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.

Prompt

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?

Start
H=40 m
Finish
L=80 m

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

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

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

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.

Response

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:

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

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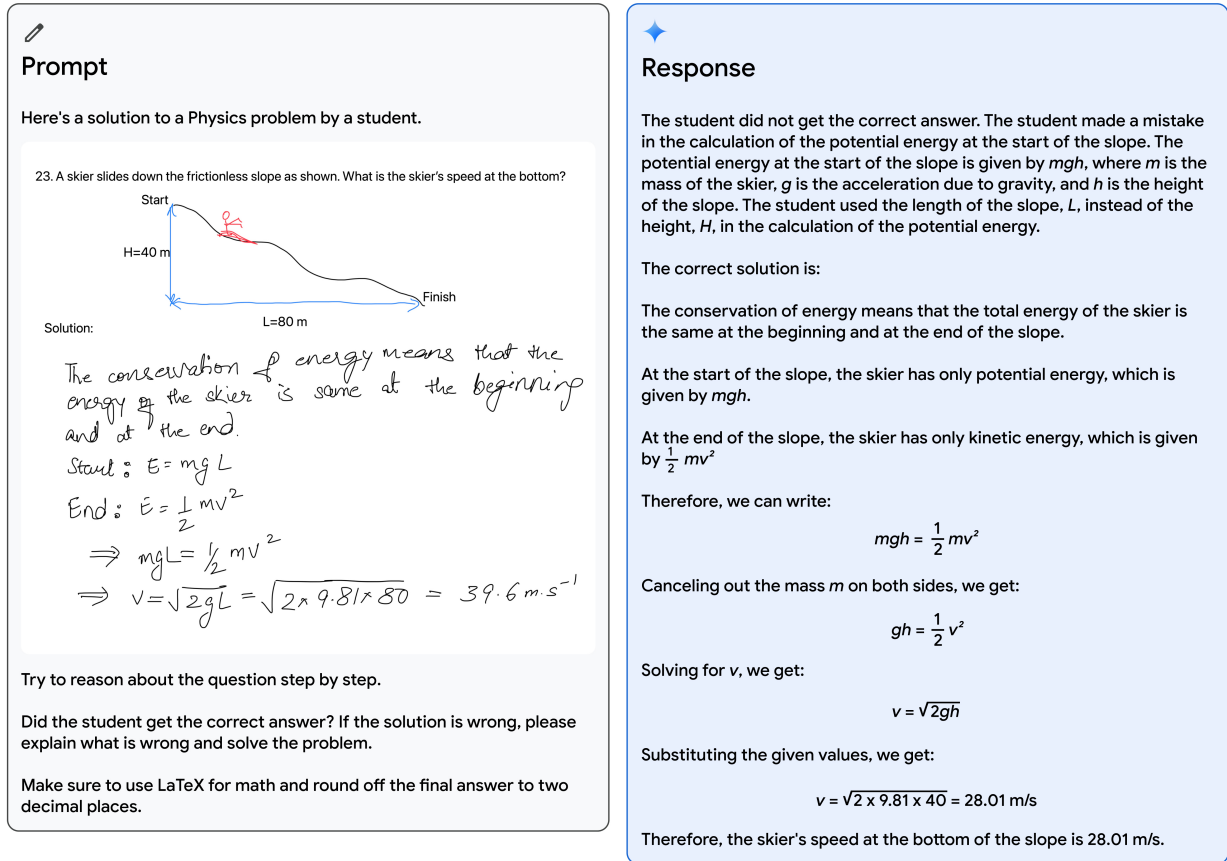

Figure 1 | Verifying a student's solution to a physics problem. The model is able to correctly recognize all of the handwritten content and verify the reasoning. On top of understanding the text in the image, it needs to understand the problem setup and correctly follow instructions to generate LaTeX.

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.

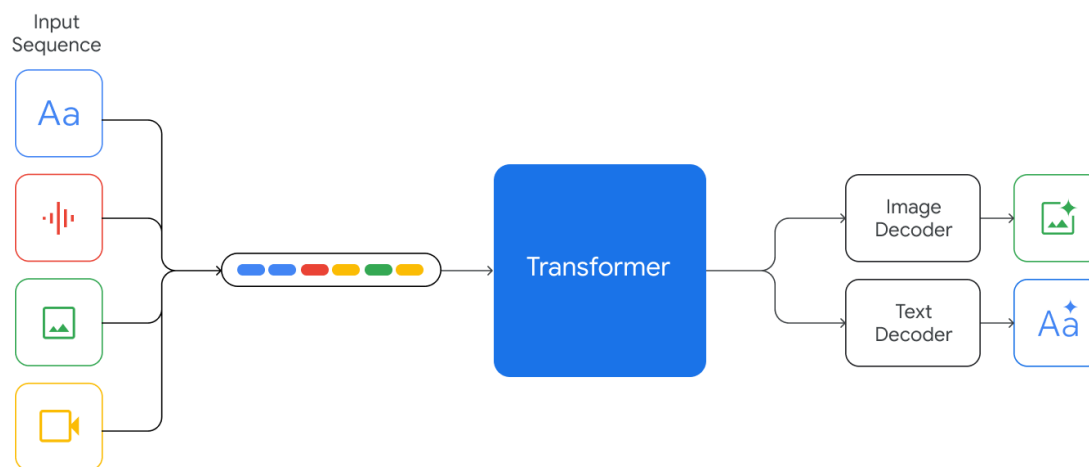Table 1 | An overview of the Gemini 1.0 model family.

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