Google DeepMind

Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities.

Gemini Team, Google

In this report, we introduce the Gemini 2.X model family: Gemini 2.5 Pro and Gemini 2.5 Flash, as well as our earlier Gemini 2.0 Flash and Flash-Lite models. Gemini 2.5 Pro is our most capable model yet, achieving SoTA performance on frontier coding and reasoning benchmarks. In addition to its incredible coding and reasoning skills, Gemini 2.5 Pro is a thinking model that excels at multimodal understanding and it is now able to process up to 3 hours of video content. Its unique combination of long context, multimodal and reasoning capabilities can be combined to unlock new agentic workflows. Gemini 2.5 Flash provides excellent reasoning abilities at a fraction of the compute and latency requirements and Gemini 2.0 Flash and Flash-Lite provide high performance at low latency and cost. Taken together, the Gemini 2.X model generation spans the full Pareto frontier of model capability vs cost, allowing users to explore the boundaries of what is possible with complex agentic problem solving.

1. Introduction

We present our latest family of natively multimodal models with advanced reasoning through thinking, long context and tool-use capabilities: Gemini 2.5 Pro and 2.5 Flash and our earlier Gemini 2.0 Flash and Gemini 2.0 Flash-Lite models. Together these form a new family of highly-capable models representing our next generation of AI models, designed to power a new era of agentic systems. Building upon the foundation of the Gemini 1.5 series (Gemini Team, 2024), this Gemini 2.X generation brings us closer to the vision of a universal AI assistant (Hassabis, 2025).

The Gemini 2.X series are all built to be natively multimodal, supporting long context inputs of >1 million tokens and have native tool use support. This allows them to comprehend vast datasets and handle complex problems from different information sources, including text, audio, images, video and even entire code repositories. These extensive capabilities can also be combined to build complex agentic systems, as happened in the case of Gemini Plays Pokémon[1] (Zhang, 2025). Different models in the series have different strengths and capabilities: (1) Gemini 2.5 Pro is our most intelligent thinking model, exhibiting strong reasoning and code capabilities. It excels at producing interactive web applications, is capable of codebase-level understanding and also exhibits emergent multimodal coding abilities. (2) Gemini 2.5 Flash is our hybrid reasoning model with a controllable thinking budget, and is useful for most complex tasks while also controlling the tradeoff between quality, cost, and latency. (3) Gemini 2.0 Flash is our fast and cost-efficient non-thinking model for everyday tasks and (4) Gemini 2.0 Flash-Lite is our fastest and most cost-efficient model, built for at-scale usage. A full comparison of the models in the Gemini 2.X model family is provided in Table 1. Taken together, the Gemini 2.X family of models cover the whole Pareto frontier of model capability vs cost, shifting it forward across a large variety of core capabilities, applications and use-cases, see Figure 1.

The Gemini 2.5 family of models maintain robust safety metrics while improving dramatically on

---

[1]Pokémon is a trademark of Nintendo Co., Ltd., Creatures Inc., and Game Freak Inc.

	Gemini 1.5 Flash	*Gemini 1.5 Pro*	**Gemini 2.0 Flash-Lite**	**Gemini 2.0 Flash**	**Gemini 2.5 Flash**	**Gemini 2.5 Pro**
Input modalities	Text, Image, Video, Audio	Text, Image, Video, Audio	Text, Image, Video, Audio	Text, Image, Video, Audio	Text, Image, Video, Audio	Text, Image, Video, Audio
Input length	1M	2M	1M	1M	1M	1M
Output modalities	Text	Text	Text	Text, Image*	Text, Audio*	Text, Audio*
Output length	8K	8K	8K	8K	64K	64K
Thinking	No	No	No	Yes*	Dynamic	Dynamic
Supports tool use?	No	No	No	Yes	Yes	Yes
Knowledge cutoff	November 2023	November 2023	June 2024	June 2024	January 2025	January 2025

Table 1 | Comparison of Gemini 2.X model family with Gemini 1.5 Pro and Flash. Tool use refers to the ability of the model to recognize and execute function calls (e.g., to perform web search, complete a math problem, execute code). *currently limited to Experimental or Preview, see Section 2.7. Information accurate as of publication date.*

helpfulness and general tone compared to their 2.0 and 1.5 counterparts. In practice, this means that the 2.5 models are substantially better at providing safe responses without interfering with important use cases or lecturing end users. We also evaluated Gemini 2.5 Pro's Critical Capabilities, including CBRN, cybersecurity, machine learning R&D, and deceptive alignment. While Gemini 2.5 Pro showed a significant increase in some capabilities compared to previous Gemini models, it did not reach any of the Critical Capability Levels in any area.

Our report is structured as follows: we begin by briefly describing advances we have made in model architecture, training and serving since the release of the Gemini 1.5 model. We then showcase the performance of the Gemini 2.5 models, including qualitative demonstrations of its abilities. We conclude by discussing the safety evaluations and implications of this model series.

2. Model Architecture, Training and Dataset

2.1. Model Architecture

The Gemini 2.5 models are sparse mixture-of-experts (MoE) (Clark et al., 2022; Du et al., 2021; Fedus et al., 2021; Jiang et al., 2024; Lepikhin et al., 2020; Riquelme et al., 2021; Roller et al., 2021; Shazeer et al., 2017) transformers (Vaswani et al., 2017) with native multimodal support for text, vision, and audio inputs. Sparse MoE models activate a subset of model parameters per input token by learning to dynamically route tokens to a subset of parameters (experts); this allows them to decouple total model capacity from computation and serving cost per token. Developments to the model architecture contribute to the significantly improved performance of Gemini 2.5 compared to Gemini 1.5 Pro (see Section 3). Despite their overwhelming success, large transformers and sparse MoE models are known to suffer from training instabilities (Chowdhery et al., 2022; Dehghani et al., 2023; Fedus et al., 2021; Lepikhin et al., 2020; Liu et al., 2020; Molybog et al., 2023; Wortsman et al., 2023; Zhai et al., 2023; Zhang et al., 2022). The Gemini 2.5 model series makes considerable progress in enhancing large-scale training stability, signal propagation and optimization dynamics, resulting in a considerable boost in performance straight out of pre-training compared to previous Gemini models.
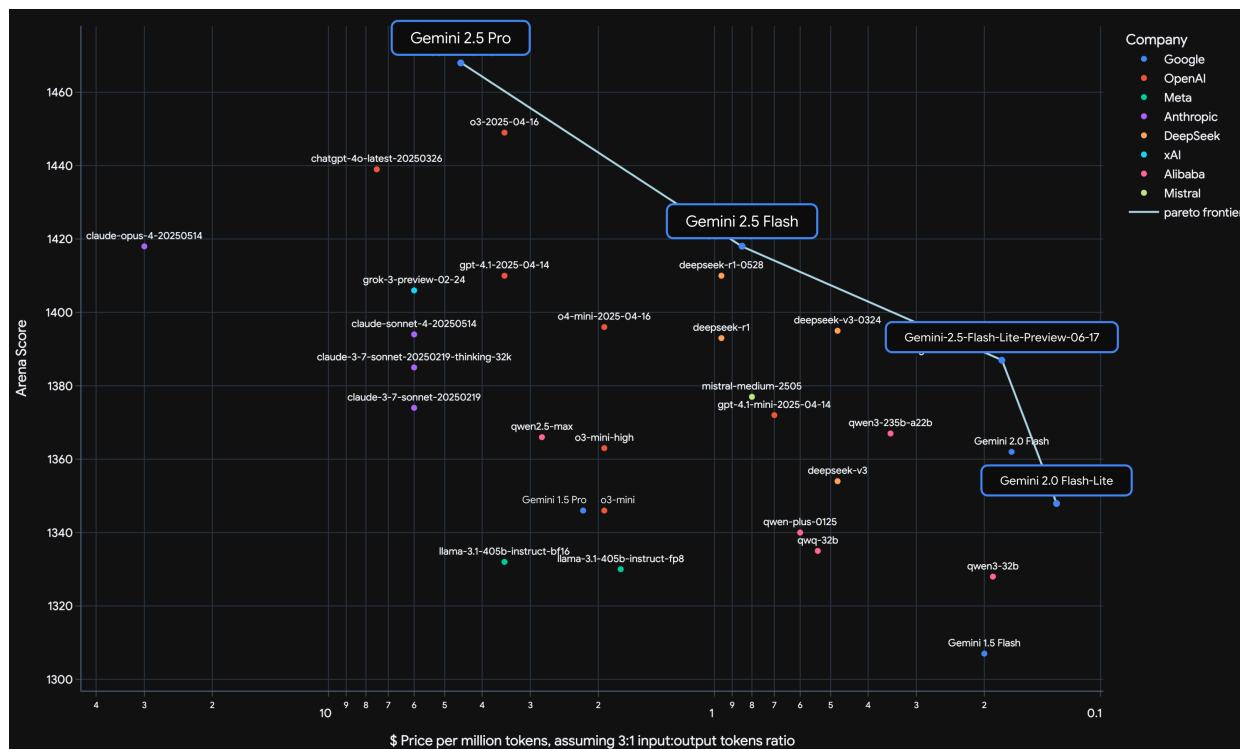
Figure 1 | Cost-performance plot. Gemini 2.5 Pro is a marked improvement over Gemini 1.5 Pro, and has an LMArena score that is over 120 points higher than Gemini 1.5 Pro. Cost is a weighted average of input and output tokens pricing per million tokens. Source: LMArena, imported on 2025-06-16.

Gemini 2.5 models build on the success of Gemini 1.5 in processing long-context queries, and incorporate new modeling advances allowing Gemini 2.5 Pro to surpass the performance of Gemini 1.5 Pro in processing long context input sequences of up to 1M tokens (see Table 3). Both Gemini 2.5 Pro and Gemini 2.5 Flash can process pieces of long-form text (such as the entirety of "Moby Dick" or "Don Quixote"), whole codebases, and long form audio and video data (see Appendix 8.5). Together with advancements in long-context abilities, architectural changes to Gemini 2.5 vision processing lead to a considerable improvement in image and video understanding capabilities, including being able to process 3-hour-long videos and the ability to convert demonstrative videos into interactive coding applications (see our recent blog post by Baddepudi et al., 2025).

The smaller models in the Gemini 2.5 series — Flash size and below — use distillation (Anil et al., 2018; Hinton et al., 2015), as was done in the Gemini 1.5 series (Gemini Team, 2024). To reduce the cost associated with storing the teacher's next token prediction distribution, we approximate it using a k-sparse distribution over the vocabulary. While this still increases training data throughput and storage demands by a factor of k, we find this to be a worthwhile trade-off given the significant quality improvement distillation has on our smaller models, leading to high-quality models with a reduced serving cost (see Figure 2).

2.2. Dataset

Our pre-training dataset is a large-scale, diverse collection of data encompassing a wide range of domains and modalities, which includes publicly available web documents, code (various programming languages), images, audio (including speech and other audio types) and video, with a cutoff date of June 2024 for 2.0 and January 2025 for 2.5. Compared to the Gemini 1.5 pre-training dataset
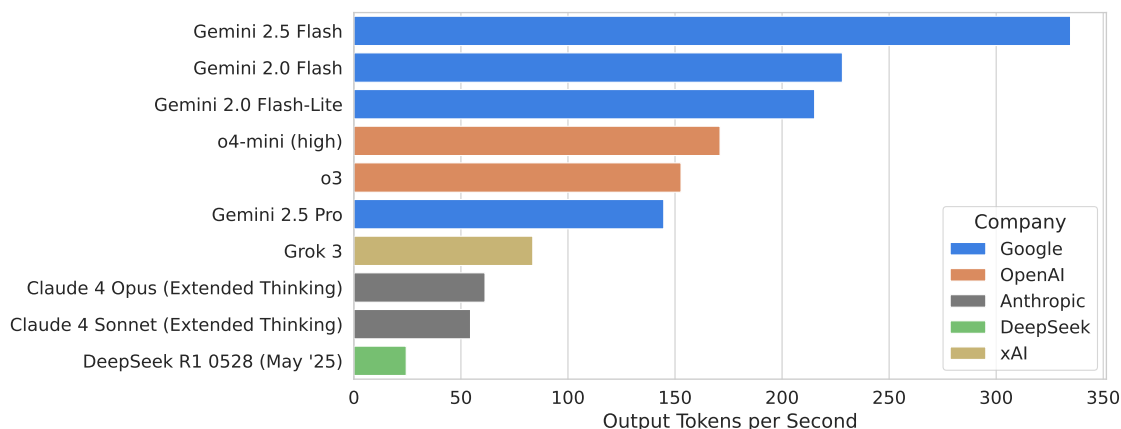
Figure 2 | Number of output tokens generated per second (after the first chunk has been received from the API) for different models. Source: `ArtificialAnalysis.ai`, imported on 2025-06-15.

we also utilized new methods for improved data quality for both filtering, and deduplication. Our post-training dataset, like Gemini 1.5, consists of instruction tuning data that is carefully collected and vetted. It is a collection of multimodal data with paired instructions and responses, in addition to human preference and tool-use data.

2.3. Training Infrastructure

This model family is the first to be trained on TPUv5p architecture. We employed synchronous data-parallel training to parallelise over multiple 8960-chip pods of Google's TPUv5p accelerators, distributed across multiple datacenters.

The main advances in software pre-training infrastructure compared with Gemini 1.5 were related to elasticity and mitigation of SDC (Silent Data Corruption) errors:

1. **Slice-Granularity Elasticity**: Our system now automatically continues training with fewer "slices" of TPU chips when there is a localized failure, and this reconfiguration results in tens of seconds of lost training time per interruption, compared with the 10 or more minute delay waiting for healthy machines to be rescheduled without elasticity; the system continues training at around 97% throughput while the failed slice is recovering. At the scale of this training run we see interruptions from hardware failures multiple times per hour, but our fault tolerance machinery is designed to tolerate the higher failure rates expected at much larger scales.

2. **Split-Phase SDC Detection**: On previous large-scale runs it could take many hours to detect and localize machines with SDC errors, requiring both downtime while debugging, and roll-back/replay of a large number of potentially corrupt training steps. We now use lightweight deterministic replay to immediately repeat any step with suspicious metrics, and compare per-device intermediate checksums to localize the root cause of any data corruption. Empirically, accelerators that start to exhibit intermittent SDCs are identified within a few minutes, and quickly excluded from the job. During this run, around 0.25% of steps were replayed due to suspected SDCs and 6% of these replays turned out to be genuine hardware corruption.

Both of the above techniques were relatively simple to implement due to the single-controller design of the Pathways system (Barham et al., 2022), which allows all accelerators to be coordinated from a single python program with a global view of the system state. The controller can make use of
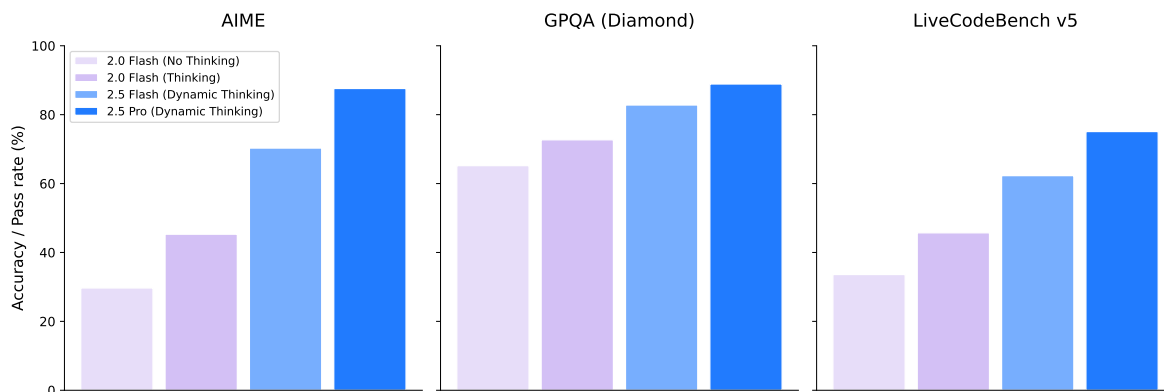
Figure 3 | Impact of "Thinking" on Gemini's performance on AIME 2025 (Balunović et al., 2025), LiveCodeBench (corresponding to 10/05/2024 - 01/04/2025 in the UI) (Jain et al., 2024) and GPQA diamond (Rein et al., 2024) benchmarks.

parallel 'remote python' operations on TPU workers to monitor training metrics, track performance stragglers, and root-cause SDC errors.

Overall during the run, 93.4% of the time was spent performing TPU computations; the remainder was approximately spent half in elastic reconfigurations, and half in rare tail cases where elasticity failed. Around 4.5% of the computed steps were replays or rollbacks for model debugging interventions.

2.4. Post-training

Since the initial announcement of Gemini 1.5, significant advancements have been made in our post-training methodologies, driven by a consistent focus on data quality across the Supervised Fine-Tuning (SFT), Reward Modeling (RM), and Reinforcement Learning (RL) stages. A key focus has been leveraging the model itself to assist in these processes, enabling more efficient and nuanced quality control.

Furthermore, we have increased the training compute allocated to RL, allowing deeper exploration and refinement of model behaviors. This has been coupled with a focus on verifiable rewards and model-based generative rewards to provide more sophisticated and scalable feedback signals. Algorithmic changes to the RL process have also improved stability during longer training. These advancements have enabled Gemini 2.5 to learn from more diverse and complex RL environments, including those requiring multi-step actions and tool use. The combination of these improvements in data quality, increased compute, algorithmic enhancements, and expanded capabilities has contributed to across-the-board performance gains (as described in Section 3) , notably reflected in the significant increase in the model's LMArena Elo scores, with both Gemini 2.5 Flash and Pro gaining more than 110 points over their Gemini 1.5 counterparts (122 for Gemini 2.5 Pro and 111 for Gemini 2.5 Flash, see Figure 1), along with significant improvements on several other frontier benchmarks.

2.5. Thinking

Past Gemini models produce an answer immediately following a user query. This constrains the amount of inference-time compute (Thinking) that our models can spend reasoning over a problem. Gemini Thinking models are trained with Reinforcement Learning to use additional compute at inference time to arrive at more accurate answers. The resulting models are able to spend tens of
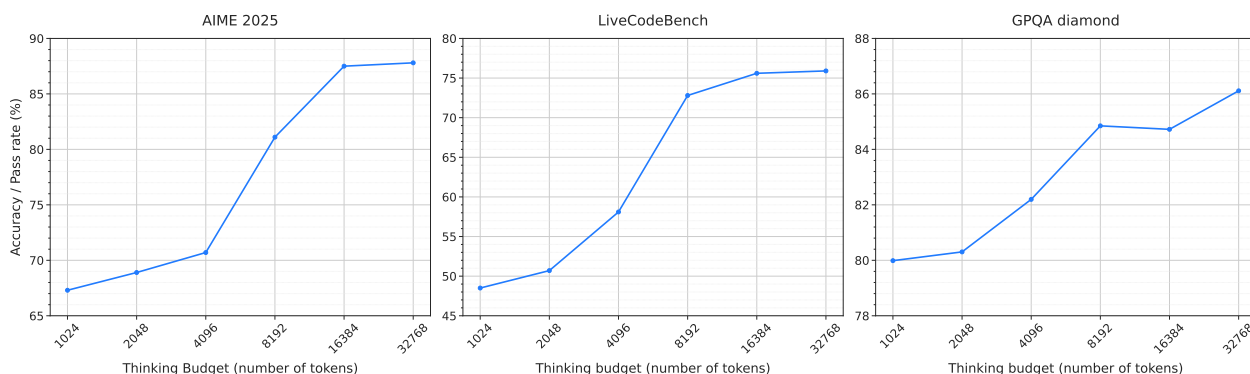
Figure 4 | Impact of thinking budget on performance on AIME 2025 (Balunović et al., 2025), Live-CodeBench (corresponding to 10/05/2024 - 01/04/2025 in the UI) (Jain et al., 2024) and GPQA diamond (Rein et al., 2024) benchmarks.

thousands of forward passes during a "thinking" stage, before responding to a question or query.

Our training recipe has evolved from the original experimental thinking model, Gemini 2.0 Flash Thinking (launched in December 2024), to the Gemini 2.5 Thinking series, which incorporates Thinking natively across all domains. The result is a single model that can achieve stronger reasoning performance across the board, and is able to scale up its performance further as a function of inference time (see Figure 3 for an example of the impact of Thinking).

We integrated Thinking with other Gemini capabilities, including native multimodal inputs (images, text, video, audio) and long context (1M+ tokens). For any of these capabilities, the model decides for itself how long to think before providing an answer. We also provide the ability to set a Thinking budget, constraining the model to respond within a desired number of tokens. This allows users to trade off performance with cost. To demonstrate this capability, we conducted experiments where we systematically varied the thinking budget, measured in the number of tokens the model is allowed to use for internal computation. As shown in Figure 4, increasing this budget allows the model to scale its performance and achieve significantly higher accuracy.

2.6. Capability-specific improvements

While most of the changes made to our training architecture and recipe since Gemini 1.5 have resulted in improvements across all capabilities, we have also made changes that have resulted in some capability-specific wins. We will now discuss these for code, factuality, long context, multilinguality, audio, video, and agentic use cases (with a particular focus on Gemini Deep Research).

Code

Gemini 2.0 and 2.5 represent a strategic shift of our development priorities towards delivering tangible real-world value, empowering users to address practical challenges and achieve development objectives within today's complex, multimodal software environments. To realize this, concerted efforts have been undertaken across both pre-training and post-training phases since Gemini 1.5. In pre-training, we intensified our focus on incorporating a greater volume and diversity of code data from both repository and web sources into the training mixture. This has rapidly expanded coverage and enabled the development of more compute-efficient models. Furthermore, we have substantially enhanced our suite of evaluation metrics for assessing code capabilities aligned with downstream use cases, alongside improving our ability to accurately predict model performance.

During post-training, we developed novel training techniques incorporating reasoning capabilities and curated a diverse set of engineering tasks, with the aim to equip Gemini with effective problem-solving skills crucial for addressing modern engineering challenges. Key applications demonstrating these advancements include IDE functionalities, code agent use cases for complex, multi-step operations within full repositories, and multimodal, interactive scenarios such as end-to-end web and mobile application development. Collectively, these efforts have yielded broad and significant improvements in Gemini's coding capabilities. This progress is evidenced by superior performance on established benchmarks: performance on LiveCodeBench (Jain et al., 2024) increased from 30.5% for Gemini 1.5 Pro to 74.2% for Gemini 2.5 Pro, while that for Aider Polyglot (Gauthier, 2025) went from 16.9% to 82.2%. Performance on SWEBench-verified (Chowdhury et al., 2024; Jimenez et al., 2024) went from 34.2% to 67.2%, see Table 3 and Figure 5 in Section 3.2. Furthermore, Gemini 2.5 Pro obtained an increase of over 500 Elo over Gemini 1.5 Pro on the LMArena WebDev Arena (Chiang et al., 2024; LMArena Team, 2025), resulting in meaningful enhancements in practical applications, including UI and web application development (Doshi, 2025a), and the creation of sophisticated agentic workflows (Kilpatrick, 2025).

Factuality

Within the context of generative models, ensuring the factuality of model responses to information-seeking prompts remains a core pillar of Gemini model development. With Gemini 1.5, our research was concentrated on enhancing the model's world knowledge and its ability to provide answers faithfully grounded in the context provided within the prompt. This effort culminated in the December 2024 release of FACTS Grounding (Jacovi et al., 2025), now an industry-standard benchmark for evaluating an LLM's capacity to generate responses grounded in user-provided documents. With Gemini 2.0 and 2.5, we have significantly expanded our scope to address multimodal inputs, long-context reasoning, and model-retrieved information. At the same time, the landscape and user expectations for factuality have evolved dramatically, shaped in part by Google's deployment of AI Overviews and AI Mode (Stein, 2025). To meet these demands, Gemini 2.0 marked a significant leap as our first model family trained to natively call tools like Google Search, enabling it to formulate precise queries and synthesize fresh information with sources. Building on this, Gemini 2.5 integrates advanced reasoning, allowing it to interleave these search capabilities with internal thought processes to answer complex, multi-hop queries and execute long-horizon tasks. The model has learned to use search and other tools, reason about the outputs, and issue additional, detailed follow-up queries to expand the information available to it and to verify the factual accuracy of the response. Our latest models now power the experiences of over 1.5B monthly active users in Google's AI Overviews and 400M users in the Gemini App. These models exhibit state-of-the-art performance across a suite of factuality benchmarks, including SimpleQA for parametric knowledge (Wei et al., 2024), FACTS Grounding for faithfulness to provided documents (Jacovi et al., 2024, 2025), and the Vectara Hallucination Leaderboard (Hughes et al., 2023), cementing Gemini as the model of choice for information-seeking demands.

Long context

Modeling and data advances helped us improve the quality of our models' responses to queries utilizing our one million-length context window, and we reworked our internal evaluations to be more challenging to help steer our modeling research. When hill-climbing, we targeted challenging retrieval tasks (like LOFT of Lee et al., 2024), long-context reasoning tasks (like MRCR-V2 of Vodrahalli et al., 2024), and multimodal tasks (like VideoMME of Fu et al., 2025). According to the results in Table 6, the new 2.5 models improve greatly over previous Gemini 1.5 models and achieve state-of-the-art quality on all of those. An example showcasing these improved capabilities for video recall can be

seen in Appendix 8.5, where Gemini 2.5 Pro is able to consistently recall a 1 second visual event out of a full 46-minute video.[2]

Multilinguality

Gemini's multilingual capabilities have also undergone a profound evolution since 1.5, which already encompassed over 400 languages via pretraining. This transformation stems from a holistic strategy, meticulously refining pre- and post-training data quality, advancing tokenization techniques, innovating core modeling, and executing targeted capability hillclimbing. The impact is particularly striking in Indic and Chinese, Japanese and Korean languages, where dedicated optimizations in data quality and evaluation have unlocked dramatic gains in both quality and decoding speed. Consequently, users benefit from significantly enhanced language adherence, responses designed to faithfully respect the requested output language, and a robust improvement in generative quality and factuality across languages, solidifying Gemini's reliability across diverse linguistic contexts.

Audio

While Gemini 1.5 was focused on native audio understanding tasks such as transcription, translation, summarization and question-answering, in addition to understanding, Gemini 2.5 was trained to perform audio generation tasks such as text-to-speech or native audio-visual to audio out dialog. To enable low-latency streaming dialog, we incorporated causal audio representations that also allow streaming audio into and out of Gemini 2.5. These capabilities derive from an increased amount of pre-training data spanning over 200 languages, and development of improved post-training recipes. Finally, through our improved post-training recipes, we have integrated advanced capabilities such as thinking, affective dialog, contextual awareness and tool use into Gemini's native audio models.

Video

We have significantly expanded both our pretraining and post-training video understanding data, improving the audio-visual and temporal understanding capabilities of the model. We have also trained our models so that they perform competitively with 66 instead of 258 visual tokens per frame, enabling using about 3 hours of video instead of 1h within a 1M tokens context window[3]. Two new applications that were not previously possible, but that have been unlocked as a result of these changes are: creating an interactive app from a video (such as a quiz to test students' understanding of the video content) and creating a p5.js animation to show the key concepts from the video. Our recent blog post (Baddepudi et al., 2025) shows examples of these applications.

Gemini as an Agent: Deep Research

Gemini Deep Research (Gemini Team, Google, 2024) is an agent built on top of the Gemini 2.5 Pro model designed to strategically browse the web and provide informed answers to even the most niche user queries. The agent is optimized to perform task prioritization, and is also able to identify when it reaches a dead-end when browsing. We have massively improved the capabilities of Gemini Deep Research since its initial launch in December 2024. As evidence of that, performance of Gemini Deep Research on the Humanity's Last Exam benchmark (Phan et al., 2025) has gone from 7.95% in December 2024 to the **SoTA score of 26.9% and 32.4% with higher compute** (June 2025).

---

[2]For further discussion on long context capabilities, challenges, and future outlook, the Release Notes podcast episode "Deep Dive into Long Context" provides additional insights and discussion: `https://youtu.be/NHMJ9mqKeMQ`.

[3]This is referred to as low media resolution in the API: `https://ai.google.dev/api/generate-content#Media Resolution`.

2.7. The path to Gemini 2.5

On the way to Gemini 2.5 Pro, we experimented with our training recipe, and tested a small number of these experimental models with users. We have already discussed Gemini 2.0 Flash Thinking (see Section 2.5). We will now discuss some of the other models briefly.

Gemini 2.0 Pro

In February 2025, we released an experimental version of Gemini 2.0 Pro. At the time, it had the strongest coding performance of any model in the Gemini model family, as well as the best understanding and world knowledge. It also came with our largest context window at 2 million tokens, which enabled it to comprehensively analyze and understand vast amounts of information. For further information about Gemini 2.0 Pro, please see our earlier blog posts (Kavukcuoglu, 2025; Mallick and Kilpatrick, 2025).

Gemini 2.0 Flash Native Image Generation Model

In March 2025, we released an experimental version of Gemini 2.0 Flash Native Image Generation. It has brought to the users new capabilities as a result of a strong integration between the Gemini model and image-generation capabilities, enabling new experiences related to image generation & image editing via natural-language prompting. Capabilities such as multi-step conversational editing or interleaved text-image generation are very natural in such a setting, and horizontal transfer related to multi-language coverage immediately allowed such experiences to happen across all the languages supported by the Gemini models. Native image generation turns Gemini into a multimodal creation partner and enables Gemini to express ideas through both text and images, and to seamlessly move between the two. For further information about Gemini 2.0 Flash Native Image Generation, please see our earlier blog posts (Kampf and Brichtova, 2025; Sharon, 2025)

Gemini 2.5 Audio Generation

With Gemini 2.5, the Controllable TTS and Native Audio Dialog capabilities are available as separate options on AI Studio (Generate Media and Stream sections respectively). Our Gemini 2.5 Preview TTS Pro and Flash models support more than 80 languages with the speech style controlled by a free formatted prompt which can specify style, emotion, pace, etc, while also being capable of following finer-grained steering instructions specified in the transcript. Notably, Gemini 2.5 Preview TTS can generate speech with multiple speakers, which enables the creation of podcasts as used in NotebookLM Audio Overviews (Wang, 2024). Our Gemini 2.5 Flash Preview Native Audio Dialog model uses native audio generation, which enables the same level of style, pacing and accent control as available in our controllable TTS offering. Our dialog model supports tool use and function calling, and is available in more than 24 languages. With native audio understanding and generation capabilities, it can understand and respond appropriately to the user's tone. This model is also capable of understanding when to respond to the user, and when not to respond, ignoring background and non-device directed audio. Finally, we also offer an advanced 'Thinking' variant that effectively handles more complex queries and provides more robust and reasoned responses in exchange for some additional latency.

Gemini 2.5 Flash-Lite

In June 2025, we released an experimental version of Gemini 2.5 Flash-Lite (`gemini-2.5-flash-lite-preview-06-17`). It comes with the same capabilities that make Gemini 2.5 helpful, including the ability to turn thinking on at different budgets, connecting to tools like Google Search and code

execution, support for multimodal inputs and a 1 million-token context length. Our goal was to provide an economical model class which provides ultra-low-latency capabilities and high throughput per dollar, echoing the initial release of 2.0 Flash-Lite (Google DeepMind, 2025b; Mallick and Kilpatrick, 2025).

Gemini 2.5 Pro Deep Think

To advance Gemini's capabilities towards solving hard reasoning problems, we developed a novel reasoning approach, called Deep Think, that naturally blends in parallel thinking techniques during response generation. Deep Think enables Gemini to creatively produce multiple hypotheses and carefully critique them before arriving at the final answer, achieving state-of-the-art performances in challenging benchmarks such as Olympiad math (USAMO 2025), competitive coding (LiveCodeBench), and multimodality (MMMU), see more details at (Doshi, 2025b). We announced Gemini 2.5 Deep Think at Google I/O and launched an experimental version to trusted testers and advanced users in June 2025.