padgett, Faro Thomas Rabe, Saikat Ray, Malveeka Tewari, Matt Tierney, Monika Zahn, Jon Zolla, Joon Ong, and Amin Vahdat. B4 and after: Managing hierarchy, partitioning, and asymmetry for availability and scale in google's software-defined wan. In *SIGCOMM'18*, 2018. URL https://conferences.sigcomm.org/sigcomm/2018/program_tuesday.html.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models, 2022a.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5, 2022b.

Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. Self-chained image-language model for video localization and question answering. *arXiv preprint arXiv:2305.06988*, 2023.

Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. ActivityNet-QA: A dataset for understanding complex web videos via question answering. In *AAAI*, 2019.

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, 2023.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*, 2019.

Yu Zhang, Wei Han, James Qin, Yongqiang Wang, Ankur Bapna, Zhehuai Chen, Nanxin Chen, Bo Li, Vera Axelrod, Gary Wang, et al. Google usm: Scaling automatic speech recognition beyond 100 languages. *arXiv preprint arXiv:2303.01037*, 2023.

Chuanyang Zheng, Zhengying Liu, Enze Xie, Zhenguo Li, and Yu Li. Progressive-hint prompting improves reasoning in large language models, 2023.

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

Luowei Zhou, Chenliang Xu, and Jason J Corso. Towards automatic learning of procedures from web instructional videos. In *AAAI Conference on Artificial Intelligence*, pages 7590–7598, 2018.

8. Contributions and Acknowledgments

Core Contributors

Chao Jia

Chenjie Gu

Qiao Zhang

Jordan Grimstad

Ale Jakse Hartman

Martin Chadwick

Gaurav Singh Tomar

Xavier Garcia

Evan Senter

Emanuel Taropa

Thanumalayan Sankaranarayana Pillai

Jacob Devlin

Michael Laskin

Diego de Las Casas

Dasha Valter

Connie Tao

Lorenzo Blanco

Adrià Puigdomènech Badia

David Reitter

Mianna Chen

Jenny Brennan

Clara Rivera

Sergey Brin

Shariq Iqbal

Gabriela Surita

Jane Labanowski

Abhi Rao

Stephanie Winkler

Emilio Parisotto

Yiming Gu

Kate Olszewska

Yujing Zhang

Ravi Addanki

Antoine Miech

Annie Louis

Laurent El Shafey

Denis Teplyashin

Geoff Brown

Elliot Catt

Nithya Attaluri

Jan Balaguer

Jackie Xiang

Pidong Wang

Zoe Ashwood

Anton Briukhov

Albert Webson

Sanjay Ganapathy

Smit Sanghavi

Core Contributors

Ajay Kannan

Ming-Wei Chang

Axel Stjerngren

Josip Djolonga

Yuting Sun

Ankur Bapna

Matthew Aitchison

Pedram Pejman

Henryk Michalewski

Tianhe Yu

Cindy Wang

Juliette Love

Junwhan Ahn

Dawn Bloxwich

Kehang Han

Peter Humphreys

Thibault Sellam

James Bradbury

Varun Godbole

Sina Samangooei

Bogdan Damoc

Alex Kaskasoli

Sébastien M. R. Arnold

Vijay Vasudevan

Shubham Agrawal

Jason Riesa

Dmitry Lepikhin

Richard Tanburn

Srivatsan Srinivasan

Hyeontaek Lim

Sarah Hodkinson

Pranav Shyam

Johan Ferret

Steven Hand

Ankush Garg

Tom Le Paine

Jian Li

Yujia Li

Minh Giang

Alexander Neitz

Zaheer Abbas

Sarah York

Machel Reid

Elizabeth Cole

Aakanksha Chowdhery

Dipanjan Das

Dominika Rogozińska

Vitaly Nikolaev

Core Contributors

Pablo Sprechmann
Zachary Nado
Lukas Zilka
Flavien Prost
Luheng He
Marianne Monteiro
Gaurav Mishra
Chris Welty
Josh Newlan
Dawei Jia
Miltiadis Allamanis
Clara Huiyi Hu
Raoul de Liedekerke
Justin Gilmer
Carl Saroufim
Shruti Rijhwani
Shaobo Hou
Disha Shrivastava
Anirudh Baddepudi
Alex Goldin
Adnan Ozturel
Albin Cassirer
Yunhan Xu
Daniel Sohn
Devendra Sachan
Reinald Kim Amplayo
Craig Swanson
Dessie Petrova
Shashi Narayan
Arthur Guez
Siddhartha Brahma
Jessica Landon
Miteyan Patel
Ruizhe Zhao
Kevin Villela
Luyu Wang
Wenhao Jia
Matthew Rahtz
Mai Giménez
Legg Yeung
Hanzhao Lin
James Keeling
Petko Georgiev
Diana Mincu
Boxi Wu
Salem Haykal
Rachel Saputro
Kiran Vodrahalli

Core Contributors

James Qin
Zeynep Cankara
Abhanshu Sharma
Nick Fernando
Will Hawkins
Behnam Neyshabur
Solomon Kim
Adrian Hutter
Priyanka Agrawal
Alex Castro-Ros
George van den Driessche
Tao Wang
Fan Yang
Shuo-yiin Chang
Paul Komarek
Ross McIlroy
Mario Lučić
Guodong Zhang
Wael Farhan
Michael Sharman
Paul Natsev
Paul Michel
Yong Cheng
Yamini Bansal
Siyuan Qiao
Kris Cao
Siamak Shakeri
Christina Butterfield
Justin Chung
Paul Kishan Rubenstein
Shivani Agrawal
Arthur Mensch
Kedar Soparkar
Karel Lenc
Timothy Chung
Aedan Pope
Loren Maggiore
Jackie Kay
Priya Jhakra
Shibo Wang
Joshua Maynez
Mary Phuong
Taylor Tobin
Andrea Tacchetti
Maja Trebacz
Kevin Robinson
Yash Katariya
Sebastian Riedel

Core Contributors

Paige Bailey
Kefan Xiao
Nimesh Ghelani
Lora Aroyo
Ambrose Slone
Neil Houlsby
Xuehan Xiong
Zhen Yang
Elena Gribovskaya
Jonas Adler
Mateo Wirth
Lisa Lee
Music Li
Thais Kagohara
Jay Pavagadhi
Sophie Bridgers
Anna Bortsova
Sanjay Ghemawat
Zafarali Ahmed
Tianqi Liu
Richard Powell
Vijay Bolina
Mariko Iinuma
Polina Zablotskaia
James Besley
Da-Woon Chung
Timothy Dozat
Ramona Comanescu
Xiance Si
Jeremy Greer
Guolong Su
Martin Polacek
Raphaël Lopez Kaufman
Simon Tokumine
Hexiang Hu
Elena Buchatskaya
Yingjie Miao
Mohamed Elhawaty
Aditya Siddhant
Nenad Tomasev
Jinwei Xing
Christina Greer
Helen Miller
Shereen Ashraf
Aurko Roy
Zizhao Zhang
Angelos Filos
Milos Besta

Core Contributors

Rory Blevins
Ted Klimenko
Chih-Kuan Yeh
Soravit Changpinyo
Jiaqi Mu
Oscar Chang
Mantas Pajarskas
Carrie Muir
Vered Cohen
Charline Le Lan
Krishna Haridasan
Amit Marathe
Steven Hansen
Sholto Douglas
Rajkumar Samuel
Mingqiu Wang
Sophia Austin
Chang Lan
Jiepu Jiang
Justin Chiu
Jaime Alonso Lorenzo
Lars Lowe Sjösund
Sébastien Cevey
Zach Gleicher
Thi Avrahami
Anudhyan Boral
Hansa Srinivasan
Vittorio Selo
Rhys May
Kostas Aisopos
Léonard Hussenot
Livio Baldini Soares
Kate Baumli
Michael B. Chang
Adrià Recasens
Ben Caine
Alexander Pritzel
Filip Pavetic
Fabio Pardo
Anita Gergely
Justin Frye
Vinay Ramasesh
Dan Horgan
Kartikeya Badola
Nora Kassner
Subhrajit Roy
Ethan Dyer
Víctor Campos

Core Contributors

Yunhao Tang
Basil Mustafa
Oran Lang
Abhishek Jindal
Sharad Vikram
Zhitao Gong
Sergi Caelles
Ross Hemsley
Gregory Thornton
Fangxiaoyu Feng
Wojciech Stokowiec
Ce Zheng
Phoebe Thacker
Çağlar Ünlü
Zhishuai Zhang
Mohammad Saleh
James Svensson
Max Bileschi
Piyush Patil
Ankesh Anand
Roman Ring
Katerina Tsihlas
Arpi Vezer
Marco Selvi
Toby Shevlane
Mikel Rodriguez
Tom Kwiatkowski
Samira Daruki
Keran Rong
Allan Dafoe
Nicholas FitzGerald
Keren Gu-Lemberg
Mina Khan
Lisa Anne Hendricks
Marie Pellat
Vladimir Feinberg
James Cobon-Kerr
Tara Sainath
Maribeth Rauh
Sayed Hadi Hashemi
Richard Ives
Yana Hasson
YaGuang Li
Eric Noland
Yuan Cao
Nathan Byrd
Le Hou
Thibault Sottiaux

Core Contributors

Michela Paganini
Alexandre Moufarek
Samer Hassan
Kaushik Shivakumar
Joost van Amersfoort
Amol Mandhane
Pratik Joshi
Anirudh Goyal
Matthew Tung
Andrew Brock
Hannah Sheahan
Vedant Misra
Cheng Li
Nemanja Rakićević
Mostafa Dehghani
Fangyu Liu
Sid Mittal
Junhyuk Oh
Seb Noury
Eren Sezener
Fantine Huot
Matthew Lamm
Nicola De Cao
Charlie Chen

Contributors

Gamaleldin Elsayed
Ed Chi
Mahdis Mahdieh
Ian Tenney
Nan Hua
Ivan Petrychenko
Patrick Kane
Dylan Scandinaro
Rishub Jain
Jonathan Uesato
Romina Datta
Adam Sadovsky
Oskar Bunyan
Alex Tomala
Dominik Rabiej
Shimu Wu
John Zhang
Betty Chan
Pam G Rabinovitch
David Steiner
Shirley Chung
Harry Askham

Contributors

Gautam Vasudevan
Edouard Leurent
Ionut Georgescu
Nan Wei
Ivy Zheng
Piotr Stanczyk
Ye Zhang
Subhajit Naskar
Michael Azzam
Christopher Choquette
Matthew Johnson
Adam Paszke
Chung-Cheng Chiu
Jaume Sanchez Elias
Afroz Mohiuddin
Faizan Muhammad
Jin Miao
Andrew Lee
Nino Vieillard
Sahitya Potluri
Jane Park
Elnaz Davoodi
Jiageng Zhang
Jeff Stanway
Drew Garmon
Abhijit Karmarkar
Zhe Dong
Jong Lee
Aviral Kumar
Luowei Zhou
Jonathan Evens
William Isaac
Zhe Chen
Johnson Jia
Anselm Levskaya
Zhenkai Zhu
Chris Gorgolewski
Peter Grabowski
Yu Mao
Alberto Magni
Kaisheng Yao
Javier Snaider
Norman Casagrande
Paul Suganthan
Evan Palmer
Michael Fink
Daniel Andor
Vikas Yadav

Contributors

Geoffrey Irving
Edward Loper
Manaal Faruqui
Isha Arkatkar
Nanxin Chen
Izhak Shafran
Rama Pasumarthi
Nathan Lintz
Anitha Vijayakumar
Lam Nguyen Thiet
Pedro Valenzuela
Cosmin Paduraru
Daiyi Peng
Katherine Lee
Shuyuan Zhang
Somer Greene
Duc Dung Nguyen
Paula Kurylowicz
Sarmishta Velury
Sebastian Krause
Cassidy Hardin
Lucas Dixon
Lili Janzer
Kiam Choo
Ziqiang Feng
Biao Zhang
Achintya Singhal
Tejasi Latkar
Mingyang Zhang
Quoc Le
Elena Allica Abellan
Dayou Du
Dan McKinnon
Natasha Antropova
Tolga Bolukbasi
Orgad Keller
David Reid
Daniel Finchelstein
Maria Abi Raad
Remi Crocker
Peter Hawkins
Robert Dadashi
Colin Gaffney
Sid Lall
Ken Franko
Egor Filonov
Anna Bulanova
Rémi Leblond

Contributors

Luis C. Cobo
Kelvin Xu
Felix Fischer
Jun Xu
Christina Sorokin
Chris Alberti
Chu-Cheng Lin
Colin Evans
Hao Zhou
Alek Dimitriev
Hannah Forbes
Dylan Banarse
Zora Tung
Jeremiah Liu
Mark Omernick
Colton Bishop
Chintu Kumar
Rachel Sterneck
Ryan Foley
Rohan Jain
Swaroop Mishra
Jiawei Xia
Taylor Bos
Geoffrey Cideron
Ehsan Amid
Francesco Piccinno
Xingyu Wang
Praseem Banzal
Petru Gurita
Ada Ma
Hila Noga
Premal Shah
Daniel J. Mankowitz
Alex Polozov
Nate Kushman
Victoria Krakovna
Sasha Brown
MohammadHossein Bateni
Dennis Duan
Vlad Firoiu
Meghana Thotakuri
Tom Natan
Anhad Mohananey
Matthieu Geist
Sidharth Mudgal
Sertan Girgin
Hui Li
Jiayu Ye

Contributors

Ofir Roval
Reiko Tojo
Michael Kwong
James Lee-Thorp
Christopher Yew
Quan Yuan
Sumit Bagri
Danila Sinopalnikov
Sabela Ramos
John Mellor
Abhishek Sharma
Aliaksei Severyn
Jonathan Lai
Kathy Wu
Nanxin Chen
Heng-Tze Cheng
David Miller
Nicolas Sonnerat
Denis Vnukov
Rory Greig
Jennifer Beattie
Emily Caveness
Libin Bai
Julian Eisenschlos
Dalia El Badawy
Alex Korchemniy
Tomy Tsai
Mimi Jasarevic
Weize Kong
Phuong Dao
Zeyu Zheng
Frederick Liu
Fan Yang
Rui Zhu
Mark Geller
Tian Huey Teh
Jason Sanmiya
Evgeny Gladchenko
Nejc Trdin
Andrei Sozanschi
Daniel Toyama
Evan Rosen
Sasan Tavakkol
Linting Xue
Chen Elkind
Oliver Woodman
John Carpenter
George Papamakarios

Contributors

Rupert Kemp
Sushant Kafle
Tanya Grunina
Alice Talbert
Abhimanyu Goyal
Diane Wu
Denese Owusu-Afriyie
Cosmo Du
Chloe Thornton
Jordi Pont-Tuset
Pradyumna Narayana
Jing Li
Saaber Fatehi
John Wieting
Omar Ajmeri
Benigno Uria
Tao Zhu
Yeongil Ko
Laura Knight
Amélie Héliou
Ning Niu
Shane Gu
Chenxi Pang
Dustin Tran
Yeqing Li
Nir Levine
Ariel Stolovich
Norbert Kalb
Rebeca Santamaria-Fernandez
Sonam Goenka
Wenny Yustalim
Robin Strudel
Ali Elqursh
Balaji Lakshminarayanan
Charlie Deck
Shyam Upadhyay
Hyo Lee
Mike Dusenberry
Zonglin Li
Xuezhi Wang
Kyle Levin
Raphael Hoffmann
Dan Holtmann-Rice
Olivier Bachem
Summer Yue
Sho Arora
Christy Koh
Soheil Hassas Yeganeh

Contributors

Siim Põder
Steven Zheng
Francesco Pongetti
Mukarram Tariq
Yanhua Sun
Lucian Ionita
Mojtaba Seyedhosseini
Pouya Tafti
Ragha Kotikalapudi
Zhiyu Liu Anmol Gulati
Jasmine Liu
Xinyu Ye
Bart Chrzaszcz
Lily Wang
Nikhil Sethi
Tianrun Li
Ben Brown
Shreya Singh
Wei Fan
Aaron Parisi
Joe Stanton
Chenkai Kuang
Vinod Koverkathu
Christopher A. Choquette-Choo
Yunjie Li
TJ Lu
Abe Ittycheriah
Prakash Shroff
Pei Sun
Mani Varadarajan
Sanaz Bahargam
Rob Willoughby
David Gaddy
Ishita Dasgupta
Guillaume Desjardins
Marco Cornero
Brona Robenek
Bhavishya Mittal
Ben Albrecht
Ashish Shenoy
Fedor Moiseev
Henrik Jacobsson
Alireza Ghaffarkhah
Morgane Rivière
Zongwei Zhou
Madhavi Yenugula
Dominik Grewe
Anastasia Petrushkina

Contributors

Tom Duerig
Antonio Sanchez
Steve Yadlowsky
Amy Shen
Amir Globerson
Adam Kurzrok
Lynette Webb
Sahil Dua
Dong Li
Preethi Lahoti
Surya Bhupatiraju
Dan Hurt
Haroon Qureshi
Ananth Agarwal
Tomer Shani
Matan Eyal
Anuj Khare Shreyas Rammohan Belle
Lei Wang
Chetan Tekur
Mihir Sanjay Kale
Jinliang Wei
Ruoxin Sang
Brennan Saeta
Tyler Liechty
Yi Sun
Yao Zhao
Stephan Lee
Pandu Nayak
Doug Fritz
Manish Reddy Vuyyuru
John Aslanides
Nidhi Vyas
Martin Wicke
Xiao Ma
Taylan Bilal
Evgenii Eltyshev
Daniel Balle
Nina Martin
Hardie Cate
Pratik Joshi
James Manyika
Keyvan Amiri
Yelin Kim

Contributors

Mandy Guo
Austin Waters
Oliver Wang
Joshua Ainslie
Jason Baldridge
Han Zhang
Garima Pruthi
Jakob Bauer
Feng Yang
Hongkun Yu
Anthony Urbanowicz
Jennimaria Palomaki
Chrisantha Fernando
Kevin Brooks
Ken Durden
Nikola Momchev
Elahe Rahimtoroghi
Maria Georgaki
Amit Raul
Morgan Redshaw
Jinhyuk Lee
Komal Jalan
Dinghua Li
Ginger Perng
Blake Hechtman
Parker Schuh
Milad Nasr
Mia Chen
Kieran Milan
Vladimir Mikulik
Trevor Strohman
Juliana Franco

Program Leads

Demis Hassabis
Koray Kavukcuoglu

Overall Technical Leads (equal contribution)
Jeffrey Dean
Oriol Vinyals

9. Appendix

9.1. Chain-of-Thought Comparisons on MMLU benchmark

We contrast several chain-of-thought approaches on MMLU and discuss their results in this section. We proposed a new approach where model produces k chain-of-thought samples, selects the majority vote if the model is confident above a threshold, and otherwise defers to the greedy sample choice. The thresholds are optimized for each model based on their validation split performance. The proposed approach is referred to as *uncertainty-routed chain-of-thought*. The intuition behind this approach is that chain-of-thought samples might degrade performance compared to the maximum-likelihood decision when the model is demonstrably inconsistent. We compare the gains from the proposed approach on both Gemini Ultra and GPT-4 in Figure 7. We find that Gemini Ultra benefits more from this approach compared to using only chain-of-thought samples. GPT-4's performance improves from 84.2% with greedy sampling to 87.3% with uncertainty-routed chain-of-thought approach with 32 samples, but it already achieves these gains from using 32 chain-of-thought samples. In contrast, Gemini Ultra improves its performance significantly from 84.0% with greedy sampling to 90.0% with uncertainty-routed chain-of-thought approach with 32 samples while it marginally improves to 85.0% with the use of 32 chain-of-thought samples only.
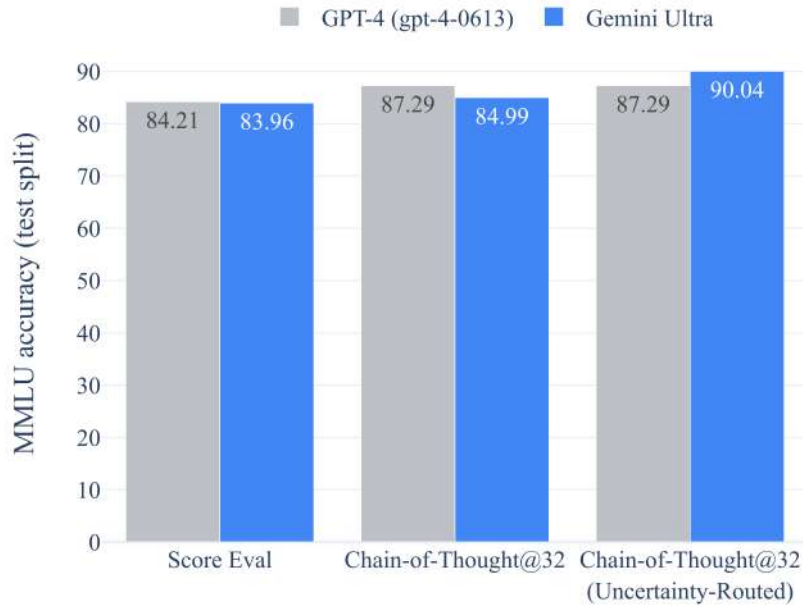

Figure 7 | Chain-of-Thought with uncertainty routing on MMLU.

9.2. Capabilities and Benchmarking Tasks

We use more than 50 benchmarks as a holistic harness to evaluate the Gemini models across text, image, audio and video. We provide a detailed list of benchmarking tasks for six different capabilities in text understanding and generation: factuality, long context, math/science, reasoning, summarization, and multilinguality. We also enumerate the benchmarks used for image understanding, video understanding, and audio understanding tasks.
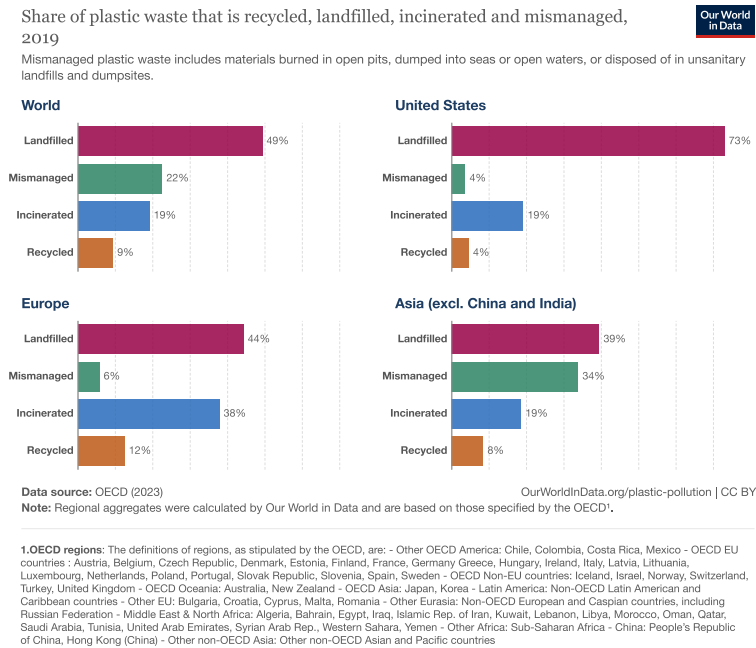
- **Factuality**: We use 5 benchmarks: BoolQ (Clark et al., 2019), NaturalQuestions-Closed (Kwiatkowski et al., 2019), NaturalQuestions-Retrieved (Kwiatkowski et al., 2019), RealtimeQA (Kasai et al., 2022), TydiQA-noContext and TydiQA-goldP (Clark et al., 2020).
- **Long Context**: We use 6 benchmarks: NarrativeQA (Kočiský et al., 2018), Scrolls-Qasper, Scrolls-Quality (Shaham et al., 2022), XLsum (En), XLSum (non-English languages) (Hasan et al., 2021), and one other internal benchmark.
- **Math/Science**: We use 8 benchmarks: GSM8k (with CoT) (Cobbe et al., 2021), Hendryck's MATH pass@1 (Hendrycks et al., 2021b), MMLU (Hendrycks et al., 2021a), Math-StackExchange, Math-AMC 2022-2023 problems, and three other internal benchmarks.
- **Reasoning**: We use 7 benchmarks: BigBench Hard (with CoT) (Srivastava et al., 2022), CLRS (Veličković et al., 2022), Proof Writer (Tafjord et al., 2020), Reasoning-Fermi problems (Kalyan et al., 2021), Lambada (Paperno et al., 2016), HellaSwag (Zellers et al., 2019), DROP (Dua et al., 2019).
- **Summarization**: We use 5 benchmarks: XL Sum (English), XL Sum (non-English languages) (Hasan et al., 2021), WikiLingua (non-English languages), WikiLingua (English) (Ladhak et al., 2020), XSum (Narayan et al., 2018).
- **Multilinguality**: We use 10 benchmarks: XLSum (Non-English languages) (Hasan et al., 2021), WMT22 (Kocmi et al., 2022), WMT23 (Tom et al., 2023), FRMT (Riley et al., 2023), WikiLingua (Non-English languages) (Ladhak et al., 2020), TydiQA (no context), TydiQA (GoldP) (Clark et al., 2020), MGSM (Shi et al., 2023), translated MMLU (Hendrycks et al., 2021a), NTREX (Federmann et al., 2022), FLORES-200 (Team et al., 2022).
- **Image and Video**: We use 9 benchmarks for image understanding: MMMU (Yue et al., 2023), TextVQA (Singh et al., 2019), DocVQA (Mathew et al., 2021), ChartQA (Masry et al., 2022), InfographicVQA (Mathew et al., 2022), MathVista (Lu et al., 2023), AI2D (Kembhavi et al., 2016), VQAv2 (Goyal et al., 2017), XM3600 (Thapliyal et al., 2022) for multi-lingual image understanding, and 6 benchmarks for video understanding: VATEX (Wang et al., 2019) for captioning in two different languages, YouCook2 (Zhou et al., 2018), NextQA (Xiao et al., 2021), ActivityNet-QA (Yu et al., 2019), and Perception Test MCQA (Pătrăucean et al., 2023).
- **Audio**: We use 5 benchmarks including automatic speech recognition (ASR) tasks such as FLEURS (Conneau et al., 2023), VoxPopuli, (Wang et al., 2021), Multi-lingual Librispeech (Panayotov et al., 2015), and automatic speech translation task such as CoVoST 2 (Wang et al., 2020).

9.3. Qualitative Examples

This section shows sample qualitative examples from prompting the Gemini Ultra model. Some illustrative examples of multimodal reasoning for image understanding tasks over charts, natural images and memes are shown in Figures 8, 9, 11, 13, 14, and 15. Figure 10 shows an example of image generation capabilities of Gemini Ultra where the user generates an interleaved sequence of image and text to design a blog post. Beyond English, Figure 16 shows model's capability to understand images in a multilingual setting. Gemini models also show strong performance on multimodal image understanding and reasoning in mathematics, as shown in Figures 12, 18 and 19. Figure 20 is an example of complex multimodal reasoning demonstrating how the model composes complex image understanding, code generation, and instruction following capabilities for a given user task. In Figure 17, we see another example of the model being able to generate working code and follow complex user instructions. Finally, Figure 21 shows an example of Gemini Ultra's capability of understanding video by reasoning over temporally connected set of frames.

9.3.1. Chart understanding and reasoning over data

Prompt

Share of plastic waste that is recycled, landfilled, incinerated and mismanaged, 2019

Mismanaged plastic waste includes materials burned in open pits, dumped into seas or open waters, or disposed of in unsanitary landfills and dumpsites.

World

Landfilled	49%
Mismanaged	22%
Incinerated	19%
Recycled	9%

United States

Landfilled	73%
Mismanaged	4%
Incinerated	19%
Recycled	4%

Europe

Landfilled	44%
Mismanaged	6%
Incinerated	38%
Recycled	12%

Asia (excl. China and India)

Landfilled	39%
Mismanaged	34%
Incinerated	19%
Recycled	8%

Data source: OECD (2023)　　　　　　OurWorldInData.org/plastic-pollution | CC BY
Note: Regional aggregates were calculated by Our World in Data and are based on those specified by the OECD[1].

1.OECD regions: The definitions of regions, as stipulated by the OECD, are: - Other OECD America: Chile, Colombia, Costa Rica, Mexico - OECD EU countries : Austria, Belgium, Czech Republic, Denmark, Estonia, Finland, France, Germany Greece, Hungary, Ireland, Italy, Latvia, Lithuania, Luxembourg, Netherlands, Poland, Portugal, Slovak Republic, Slovenia, Spain, Sweden - OECD Non-EU countries: Iceland, Israel, Norway, Switzerland, Turkey, United Kingdom - OECD Oceania: Australia, New Zealand - OECD Asia: Japan, Korea - Latin America: Non-OECD Latin American and Caribbean countries - Other EU: Bulgaria, Croatia, Cyprus, Malta, Romania - Other Eurasia: Non-OECD European and Caspian countries, including Russian Federation - Middle East & North Africa: Algeria, Bahrain, Egypt, Iraq, Islamic Rep. of Iran, Kuwait, Lebanon, Libya, Morocco, Oman, Qatar, Saudi Arabia, Tunisia, United Arab Emirates, Syrian Arab Rep., Western Sahara, Yemen - Other Africa: Sub-Saharan Africa - China: People's Republic of China, Hong Kong (China) - Other non-OECD Asia: Other non-OECD Asian and Pacific countries

Spot a data point that stands out in these charts and what that implicates. Then produce a detailed markdown table for all the data shown.

Model Response (rendered Markdown)

The data point that stands out is the United States' landfilled plastic waste, which is 73%. This is significantly higher than the other countries and regions shown. This implies that the United States is not doing as much to recycle or incinerate its plastic waste as other countries and regions.

Country/Region	Landfilled (%)	Mismanaged (%)	Incinerated (%)	Recycled (%)
World	49	22	19	9
United States	73	4	19	4
Europe	44	6	38	12
Asia (excl. China and India)	39	34	19	8

Figure 8 | Solving a problem requiring multimodal chart understanding.The model has to read the text, understand the connections between different data points and reason over them to recommend an interesting point and follow the instructions to generate a markdown table (shown correctly rendered).
Source: Our World In Data (Ritchie et al., 2023).

9.3.2. Multimodal question answering

Prompt

Do you know what it this plant? How do I best take care of it?

Model Response

This is a Persian shield plant. It is a tropical plant that is native to Southeast Asia. It is a popular houseplant because it is easy to care for and has beautiful purple leaves.

Persian shield plants prefer bright, indirect light. They can tolerate some direct sunlight, but too much direct sunlight can scorch their leaves. They also prefer warm temperatures and high humidity.

Persian shield plants need to be watered regularly, but they should not be allowed to sit in water. The soil should be kept moist but not soggy. They also benefit from being misted regularly.

Persian shield plants can be fertilized every two weeks during the growing season. A balanced fertilizer is a good choice.

Persian shield plants are susceptible to pests, such as mealybugs and spider mites. These pests can be controlled with insecticidal soap or neem oil.

Persian shield plants can be propagated by stem cuttings. The cuttings should be taken in the spring or summer and should be rooted in water or a moist potting mix.

Figure 9 | Answering a multimodal information-seeking query. The model is able to recognize the specific plant shown in the image and provide information about it. The model shows robustness to typos as it is able to understand the user question despite them.
Source: photo taken by an author from the Gemini team.

9.3.3. Interleaved image and text generation

Prompt

Please create a blog post about a trip to New York, where a dog and his owner had lots of fun. Include a few pictures of the dog posing happily at different landmarks.

Model Response

Hi, I'm Cooper. I'm a 2 year old Goldendoodle who loves to travel. Last month, my family and I went to New York City. It was my first time in the Big Apple, and I had a blast!

We stayed in a hotel near Times Square. It was so much fun to see all the bright lights and people. We went to see a Broadway show, and I even got to meet some of the actors.

One of my favorite parts of New York City was Central Park. It was so big and beautiful. I loved running around and playing fetch. We also went to the Statue of Liberty and the Empire State Building. I was so excited to see all the famous landmarks.

I had so much fun in New York City. It's definitely one of my favorite places to travel. I can't wait to go back someday!

Figure 10 | Generating interleaved text and images. The model is able to follow the instructions of generating a blog post with images closely related to the text and with dog images showing high levels of consistency across all images.

9.3.4. Image understanding and reasoning

Prompt
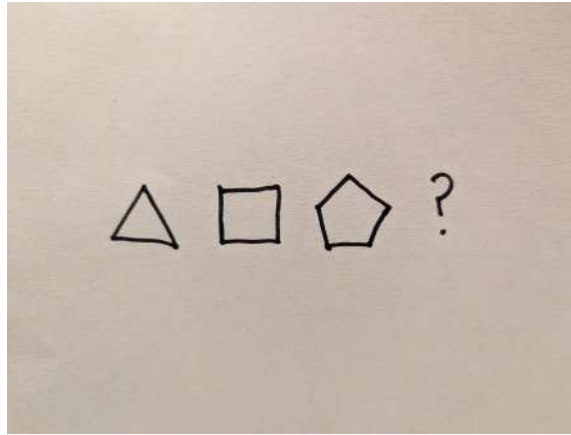
Look at this sequence of three shapes. What shape should come as the fourth shape? Explain your reasoning with detailed descriptions of the first shapes.

Model Response

The fourth shape should be a hexagon. The first shape is a triangle, the second shape is a square, and the third shape is a pentagon. The number of sides in each shape is increasing by one

---

Figure 11 | Solving a multimodal reasoning problem.The model is able to recognize shapes in the image, understand their properties and reason about the relationship between them to predict the next object.
Source: photo taken by an author from the Gemini team.

9.3.5. Geometrical reasoning

Prompt

Find the height of the parallelogram given its area with 100 square units.

Model Response

The area of the parallelogram is equal to the product of the base and the height. Hence $100 = (x + 15)x$. We get $x^2 + 15x - 100 = 0$. The roots of this equation are $x_1 = 5$ and $x_2 = -20$. The second root is negative and is not acceptable. Final Answer: The final answer is 5. I hope it is correct.

Figure 12 | Solving a geometrical reasoning task. The model shows good understanding of the task and is able to provide meaningful reasoning steps despite slightly unclear instructions.
Source: Lu et al. (2021).

9.3.6. Information seeking about objects

Prompt

Find a connection between these. Hint: think about historical events.

Model Response

The Moon is the only celestial body in which humans have played golf. In 1971, the Apollo 14 crew hit two golf balls on the lunar surface.

Figure 13 | Solving a puzzle using multimodal inputs. The model recognizes the objects in the images and identifies a commonality that connects the two objects.
Source: photo taken by an author from the Gemini team.

9.3.7. Multimodal reasoning based on visual cues

Prompt

Do you know the precise location where this image was taken?

Model Response

This image was taken in New York City. The Empire State Building is visible in the background. The street in the foreground is 8th Avenue. The cross street is West 34th Street.

Figure 14 | Identifying the objects in the image (the Empire State Building) and recognizing what those are even with small levels of visual distortion in the image. Based on the image, the model is also able to correctly identify the precise location of the person taking the photo.
Source: photo taken by an author from the Gemini team.

9.3.8. Multimodal humor understanding

Prompt

Can you explain this meme?

Model Response

The meme is about a person who is playing a game at 300 FPS, but their monitor is only 75Hz. The person is trying to catch the game, but it is too fast for their monitor to display. The meme is funny because it is relatable to anyone who has ever played a game on a monitor that is not fast enough.

Figure 15 | Explanation of humor in a meme. The model is showing the ability to not only describe what is happening in the image but also what it means even though the cultural context is not mentioned explicitly in the image or the prompt.
Source: Hwang and Shwartz (2023).

9.4. Commonsense reasoning in a multilingual setting

Prompt
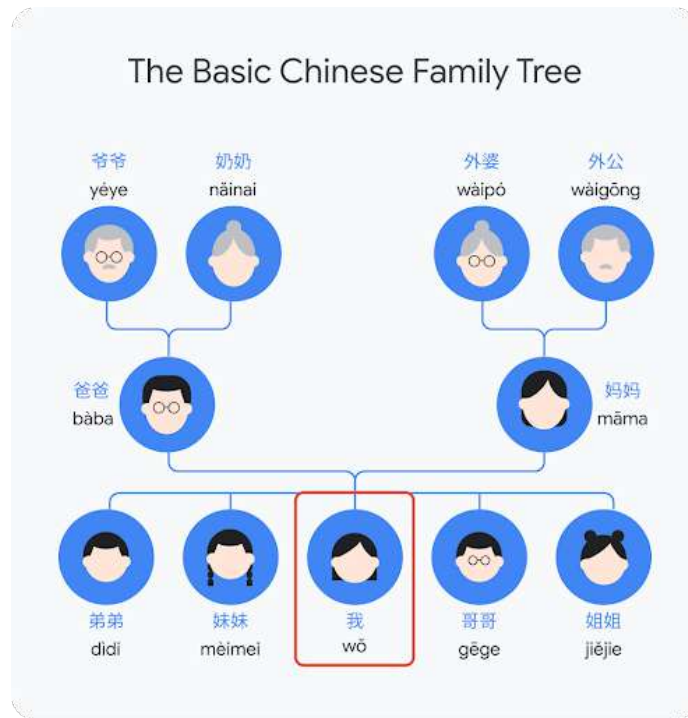
Explain what i should call my grandparents on each side of my family based on this chart

Model Response

On your father's side, you would call your grandparents Yéye and Nǎinai. On your mother's side, you would call your grandparents Wàigōng and Wàipó.

Figure 16 | Common-sense reasoning in images. The model is able to understand the relationships represented in the graphs and reason about them in a multilingual setting.
Source: image created by an author from the Gemini team.

9.4.1. Reasoning and code generation

Prompt

Create a web app called "Opossum Search":

1. Every time you make a search query, it should redirect you to a google search with the same query, but the word opossum before it.
2. It should be visually similar to Google search,
3. Instead of the google logo, it should have a picture of an opossum from the internet.
4. It should be a single html file, no separate js or css files.
5. It should say "Powered by google search" in the footer.

Model Response (rendered code)

Here is the code for Opossum Search:

```html
1  <!DOCTYPE html>
2  <html>
3  <head>
4    <title>Opossum Search</title>
5    <style>
6      body {
7        font-family: Arial, sans-serif;
8        margin: 0;
9        padding: 0;
10     }
11
12     header {
13       background-color: #333;
14       color: #fff;
15       text-align: center;
16       padding: 1rem 0;
17     }
18
19     .search-bar {
20       display: flex;
21       justify-content: center;
22       align-items: center;
23       height: 50px;
24       margin-top: 1rem;
25     }
26
27     .search-input {
28       width: 50%;
29       padding: 0.5rem;
30       border: 1px solid #ccc;
31       border-radius: 5px;
32     }
33
34     .search-button {
35       margin-left: 1rem;
36       padding: 0.5rem 1rem;
37       border: 1px solid #ccc;
38       border-radius: 5px;
39       cursor: pointer;
40     }
41
42     footer {
43       background-color: #333;
44       color: #fff;
45       text-align: center;
46       padding: 1rem 0;
47       margin-top: 1rem;
48     }
49    </style>
50  </head>
51  <body>
52    <header>
53      <h1>Opossum Search</h1>
54      <img src="https://upload.wikimedia.org/wikipedia/commons/thumb/a/a4/Didelphis_virginiana.jpg/
55      1200px-Didelphis_virginiana.jpg" alt="Opossum" width="100">
56    </header>
57
58    <div class="search-bar">
59      <input type="text" class="search-input" placeholder="Search the web">
60      <button class="search-button">Search</button>
61    </div>
```

```
62    <footer>
63      Powered by Google Search
64    </footer>
65
66    <script>
67      const searchInput = document.querySelector('.search−input');
68      const searchButton = document.querySelector('.search−button');
69
70      searchButton.addEventListener('click',  () => {
71        const query = searchInput.value;
72        if (query) {
73          window.location.href = `https://www.google.com/search?q=opossum+${query}`;
74        }
75      });
76    </script>
77  </body>
```

This code creates a simple web app that looks similar to Google Search, but with an opossum logo. When you enter a search query and click the "Search" button, it will redirect you to a Google search with the word "opossum" added to the beginning of your query. The app is powered by Google Search, as indicated in the footer.

Rendered Website

Figure 17 | Writing code for a website based on user request. The model follows the instructions and requirements defined by the user and converts them to functioning code.

9.4.2. Mathematics: Calculus

> **Prompt**
>
> Find the derivative of $\sinh x + \cosh y = x + y$.

> **Model Response (rendered latex)**
>
> We have that $\frac{d}{dx} \sinh x = \cosh x$ and $\frac{d}{dx} \cosh y = \sinh y \frac{dy}{dx}$. Therefore,
>
> $$\cosh x + \sinh y \frac{dy}{dx} = 1 + \frac{dy}{dx}.$$
>
> Solving for $\frac{dy}{dx}$, we find that
>
> $$\frac{dy}{dx} = \frac{1 - \cosh x}{\sinh y - 1}.$$
>
> Final Answer: The final answer is $\frac{1-\cosh x}{\sinh y-1}$.

Figure 18 | Solving a calculus problem. The model is able to get a solution to a calculus problem with step-by-step explanation and correctly defined LaTeX equations.
Source: question is provided by Macmillan Learning.

9.5. Multi-step reasoning and mathematics

Prompt

In the year 2018 factory A produced goods worth 500 million forints, while factory B produced goods worth 400 million forints. According to their long-term plans, factory A will increase production by 5% each year, while factory B will increase production by 6% per year.

The above plans were published on an economy website. A debate started in the comment section of the article: one user claimed that the difference between the volumes of production between the two factories would get smaller and smaller in the next couple of years.

Calculate the missing values of the table below and show that the above claim is not true.

Header: | 2018 | 2019 | 2020 | 2021 |
Production volume of factory A (mln) | 500 | | | |
Production volume of factory B (mln) | 400 | | | |

Model Response (rendered Markdown)

Year	Production volume of factory A (mln)	Production volume of factory B (mln)	Difference (mln)
2018	500	400	100
2019	525	424	101
2020	551.25	449.44	101.81
2021	578.81	476.41	102.4

The production volume of factory A in 2019 is 500 * 1.05 = 525 million forints.
The production volume of factory B in 2019 is 400 * 1.06 = 424 million forints.
The difference in production volume in 2019 is 525 - 424 = 101 million forints.

The production volume of factory A in 2020 is 525 * 1.05 = 551.25 million forints.
The production volume of factory B in 2020 is 424 * 1.06 = 449.44 million forints.
The difference in production volume in 2020 is 551.25 - 449.44 = 101.81 million forints.

The production volume of factory A in 2021 is 551.25 * 1.05 = 578.81 million forints.
The production volume of factory B in 2021 is 449.44 * 1.06 = 476.41 million forints.
The difference in production volume in 2021 is 578.81 - 476.41 = 102.4 million forints.

As you can see, the difference in production volume between the two factories is actually increasing, not decreasing. Therefore, the claim that the difference between the volumes of production between the two factories would get smaller and smaller in the next couple of years is not true.

Figure 19 | Solving a multi-step math problem. The model is able to understand the task and generate a markdown table with correctly calculated values. It also explicitly follows the instructions to show where the numbers come from and answer the question given in the task.
Source: Oktatási Hivatal (2023, p. 20)

9.5.1. Complex image understanding, code generation, and instruction following

Prompt
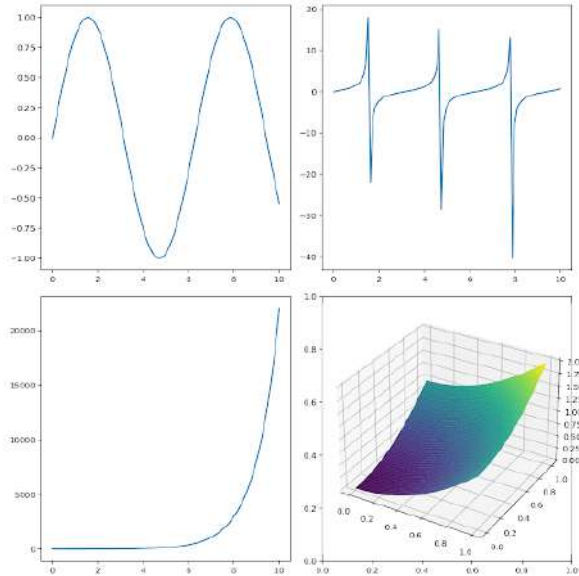
I want you to take the function depicted in the top left subplot, multiply it by 1000, and then add it to the function depicted in the bottom left subplot. Generate matplotlib code for the single resulting plot.

Model Response (rendered code)

```
1  import numpy as np
2  import matplotlib.pyplot as plt
3
4  x = np.linspace(0, 10, 1000)
5  y1 = np.sin(x)
6  y2 = np.exp(x)
7
8  plt.plot(x,  1000*y1 + y2)
9  plt.show()
```
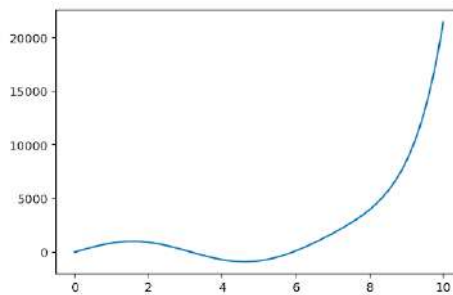
Rendered Graph

Figure 20 | Multimodal reasoning capabilities applied to code generation. Gemini Ultra needs to perform inverse graphics task to infer the code that would have generated the plots, perform additional mathematical transformations, and generate relevant code.
Source: figure generated by an author from the Gemini team.

9.5.2. *Video understanding and reasoning*

Prompt (video)

How could this person improve their technique?

Model Response

This soccer player needs to work on their ball-striking mechanics. Their foot is not making good contact with the ball, which is causing the ball to fly up and away from the goal. They also need to work on their body positioning. They should be leaning into the shot with their non-kicking leg in front of their kicking leg, and they should be following through with their kicking leg.

Figure 21 | Video understanding and reasoning over the situation presented in the video. Here, we provide a video as input to the model together with a text prompt (images are provided here only for visualization purposes). The model is able to analyze what happened in the video and provide recommendations on how the actions in the video could have been better.

Video source: "Football/Soccer Penalty Miss"

https://www.youtube.com/watch?v=VmWxjmJ3mvs