Google DeepMind

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 the Gemini family in cross-modal reasoning and language understanding will enable a wide variety of use cases. We discuss our approach toward post-training and deploying Gemini models responsibly to users through services including Gemini, Gemini Advanced, Google AI Studio, and Cloud Vertex AI.

1. Introduction

We present Gemini, a family of highly capable multimodal models developed at Google. We trained Gemini models 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.

After large-scale pre-training, we post-train our models to improve overall quality, enhance target capabilities, and ensure alignment and safety criteria are met. Due to the varied requirements of our downstream applications, we have produced two post-trained Gemini model family variants. Chat-focused variants, referred to as Gemini Apps models, are optimized for Gemini and Gemini Advanced, our conversational AI service formerly known as Bard. Developer-focused variants, referred to as Gemini API models, are optimized for a range of products and are accessible through Google AI Studio and Cloud Vertex AI.

We evaluate the performance of pre- and post-trained Gemini models on a comprehensive suite of internal and external benchmarks covering a wide range of language, coding, reasoning, and multimodal tasks.

The Gemini family 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

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

(Barham et al., 2022; Bradbury et al., 2018; Dean et al., 2012) that enable large-scale training.

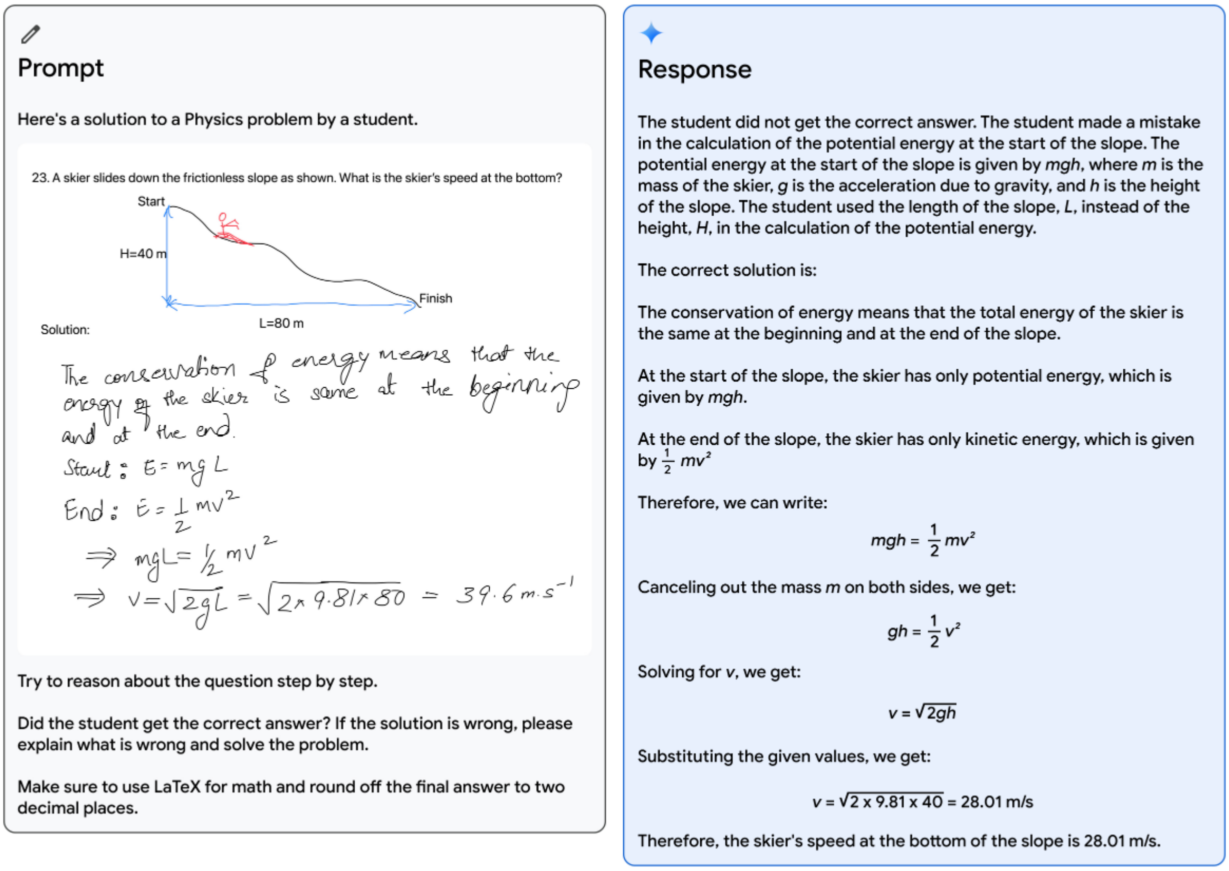
Our most capable model, Gemini Ultra, achieves new state-of-the-art results in 30 of 32 benchmarks we report on, including 10 of 12 popular text and reasoning benchmarks, 9 of 9 image understanding benchmarks, 6 of 6 video understanding benchmarks, and 5 of 5 speech recognition and speech translation 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 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 models' 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.

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-model-powered agent, that combines Gemini models' 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 pre-training dataset. We then present detailed *evaluations* of the pre- and post-trained Gemini model family, covering well-studied benchmarks across text, code, image, audio and video — which include both English performance and multilingual capabilities. Next we discuss our approach to post-training, highlight common and distinct aspects of the Gemini Apps and Gemini API model variants, and benchmark their performance on key capabilities. *Responsible deployment* is critical: we explain our process for impact assessments, developing model policies, evaluations, and mitigations of harm before deployment decisions. Finally, we discuss the broader implications of Gemini models, their limitations alongside their potential applications — paving the way for a new era of research and innovation in AI.



Here's a solution to a Physics problem by a student.

23. A skier slides down the frictionless slope as shown. What is the skier's speed at the bottom?

Solution:

The conservation of energy means that the energy of the skier is same at the beginning and at the end.

Start: $E = mgL$

End: $E = \frac{1}{2} mv^2$

$\Rightarrow mgL = \frac{1}{2} mv^2$

$\Rightarrow v = \sqrt{2gL} = \sqrt{2 \times 9.81 \times 80} = 39.6\ m.s^{-1}$

Try to reason about the question step by step.

Did the student get the correct answer? If the solution is wrong, please explain what is wrong and solve the problem.

Make sure to use LaTeX for math and round off the final answer to two decimal places.

The student did not get the correct answer. The student made a mistake in the calculation of the potential energy at the start of the slope. The potential energy at the start of the slope is given by *mgh*, where *m* is the mass of the skier, *g* is the acceleration due to gravity, and *h* is the height of the slope. The student used the length of the slope, *L*, instead of the height, *H*, in the calculation of the potential energy.

The correct solution is:

The conservation of energy means that the total energy of the skier is the same at the beginning and at the end of the slope.

At the start of the slope, the skier has only potential energy, which is given by *mgh*.

At the end of the slope, the skier has only kinetic energy, which is given by $\frac{1}{2} mv^2$

Therefore, we can write:

$$mgh = \frac{1}{2} mv^2$$

Canceling out the mass *m* on both sides, we get:

$$gh = \frac{1}{2} v^2$$

Solving for *v*, we get:

$$v = \sqrt{2gh}$$

Substituting the given values, we get:

$$v = \sqrt{2 \times 9.81 \times 40} = 28.01 \text{ m/s}$$

Therefore, the skier's speed at the bottom of the slope is 28.01 m/s.

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.

2. Model Architecture

Gemini models build on top of Transformer decoders (Vaswani et al., 2017b) 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, 2019a)). Our first version, Gemini 1.0, comprises three main sizes to support a wide range of applications as discussed in Table 1.

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 tasks that require fine-grained understanding. In addition, Gemini models can directly ingest audio

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.

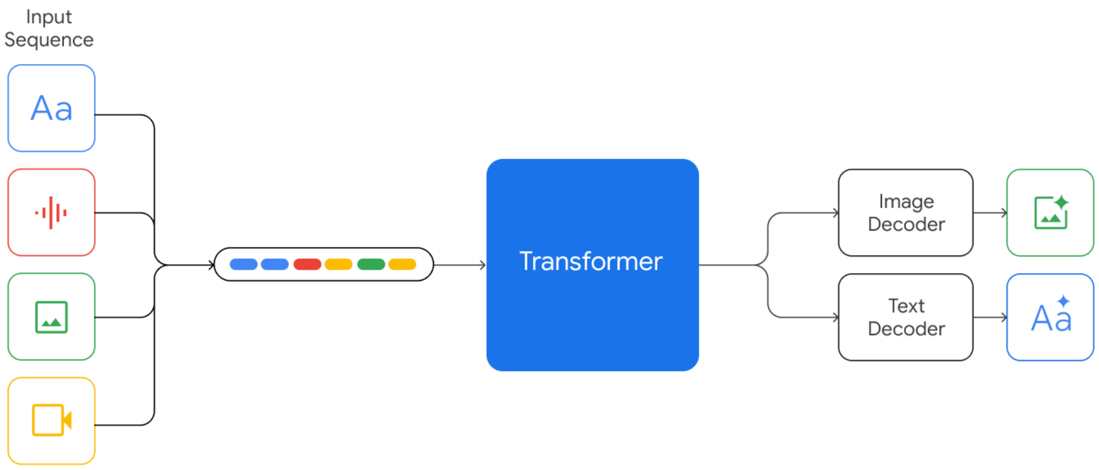


Figure 2 | Gemini models support interleaved sequences of text, image, audio, and video as inputs (illustrated by tokens of different colors in the input sequence). They can output responses with interleaved image and text.

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 pre-training 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 owned by Google 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.

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[2] at this scale would have been impossible using the conventional approach of periodic checkpointing of weights to persistent cluster storage. For Gemini models, 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 models 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. Pre-Training Dataset

Gemini models are trained on a dataset that is both multimodal and multilingual. Our pre-training 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

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

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 based on our policies. To maintain the integrity of evaluations, we search for and remove any evaluation data that may have been in our training corpus before using data for training. 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 an important factor for highly-performing models, and believe that many interesting questions remain around finding the optimal dataset distribution for pre-training.

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 models set a new state of the art across a wide range of text, image, audio, and video benchmarks. ww

5.1. Text

5.1.1. Academic Benchmarks

We compare pre- and post-trained Gemini Pro and Ultra models 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., 2022b) 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%.

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 fine-tuning steps on specific website extracts corresponding to the HellaSwag training set (which were not included in the Gemini model 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 Appendix 10.3 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[3]. 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. Several of these capabilities are targeted by post-training (Section 6). Please see Appendix 10.3 for a detailed list of tasks included for each capability.

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

	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%***
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 (PT****)	67.7% 0-shot (PT****)	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 (PT****)	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 10.2.

** 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.)

**** PT denotes a post-trained Gemini API 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 fine-tuning. Figure 3 shows the performance of these pre-trained 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

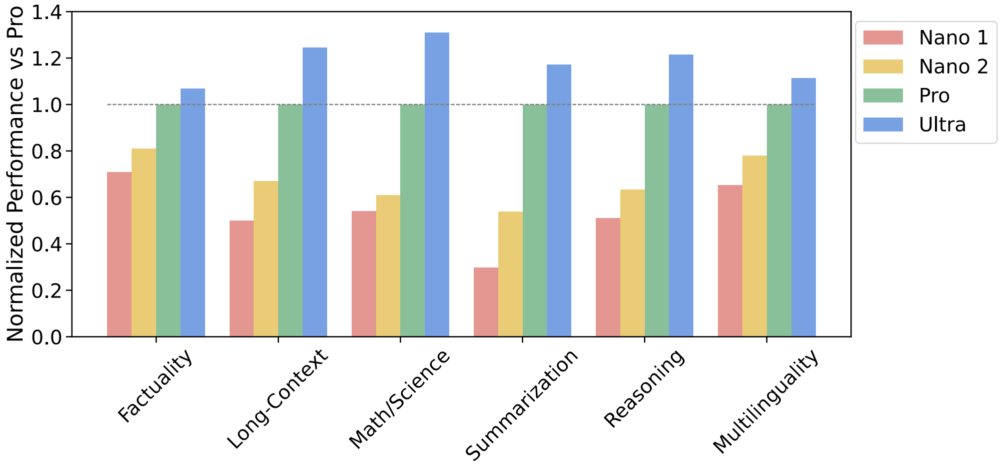


Figure 3 | Language understanding and generation performance of Gemini model family across different capabilities (normalized by the Gemini Pro model).

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).

5.1.4.1 Machine Translation

Translation is a canonical benchmark in machine learning with a rich history. We evaluated a post-trained Gemini API Ultra model (see Section 6.5.3) 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.

5.1.4.2 Multilingual Math and Summarization

Beyond translation, we evaluated how well Gemini models perform 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 models 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