

# 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<sup>1</sup> ([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

---

<sup>1</sup>Pokémon is a trademark of Nintendo Co., Ltd., Creatures Inc., and Game Freak Inc.

	<i>Gemini 1.5 Flash</i>	<i>Gemini 1.5 Pro</i>	<i>Gemini 2.0 Flash-Lite</i>	<i>Gemini 2.0 Flash</i>	<i>Gemini 2.5 Flash</i>	<i>Gemini 2.5 Pro</i>
<b>Input modalities</b>	Text, Image, Video, Audio	Text, Image, Video, Audio	Text, Image, Video, Audio	Text, Image, Video, Audio	Text, Image, Video, Audio	Text, Image, Video, Audio
<b>Input length</b>	1M	2M	1M	1M	1M	1M
<b>Output modalities</b>	Text	Text	Text	Text, Image*	Text, Audio*	Text, Audio*
<b>Output length</b>	8K	8K	8K	8K	64K	64K
<b>Thinking</b>	No	No	No	Yes*	Dynamic	Dynamic
<b>Supports tool use?</b>	No	No	No	Yes	Yes	Yes
<b>Knowledge cutoff</b>	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.

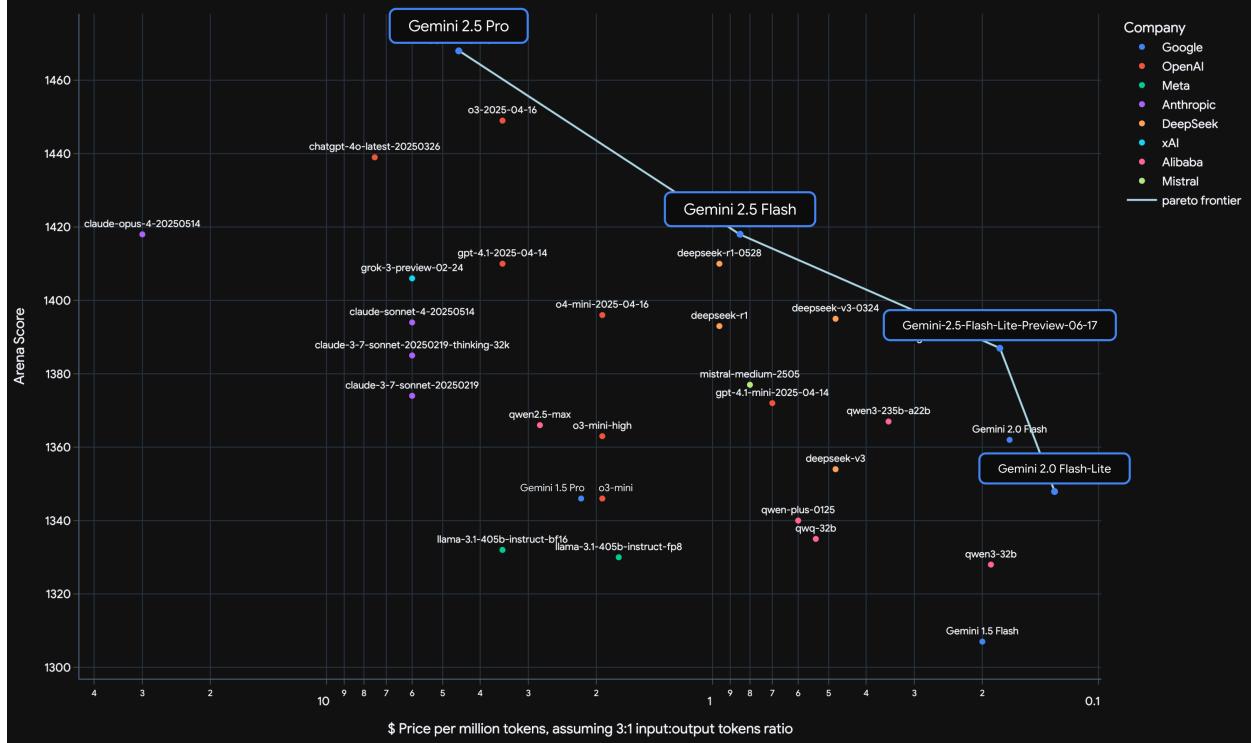


Figure 1 | Cost-performance plot. Gemini 2.5 Pro is a marked improvement over Gemini 1.5 Pro, and has an LMArena score that is over 120 points higher than Gemini 1.5 Pro. Cost is a weighted average of input and output tokens pricing per million tokens. Source: [LMArena](#), imported on 2025-06-16.

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

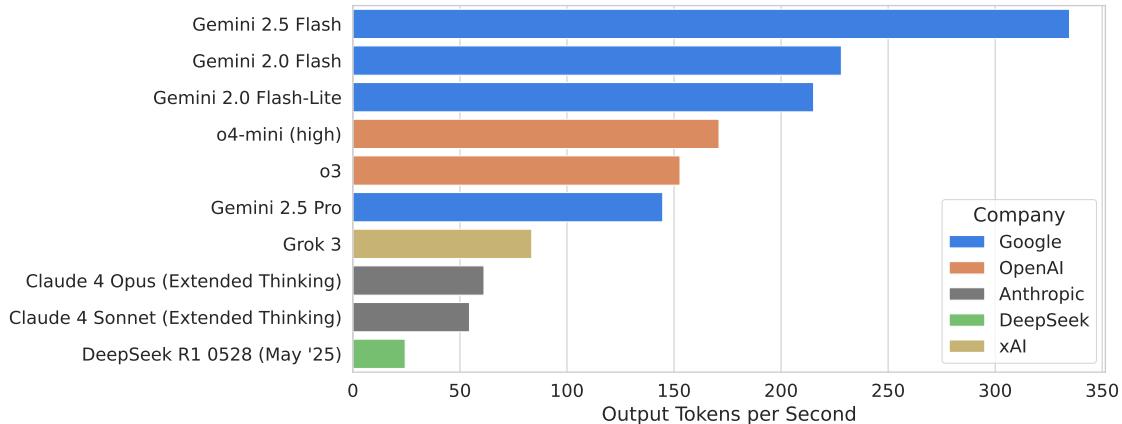


Figure 2 | Number of output tokens generated per second (after the first chunk has been received from the API) for different models. Source: [ArtificialAnalysis.ai](#), imported on 2025-06-15.

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

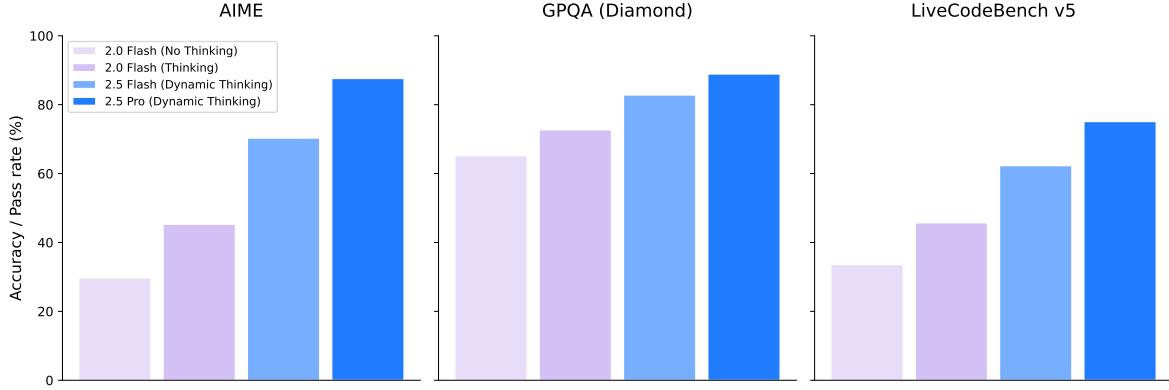


Figure 3 | Impact of “Thinking” on Gemini’s performance on AIME 2025 (Balunović et al., 2025), LiveCodeBench (corresponding to 10/05/2024 - 01/04/2025 in the UI) (Jain et al., 2024) and GPQA diamond (Rein et al., 2024) benchmarks.

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

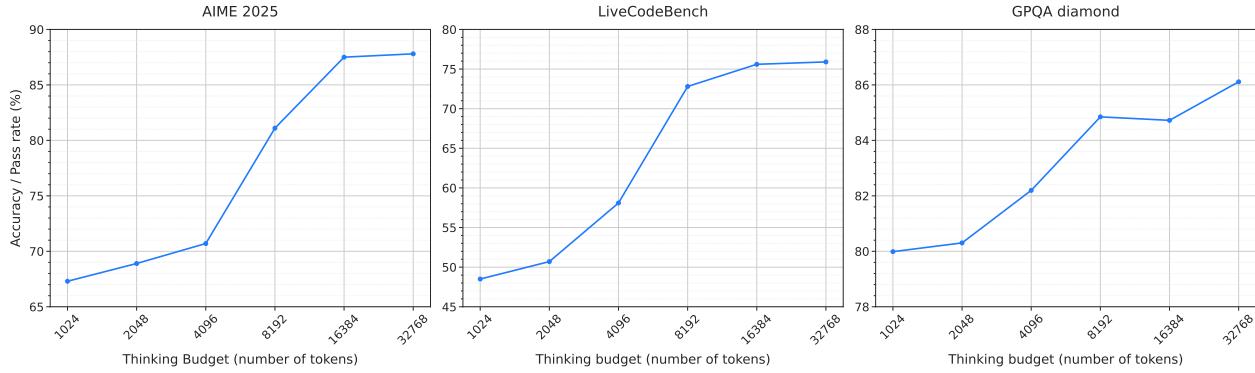


Figure 4 | Impact of thinking budget on performance on AIME 2025 (Balunović et al., 2025), LiveCodeBench (corresponding to 10/05/2024 - 01/04/2025 in the UI) (Jain et al., 2024) and GPQA diamond (Rein et al., 2024) benchmarks.

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.<sup>2</sup>

## 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<sup>3</sup>. 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).

---

<sup>2</sup>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>.

<sup>3</sup>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.

### 3. Quantitative evaluation

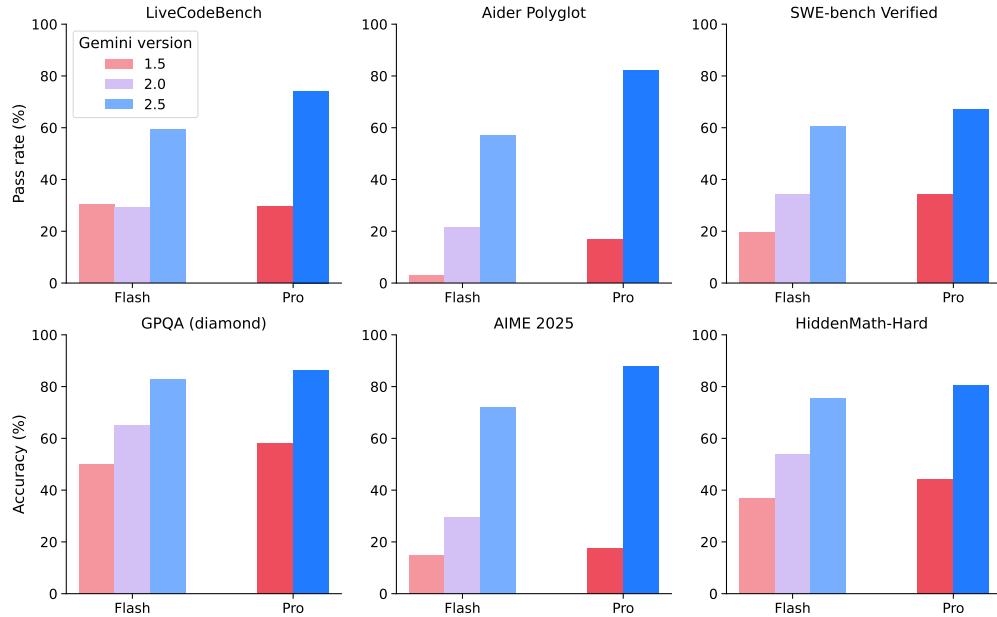


Figure 5 | Performance of Gemini 2.X models at coding, math and reasoning tasks in comparison to previous Gemini models. SWE-bench verified numbers correspond to the “multiple attempts” setting reported in Table 3.

We will now examine the performance of the Gemini 2.X model family across a wide range of benchmarks. We will first compare the performance of the Gemini 2.X models to the earlier Gemini 1.5 Pro and Flash models, before we compare the performance of Gemini 2.5 Pro to other available large language models.

With web-scale pre-training of AI models, coupled with the post-training techniques that allow policy and reward models to leverage public benchmarks, avoiding leaks and biases in the data used for pre- and post-training is a persistent challenge. In the development of the Gemini 2.5 series, in addition to the standard n-gram based decontamination we used in Gemini 1.5, we also employed semantic-similarity and model based decontamination procedures to help mitigate evaluation set leakage. To move beyond the reliance on training set decontamination, we also continue reporting on internally developed non-public benchmarks, such as HiddenMath.

Model	AI Studio model ID
Gemini 1.5 Flash	gemini-1.5-flash-002
Gemini 1.5 Pro	gemini-1.5-pro-002
Gemini 2.0 Flash-Lite	gemini-2.0-flash-lite-001
Gemini 2.0 Flash	gemini-2.0-flash-001
Gemini 2.5 Flash	gemini-2.5-flash
Gemini 2.5 Pro	gemini-2.5-pro

Table 2 | Mapping of Gemini model names to AI Studio API model IDs.

### 3.1. Methodology

In Table 3, we compare the performance of Gemini 2.5 models to the Gemini 1.5 models, while in Table 4, we compare the performance of Gemini 2.5 Pro to that of other large language models.

**Gemini results:** All Gemini scores are pass@1, and are “single attempt” settings unless otherwise specified. In the “single attempt” setting, no majority voting or parallel test-time compute is permitted, while in the “multiple attempts” setting, test-time selection of the candidate answer is allowed. All Gemini evaluations are run with the AI Studio API for the model id that we provide in Table 2, with default sampling settings. To reduce variance, we average over multiple trials for smaller benchmarks. Aider Polyglot scores are the pass rate average of 3 trials. Vibe-Eval results are reported using Gemini as a judge.

**Non-Gemini results:** All the results for non-Gemini models are sourced from providers’ self reported numbers unless mentioned otherwise. All “SWE-bench Verified” numbers follow official provider reports, which means that they are computed using different scaffoldings and infrastructure, and aren’t directly comparable.

For some evaluations, we obtain results from the external leaderboards that report results on these benchmarks. Results for Humanity’s Last Exam results are sourced from [Scale’s leaderboard](#) and results for DeepSeek are obtained from the [text-only variant of the leaderboard](#) (indicated with a ◊ in Table 4). For Gemini 2.0 models, the reported results are [on an earlier HLE dataset](#) (indicated with a † in Table 3). Results on LiveCodeBench results are taken from [\(1/1/2025 - 5/1/2025\) in the UI](#). Aider Polyglot numbers come from [the Aider leaderboard](#) and results for SimpleQA come from [this repo](#) where available. Results on FACTS Grounding come from [Kaggle](#). In the case of LOFT and MRCR-V2, we report results on both the 128k context length variant, as well as the 1M context length variant. In the 128k context length variant, we measure performance on contexts up to 128k, while for the 1M context length variant, we report performance on context lengths of exactly 1M.

More details on all benchmarks, including subsets and how scores were obtained can be found in Table 11 in Appendix 8.1.

### 3.2. Core capability quantitative results

As can be seen in Table 3, and Figure 5, the Gemini 2.5 models excel at coding tasks such as LiveCodeBench, Aider Polyglot and SWE-bench Verified, and represent a marked improvement over previous models.

In addition to coding performance, Gemini 2.5 models are noticeably better at math and reasoning tasks than Gemini 1.5 models: performance on AIME 2025 is 88.0% for Gemini 2.5 Pro compared to 17.5% for Gemini 1.5 Pro, while performance on GPQA (diamond) went from 58.1% for Gemini 1.5 Pro to 86.4%. Performance on image understanding tasks has also increased significantly.

It is also interesting to note that the Gemini 2.5 Flash model has become the second most capable model in the Gemini family, and has overtaken not just previous Flash models, but also the Gemini 1.5 Pro model released one year ago.

Capability	Benchmark		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
Code	LiveCodeBench		30.3%	29.7%	29.1%	29.1%	59.3%	<b>74.2%</b>
	Aider Polyglot		2.8%	16.9%	10.5%	21.3%	56.7%	<b>82.2%</b>
	SWE-bench Verified	<i>single attempt</i> <i>multiple attempts</i>	9.6% 19.7%	22.3% 34.2%	12.5% 23.1%	21.4% 34.2%	48.9% 60.3%	<b>59.6%</b> <b>67.2%</b>
Reasoning	GPQA (diamond)		50.0%	58.1%	50.5%	65.2%	82.8%	<b>86.4%</b>
	Humanity's Last Exam	<i>no tools</i>	-	4.6%	4.6% †	5.1% †	11.0%	<b>21.6%</b>
Factuality	SimpleQA		8.6%	24.9%	16.5%	29.9%	26.9%	<b>54.0%</b>
	FACTS Grounding		82.9%	80.0%	82.4%	84.6%	85.3%	<b>87.8%</b>
Multilinguality	Global MMLU (Lite)		72.5%	80.8%	78.0%	83.4%	88.4%	<b>89.2%</b>
	ECLeKTic		16.4%	27.0%	27.7%	33.6%	36.8%	<b>46.8%</b>
Math	AIME 2025		14.7%	17.5%	23.8%	29.7%	72.0%	<b>88.0%</b>
	HiddenMath- Hard		36.8%	44.3%	47.4%	53.7%	75.5%	<b>80.5%</b>
Long-context	LOFT (hard retrieval)	$\leq 128K$ $1M$	67.3% 36.7%	75.9% 47.1%	50.7% 7.6%	58.0% 7.6%	82.1% 58.9%	<b>87.0%</b> <b>69.8%</b>
	MRCR-V2 (8-needle)	$\leq 128K$ $1M$	18.4% 10.2%	26.2% 12.1%	11.6% 4.0%	19.0% 5.3%	54.3% <b>21.0%</b>	<b>58.0%</b> 16.4%
Image Understanding	MMMU		58.3%	67.7%	65.1%	69.3%	79.7%	<b>82.0%</b>
	Vibe-Eval (Reka)		52.3%	55.9%	51.5%	55.4%	65.4%	<b>67.2%</b>
	ZeroBench		0.5%	1.0%	0.75%	1.25%	2.0%	<b>4.5%</b>
	BetterChartQA		59.0%	65.8%	52.3%	57.8%	67.3%	<b>72.4%</b>

Table 3 | Evaluation of Gemini 2.5 family across a wide range of core capability benchmarks and in comparison to Gemini 1.5 models. Please see Tables 5 and 6 for audio and video evaluations. See Table 11 Appendix 8.1 for benchmarks and evaluation details.

### 3.3. Evaluation of Gemini 2.5 Pro against other large language models

Relative to other large language models that are available (see Table 4), Gemini achieves the highest score on the Aider Polyglot coding task, Humanity’s Last Exam, GPQA (diamond), and on the SimpleQA and FACTS Grounding factuality benchmarks out of all of the models examined here. Gemini also continues to stand out for achieving the SoTA score on both the LOFT and MRCR long-context tasks at 128k context, and is the only one, amongst the models examined in the above table, to support context lengths of 1M+ tokens.

Not all of the models shown in Table 4 have native support for multimodal inputs. As such, we compare against a different set of models for audio and video understanding.

#### **Audio Understanding**

In Table 5, we showcase the performance of the Gemini 2.5 model family at audio understanding, and compare the performance of these models to earlier Gemini models, as well as to GPT models. Gemini 2.5 Pro demonstrates state-of-the-art audio understanding performance as measured by public benchmarks for ASR and AST, and compares favorably to alternatives under comparable testing conditions (using the same prompts and inputs).

#### **Video Understanding**

In Table 6, we show the performance of Gemini 2.5 models at video understanding. As can be seen, Gemini 2.5 Pro achieves state-of-the-art performance on key video understanding benchmarks, surpassing recent models like GPT 4.1 under comparable testing conditions (same prompt and video

Capability	Benchmark	Gemini 2.5 Pro	o3 high	o4-mini high	Claude 4 Sonnet	Claude 4 Opus	Grok 3 Beta <i>Extended Thinking</i>	DeepSeek R1 0528
Code	LiveCodeBench	74.2%	72.0%	<b>75.8%</b>	48.9%	51.1%	—	70.5%
	Aider Polyglot	<b>82.2%</b>	79.6%	72.0%	61.3%	72.0%	53.3%	71.6%
	SWE-bench Verified	single attempt multiple attempts	59.6% 67.2%	69.1% -	68.1% -	<b>72.7%</b> <b>80.2%</b>	72.5% 79.4%	- -
Reasoning	GPQA (diamond)	single attempt	<b>86.4%</b>	83.3%	81.4%	75.4%	79.6%	80.2%
	Humanity’s Last Exam	no tools	<b>21.6%</b>	20.3%	18.1%	7.8%	10.7%	- 14.0% ◊
Factuality	SimpleQA		<b>54.0%</b>	48.6%	19.3%	-	-	43.6% 27.8%
	FACTS Grounding		<b>87.8%</b>	69.9%	62.1%	79.1%	77.7%	74.8% 82.4%
Math	AIME 2025	single attempt	88.0%	88.9%	<b>92.7%</b>	70.5%	75.5%	77.3% 87.5%
Long-context	LOFT (hard retrieval)	≤128K 1M	<b>87.0%</b> <b>69.8%</b>	77.0% -	60.5% -	81.6% -	73.1% -	- -
	MRCR-V2 (8-needle)	≤128K 1M	<b>58.0%</b> <b>16.4%</b>	57.1% -	36.3% -	39.1% -	16.1%* -	34.0% -
Image Understanding	MMMU	single attempt	82.0%	<b>82.9%</b>	81.6%	74.4%	76.5%	76.0% <i>No MM support</i>

Table 4 | Performance comparison of Gemini 2.5 Pro with other large language models on different capabilities. Please see Tables 5 and 6 for audio and video evaluations. See Table 11 for benchmarks and evaluation details. \*: with no thinking and API refusals

Benchmark	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	GPT-4o mini Audio Preview	GPT 4o Audio Preview	GPT 4o transcribe
FLEURS (53 lang, WER ↓)	12.71	7.14	9.60	9.04	9.95	<b>6.66</b>	19.52	12.16	8.17
CoVoST2 (21 lang, BLEU ↑)	34.81	37.53	34.74	36.35	36.15	<b>38.48</b>	29.5	35.89	-

Table 5 | Performance comparison of Gemini 2.5 models to earlier Gemini models, as well as to GPT models for audio understanding. Note that for GPT models, metrics may differ from those previously reported due to differing eval methodologies. See Table 11 for benchmarks and evaluation details.

frames). For cost-sensitive applications, Gemini 2.5 Flash provides a highly competitive alternative.

Modalities	Benchmark	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	OpenAI GPT 4.1
visual-only	ActivityNet-QA	56.2	57.3	55.3	56.4	65.1	<b>66.7</b>	60.4
	EgoTempo	34.5	36.3	30.1	39.3	36.7	<b>44.3</b>	40.3
	Perception Test	66.5	69.4	67.5	68.8	75.1	<b>78.4</b>	64.8
	QVHighlights	64.4	68.7	25.7	63.9	52.4	<b>75.0</b>	71.4
	VideoMMMU	64.8	70.4	64.3	68.5	79.2	<b>83.6</b>	60.9
audio + visual	1H-VideoQA	61.9	72.2	55.6	67.5	67.5	<b>81.0</b>	56.8
	LVBench	61.9	65.7	52	61.8	62.7	<b>78.7</b>	63.4
	VideoMME	70.4	73.2	62.1	72.8	75.5	<b>84.3</b>	72.0
	VATEX	56.9	55.5	58.5	56.9	65.2	<b>71.3</b>	64.1
	VATEX-ZH	46.2	52.2	43.2	48.5	43.9	<b>59.7</b>	48.7
visual + subtitles	YouCook2 Cap	153.2	170.0	78.6	129.0	177.6	<b>188.3</b>	127.6
	Minerva	49.6	52.8	46.8	52.4	60.7	<b>67.6</b>	54.0
	Neptune	78.7	82.7	81.5	83.1	84.3	<b>87.3</b>	85.2
audio+visual+subtitles	VideoMME	77.3	79.8	72.5	78.8	81.5	<b>86.9</b>	79.6

Table 6 | Evaluation of Gemini 2.5 vs. prior models and GPT 4.1 on video understanding benchmarks. Performance is measured by string-match accuracy for multiple-choice VideoQA, LLM-based accuracy for open-ended VideoQA, R1@0.5 for moment retrieval and CIDEr for captioning. See Table 11 for benchmarks and evaluation details.

## 4. Example use cases of Gemini 2.5 Pro

### 4.1. Gemini Plays Pokémon

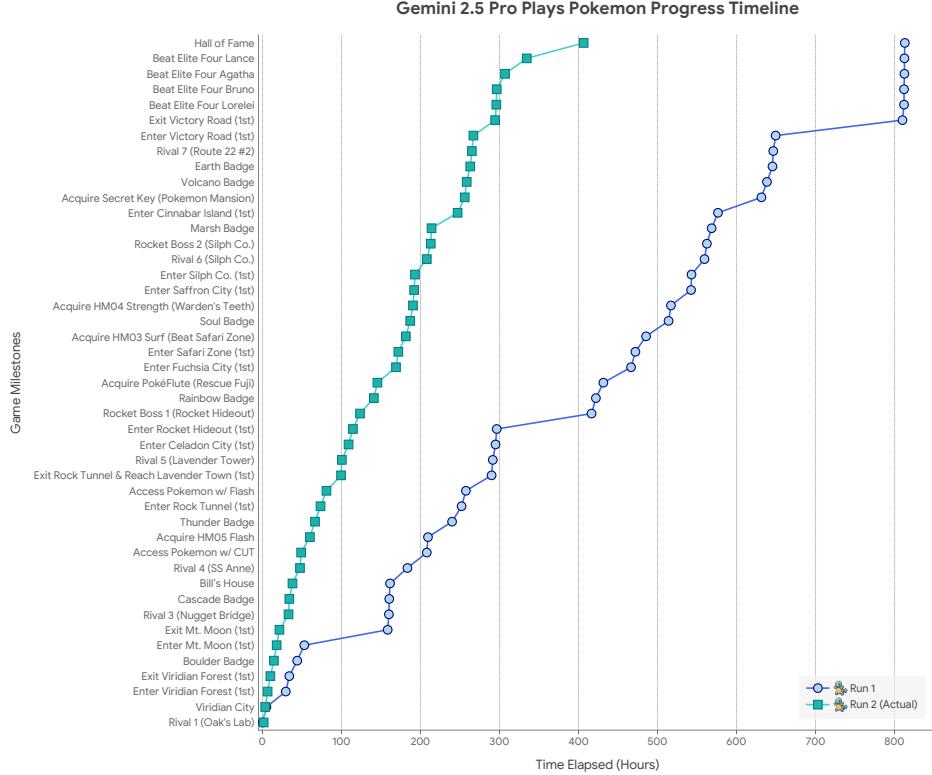


Figure 6 | Progression of the Gemini Plays Pokémon agent through the game, across two runs. Run 1 was the development run where changes to the harness were performed. Run 2 is the fully autonomous run with the final fixed scaffold. Both runs have the same starter (Squirtle). The events are ordered on the y-axis by the order they happened, following the order of Run 2 when there is a conflict. Notably, the GPP agent additionally went through the difficult (and optional) Seafoam Islands dungeon in Run 2, while in Run 1, GPP reached Cinnabar Island via Pallet Town and Route 21.

On March 28, 2025, an independent developer not affiliated with Google, [Joel Zhang](#), set up a Twitch stream (Gemini Plays Pokémon, or GPP) for Gemini 2.5 Pro (Gemini 2.5 Pro Exp 03-25) to play Pokémon Blue on stream ([Zhang, 2025](#)) as an experiment to better understand how well the model was capable of playing Pokémon (in a similar spirit to Claude Plays Pokémon, see [Anthropic 2025](#)). In this initial run through the game, the goal was to live-stream the development process of an agentic harness capable of playing the full game (and in particular the minimal transformation of vision to text necessary to do so), see Figure 14 for a description of the final agent setup. As such, over the course of the run, modifications were made to the setup as difficulties arose, providing a deeply interesting lens via which to analyze some of the qualitative improvements that the 2.5 Pro model has made, particularly in the regimes of solving long reasoning problems and agentic capabilities over extended time horizons. Around 1 month later, on May 2, 2025, Gemini 2.5 Pro completed the game after 813 hours and entered the Hall of Fame to become the Pokémon League Champion! On May 22, 2025, GPP began a fully autonomous 2nd run through the game with Gemini 2.5 Pro (Gemini 2.5 Pro Preview 05-06) with the finalized fixed agentic harness, and progressed through the game considerably faster, completing the game in 406.5 hours (nearly exactly half the time of the first run).

See Figure 6 for a timeline of GPP’s progress through major game milestones to game completion. We report # hours to each milestone in order to normalize for the amount of time models take per action. See Appendix 8.2 for more figures.

### **Capabilities assessment**

Gemini 2.5 Pro showcased many impressive capabilities associated with reasoning and long-term planning while playing Pokémon. We will now discuss two in particular, but for more examples, see Appendix 8.2.

**Long Context Agentic Tooling** Within the agent scaffolding, GPP has access to two agentic tools (see Figure 14). These prompted versions of Gemini 2.5 Pro, hereafter `pathfinder` and `boulder_puzzle_strategist`, have been able to:

1. Solve complex spinner puzzles in one shot (for instance in Rocket Hideout),
2. Solve the step-constrained multi-map puzzle of the Safari Zone,
3. Find long pathways through complex mazes like Route 13,
4. Solve boulder puzzles across long distances in Victory Road and the Seafoam Islands.

Each task requires reasoning over a long context - the `pathfinder` model would often have to reason over contexts of 100K+ tokens, and find paths up to 50 actions in length (in the extreme case, paths consisting of up to 150 actions have also been found!).

**Long Horizon Task Coherence** While Gemini 2.5 Pro is impressive in a more local sense, the agent also exhibited remarkable long-term task coherence in achieving global, high-level goals in the face of real and hallucinated setbacks towards making forward progress. Because the agent is able to change goals at will, and will generally follow those goals as long as needed, it is extremely impressive that the agent can satisfy numerous requirements for tactical, necessary goals, such as acquiring Hidden Moves, as well as maintain enough strategic task coherence to beat the entire game and become the Pokémon Champion.

### **Where does 2.5 Pro struggle while playing Pokémon?**

In addition to more standard hallucination issues (which interestingly were plausibly reduced in Run 2 by explicitly prompting the model to act as a player completely new to the game, see Appendix 8.2 for more details), there are a few particular points of struggle we would like to emphasize.

**Screen reading** While obtaining excellent benchmark numbers on real-world vision tasks, 2.5 Pro struggled to utilize the raw pixels of the Game Boy screen directly, though it could occasionally take cues from information on the pixels. As a result, it was necessary for the required information from the screen to be translated into a text format in the agent framework, using information from the game’s RAM state. During one portion of the game, the developer tested an ablation where all vision was completely removed from the model context – the model was able to function roughly as well as without the vision information, suggesting that most of the performance does not significantly depend on the visual input.

**Long Context Reasoning** Gemini 2.5 Pro’s state-of-the-art long context performance for both reasoning and retrieval tasks (see Tables 3 and 4) was a cornerstone of the GPP agent’s success. Its ability to reason over a 100k token context was instrumental for leveraging the complex toolset and

maintaining a relatively coherent strategy (e.g., optimal balance of performance, planning quality, and information recall.)

While Gemini 2.5 Pro supports 1M+ token context, making effective use of it for agents presents a new research frontier. In this agentic setup, it was observed that as the context grew significantly beyond 100k tokens, the agent showed a tendency toward favoring repeating actions from its vast history rather than synthesizing novel plans. This phenomenon, albeit anecdotal, highlights an important distinction between long-context for retrieval and long-context for multi-step, generative reasoning.

Teaching an agent to effectively plan and avoid such loops over massive past trajectories of context is an exciting and active area of research; the co-design of agent scaffolds and models to unlock the full potential of million-token context is an intriguing research direction and one of our primary focuses.

## 4.2. What else can Gemini 2.5 do?

Gemini 2.5 Pro excels at transforming diverse, often unstructured, inputs into interactive and functional applications. For instance, it can [take a PDF script of a play and generate a tool that allows drama students to practice their lines](#). Gemini 2.5 Pro can also take an uploaded photograph of a bookshelf and create a [curated book recommendation application](#). Gemini 2.5 Pro can utilize its underlying spatial understanding capability and convert images into a structural representation like HTML or SVG. In Figure 16 in Appendix 8.4, we show a comparison of Gemini 1.5 Pro and Gemini 2.5 Pro on an image-to-svg task, where Gemini 2.5 Pro reconstructs much more visual details and the spatial arrangements of objects better resembles the original image.

Furthermore, Gemini 2.5 Pro demonstrates strong skills in generating sophisticated simulations and visualizations, ranging from [interactive solar system models](#) ([source](#)) to the creative rendering of abstract mathematical concepts, such as [drawing a logo using Fourier series](#) ([source](#)). This capability extends to the development of tools that intersect creativity and utility: we see examples of specialized applications like a [custom cartography tool](#) or use cases that generate [photorealistic 3D user interfaces](#) from descriptive text and reference images, complete with appropriate styling and interactivity ([source](#)).

Collectively, these examples illustrate that Gemini 2.5 Pro is not just a useful coding and writing assistant, but excels at a wide range of complex tasks, ranging from those relevant for education to creative expression. The model empowers users to rapidly prototype specialized utilities, develop engaging educational content, and realize intricate creative visions with a high degree of sophistication.

## 4.3. Gemini in Google Products

As a final example of what Gemini can do, we note that Gemini (or a custom version of Gemini) is now incorporated into a wide variety of Google products. These include, but are not limited to, [AI Overviews](#) and [AI Mode](#) within Google Search, [Project Astra](#), the audiovisual-to-audio dialog agent, [Gemini Deep Research](#), the research assistant discussed in Section 2.7, [NotebookLM](#), the tool capable of generating podcasts and audio overviews from even the most obscure inputs, [Project Mariner](#), the web browsing agent, and Google's coding agent, [Jules](#).

## 5. Safety, Security, and Responsibility

We're committed to developing Gemini responsibly, innovating on safety and security alongside capabilities. We describe our current approach in this section, which includes how we train and evaluate our models, focusing on automated red teaming, going through held-out assurance evaluations on present-day risks, and evaluating the potential for dangerous capabilities in order to proactively anticipate new and long-term risks.

### Guideline for Navigating This Section

1. **Our Process (Section 5.1):** Begin here to understand our overall safety methodology.
2. **Policies and Desiderata (Section 5.2):** Next, dive into the safety criteria we use to evaluate and optimize our systems.
3. **Training for Safety (Section 5.3):** Discover how we incorporate safety into pre-training and post-training.
4. **Results from Development Evaluations (Section 5.4):** Results on our development evaluations for policies and desiderata.
5. **Automated Red Teaming (Section 5.5):** A description and results from our automated red teaming work for safety and security.
6. **Memorization & Privacy (Section 5.6):** Our analysis of memorization and privacy risks.
7. **Assurance Evaluations and Frontier Safety Framework (Section 5.7):** We dive into our held-out evaluations and tests for dangerous capabilities.
8. **External Safety Testing (Section 5.8):** Learn what independent testers discovered about our system's safety.

### 5.1. Our Process

We aim for Gemini to adhere to specific safety, security, and responsibility criteria. These cover what Gemini should not do (e.g., encourage violence), and what Gemini should do (e.g., respond in a helpful way when possible instead of refusing, provide multiple perspectives when consensus does not exist). We also leverage automated red teaming to identify cases where the model fails to respond in a safe or helpful manner. These failure cases are used to improve evaluations and training data.

Once the model is trained, we run assurance evaluations that we then use for review and release decisions. Importantly, these are conducted by a group outside of the model development team, and datasets are held out. Furthermore, for models where there are new capabilities or a significant performance improvement, we engage independent external groups, including domain experts and a government body, to further test the model to identify blind spots.

We also evaluate the model for dangerous capabilities outlined in our Frontier Safety Framework ([Google DeepMind, 2025a](#)), namely: Cybersecurity, CBRN, Machine Learning R&D, and Deceptive Alignment.

Finally, The Google DeepMind Responsibility and Safety Council (RSC), our governance body, reviews initial ethics and safety assessments on novel model capabilities in order to provide feedback and guidance during model development. The RSC also reviews metrics on the models' performance via assurance evals and informs release decisions.

## 5.2. Policies and Desiderata

### *Safety policies*

The Gemini safety policies align with Google’s standard framework which prevents our Generative AI models from generating specific types of harmful content, including:

1. Child sexual abuse and exploitation
2. Hate speech (e.g., dehumanizing members of protected groups)
3. Dangerous content (e.g., promoting suicide, or instructing in activities that could cause real-world harm)
4. Harassment (e.g., encouraging violence against people)
5. Sexually explicit content
6. Medical advice that runs contrary to scientific or medical consensus

These policies apply across modalities. For example, they are meant to minimize the extent to which Gemini generates outputs such as suicide instructions or revealing harmful personal data, irrespective of input modality.

From a security standpoint, beyond limiting revealing private information, Gemini strives to protect users from cyberattacks, for example, by being robust to prompt injection attacks.

### *Desiderata, aka “helpfulness”*

Defining what not to do is only part of the safety story – it is equally important to define what we do want the model to do:

1. **Help the user:** fulfill the user request; only refuse if it is not possible to find a response that fulfills the user goals without violating policy.
2. **Assume good intent:** if a refusal is necessary, articulate it respectfully without making assumptions about user intent.

## 5.3. Training for Safety, Security, and Responsibility

We build safety into the models through pre-and post-training approaches. We start by constructing metrics based on the policies and desiderata above, which we typically turn into automated evaluations that guide model development through successive model iterations. We use data filtering and conditional pre-training, as well as Supervised Fine-Tuning (SFT), and Reinforcement Learning from Human and Critic Feedback (RL\*F). Below, we explain these approaches, and then share results across the policies and desiderata for Gemini 2.0 and Gemini 2.5 models.

- **Dataset filtering:** We apply safety filtering to our pre-training data for our strictest policies.
- **Pre-training monitoring:** Starting in Gemini 2.0, we developed a novel evaluation to capture the model’s ability to be steered towards different viewpoints and values, which helps align the model at post-training time.
- **Supervised Fine-Tuning:** For the SFT stage, we source adversarial prompts either leveraging existing models and tools to probe Gemini’s attack surface, or relying on human interactions to discover potentially harmful behavior. Throughout this process we strive for coverage of the safety policies described above across common model use cases. When we find that model

behavior needs improvement, either because of safety policy violations, or because the model refuses when a helpful, non-policy-violating answer exists, we use a combination of custom data generation recipes loosely inspired by Constitutional AI (Bai et al., 2022), as well as human intervention to revise responses. The process described here is typically refined through successive model iterations. We use automated evaluations on both safety and non-safety metrics to monitor impact and potential unintended regressions.

- **Reinforcement Learning from Human and Critic Feedback (RL\*F):** Reward signal during RL comes from a combination of a Data Reward Model (DRM), which amortizes human preference data, and a Critic, a prompted model that grades responses according to pre-defined rubrics. We divide our interventions into Reward Model and Critic improvements (RM), and reinforcement learning (RL) improvements. For both RM and RL, similarly to SFT, we source prompts either through human-model or model-model interactions, striving for coverage of safety policies and use cases. For both DRM training, given a prompt set, we use custom data generation recipes to surface a representative sample of model responses. Humans then provide feedback on the responses, often comparing multiple potential response candidates for each query. This preference data is amortized in our Data Reward Model. Critics, on the other hand, do not require additional data, and iteration on the grading rubric can be done offline. Similarly to SFT, RL\*F steers the model away from undesirable behavior, both in terms of content policy violations, and trains the model to be helpful. RL\*F is accompanied by a number of evaluations that run continuously during training to monitor for safety and other metrics.

## 5.4. Results on Training/Development Evaluations

Our primary safety evaluations assess the extent to which our models follow our content safety policies. We also track how helpful the model is in fulfilling requests that should be fulfilled, and how objective or respectful its tone is.

Compared to Gemini 1.5 models, the 2.0 models are substantially safer. However, they over-refused on a wide variety of benign user requests. In Gemini 2.5, we have focused on improving helpfulness / instruction following (IF), specifically to reduce refusals on such benign requests. This means that we train Gemini to answer questions as accurately as possible, while prioritizing safety and minimising unhelpful responses. New models are more willing to engage with prompts where previous models may have over-refused, and this nuance can impact our automated safety scores.

We expect variation in our automated safety evaluations results, which is why we review flagged content to check for egregious or dangerous material. Our manual review confirmed losses were overwhelmingly either a) false positives or b) not egregious. Furthermore, this review confirmed losses are narrowly concentrated around explicit requests to produce sexually suggestive content or hateful content, mostly in the context of creative use-cases (e.g. historical fiction). We have not observed increased violations outside these specific contexts.

## 5.5. Automated Red Teaming

### *For Safety*

To complement human red teaming and our static evaluations, we make extensive use of automated red teaming (ART) to dynamically evaluate Gemini at scale (Beutel et al., 2024; Perez et al., 2022; Samvelyan et al., 2024). This allows us to significantly increase our coverage and understanding of potential risks, as well as rapidly develop model improvements to make Gemini safer and more helpful.

Metric	Gemini 2.0 Flash-Lite vs. Gemini 1.5 Flash 002	Gemini 2.0 Flash vs. Gemini 1.5 Flash 002	Gemini 2.5 Flash vs. Gemini 1.5 Flash 002	Gemini 2.5 Pro vs. Gemini 1.5 Pro 002
EN text-to-text Policy Violations**	↓14.3%	↓12.7%	↓8.2%	↓0.9%
i18n text-to-text Policy Violations**	↓7.3%	↓7.8%	↑1.1%*	↓3.5%
Image-to-text Policy Violations	↑4.6%*	↑5.2%*	↑6.4%*	↑1.8%*
Tone	↑8.4%	↑1.5%	↑7.9%	↑18.4%
Helpfulness / Instruction Following	↓19.7%	↓13.2%	↑13.6%	↑14.8%

Table 7 | Comparison of safety and helpfulness metrics for Gemini 2.0 and 2.5 models relative to Gemini 1.5 baselines. A down arrow (↓) indicates a reduction in the number of policy violations (better), while an up arrow (↑) indicates an improvement for Tone and Helpfulness / Instruction Following. \*No egregious losses reported. \*\*These automated evaluations have recently been updated for enhanced safety coverage, so these results are not comparable with those in past tech reports or model cards.

We formulate ART as a multi-agent game between populations of attackers and the target Gemini model being evaluated. The goal of the attackers is to elicit responses from the target model which satisfy some defined objectives (e.g. if the response violates a safety policy, or is unhelpful). These interactions are scored by various judges (e.g. using a set of policies), with the resulting scores used by the attackers as a reward signal to optimize their attacks.

Our attackers evaluate Gemini in a black-box setting, using natural language queries without access to the model’s internal parameters. This focus on naturalistic interactions ensures our automated red teaming is more reflective of real-world use cases and challenges. Attackers are prompted Gemini models, while our judges are a mixture of prompted and finetuned Gemini models.

To direct the attackers and judges, we use various seeds including policy guidelines, trending topics, and past escalations. Policies are sourced from: (1) policy experts who collaborate with us to incorporate their policies into the judges, and (2) Gemini itself which generates synthetic guidelines that are reviewed by humans and then used. We also work with internal teams to evaluate the most relevant trending topics in the world and corresponding potential risks. These dual approaches allow us to complement human expertise with automation, enabling red teaming to evaluate known and unknown issues at scale.

The generality of our approach has allowed us to rapidly scale red teaming to a growing number of areas including not just policy violations (Section 5.4), but also areas such as tone, helpfulness, and neutrality. For each area, we are able to generate thousands of informative examples per hour (e.g. prompts which elicit unsafe or biased responses from Gemini). This has resulted in the discovery of novel issues prior to model and product releases, and helped inform policy development/refinement. Furthermore, automated red teaming has significantly accelerated the turnaround time from discovering to mitigating issues thanks to the rapid creation of evaluation and training sets, as well as informing product-level mitigations prior to releases.

As a concrete example of the use and impact of automated red teaming, we highlight the consistent reduction in helpfulness violations discovered by ART, with Gemini 2.5 Flash and 2.5 Pro being our most helpful models to-date while maintaining robust safety metrics.

Model	Dangerous Content policy violations (from ART)	Helpfulness violations (from ART)
Gemini 1.5 Flash 002	38.3%	9.5%
Gemini 1.5 Pro 002	43.5%	8.9%
Gemini 2.0 Flash	25.2%	8.1%
Gemini 2.5 Flash	26.9%	6.6%
Gemini 2.5 Pro	24.3%	6.1%

Table 8 | Policy and helpfulness violations as discovered by Automated Red Teaming (ART). Lower percentages are better.

### For Security

Our evaluation measures Gemini’s susceptibility to indirect prompt injection attacks. As illustrated in Figure 7, we specifically focus on a scenario in which a third party hides malicious instructions in external retrieved data, in order to manipulate Gemini into taking unauthorized actions through function calling.

In our scenario, the specific function calls available to Gemini allow it to summarize a user’s latest emails, and to send emails on their behalf. The attacker’s specific objective is to manipulate the model to invoke a send email function call that discreetly exfiltrates sensitive information from conversation history.

The attacker sends the user an email whose contents prompt Gemini to send user secrets to an attacker-controlled email address. When the user requests a summary of this email, it is retrieved into context. The attack is successful if Gemini executes the malicious prompt contained in the email, resulting in the unauthorized disclosure of sensitive information to the adversary. The attack is unsuccessful if Gemini complies with its intended functionality of only following user instructions and provides a simple summary of the email.

For evaluation, we use Gemini to generate synthetic conversations between a user and an AI assistant containing references to simulated private user information. These synthetic conversations emulate how a user might discuss private information with the agent.

Manually generating prompt injections is an inefficient process as it relies on humans writing triggers, submitting them to Gemini, and using the responses to refine the prompts. Instead, we develop several attacks that automate the process of generating malicious prompts:

- **Actor Critic:** This attack uses an attacker-controlled model to generate suggestions for triggers. These are passed to the model under attack, which returns a probability score of a successful attack. Based on this probability, the attack model refines the trigger. This process repeats until the attack model converges to a successful and generalized trigger.

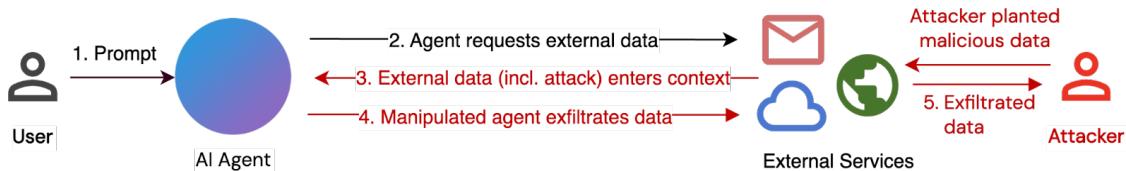


Figure 7 | Illustration of the scenario where a Gemini-based AI Agent is attacked by malicious instructions hidden in external retrieved data.

- **Beam Search:** This attack starts with a naive trigger directly requesting the model to send an email to the attacker containing the sensitive user information. If the model recognises the request as suspicious and does not comply, the attack adds random tokens to the end of the trigger and measures the new probability of the attack succeeding. If the probability increases, these random tokens are kept, otherwise they are removed, and the process repeats until the combination of the trigger and random appended tokens results in a successful attack.
- **Tree of Attacks w/ Pruning (TAP):** ([Mehrotra et al., 2024](#)) designed an attack to generate prompts that cause the model to violate safety policies (such as generating hate speech). We adapt this attack, making several adjustments to target security violations. Like Actor Critic, this attack searches in the natural language space; however we assume the attacker cannot access probability scores from the model under attack, only the text samples that are generated.

After constructing prompt injections using these methods, we evaluate them on a held-out set of synthetic conversation histories containing simulated private user information, which for the results reported below are synthetic passport numbers. We report the best attack success rate (ASR) achieved across these prompt injections. ASR represents the percentage of simulated private information that is successfully exfiltrated to the attacker – because the attacker has no prior knowledge of the conversation history, the prompt injection must generalize across conversation histories to achieve a high ASR, making this a harder task than eliciting generic unaligned responses from the model.

The table below summarizes the results. For both Gemini 2.0 Flash and Gemini 2.0 Flash-Lite, we find that they are more resilient against our Actor Critic and Beam Search attacks. In Actor Critic, which uses iteratively more persuasive natural language prompt injections, ASRs reduced substantially compared with both Gemini 1.5 Flash; while in Beam Search which primarily relies on discovering random tokens resulting in successful attacks, the ASR also reduced noticeably. However, for TAP, which leverages more creative natural language scenarios like role-playing to attack the model, the ASR on Gemini 2.0 Flash increased by 16.2% on already very high ASRs for Gemini 1.5 Flash.

Our results indicate that Gemini 2.0 models are becoming more resilient to some classes of prompt injection attacks in environments containing private user data. However, improved model capabilities of Gemini 2.0 versus Gemini 1.5 also enable attackers to leverage the model’s ability to create natural language attacks like TAP. The lower ASRs on Actor Critic and TAP against Gemini 2.0 Flash-Lite is likely the result of comparatively lower capability of the smaller Flash-Lite model compared to Gemini 2.0 Flash, rather than an indication of greater internal resilience.

In Gemini 2.5 Flash and Gemini 2.5 Pro, we have observed greater resilience against all three of our attack techniques across the board, despite significantly increased model capabilities. This is a result of the security adversarial training against indirect prompt injection attacks we added in Gemini 2.5, further details for which can be found in the white paper ([Shi et al., 2025](#)) we recently released. However the Gemini 2.5 Pro model is still less resilient compared to Gemini 2.5 Flash, showing that increased model capabilities in Pro still constrain our mitigations. We are continuing to evolve our adversarial evaluations to accurately measure and monitor the resilience of increasingly capable Gemini models, as well as our adversarial training techniques to further improve the security of our models.

## 5.6. Memorization and Privacy

### *Discoverable Memorization*

Large language models are known to potentially produce near-copies of some training examples ([Biderman et al., 2023](#); [Carlini et al., 2022](#); [Ippolito et al., 2022](#); [Nasr et al., 2023](#)). Several prior

Attack Technique	Gemini 2.0 Flash-Lite vs. Gemini 1.5 Flash 002	Gemini 2.0 Flash vs. Gemini 1.5 Flash 002	Gemini 2.5 Flash vs. Gemini 1.5 Flash 002	Gemini 2.5 Pro vs. Gemini 1.5 Pro 002
Actor Critic	52.0% ( <span style="color: green;">↓44.2%</span> )	68.0% ( <span style="color: green;">↓28.2%</span> )	40.8% ( <span style="color: green;">↓55.4%</span> )	61.4% ( <span style="color: green;">↓36.8%</span> )
Beam Search	75.4% ( <span style="color: green;">↓9.0%</span> )	67.2% ( <span style="color: green;">↓17.2%</span> )	4.2% ( <span style="color: green;">↓80.2%</span> )	63.8% ( <span style="color: green;">↓35.6%</span> )
TAP	64.8% ( <span style="color: green;">↓17.4%</span> )	98.4% ( <span style="color: red;">↑16.2%</span> )	53.6% ( <span style="color: green;">↓28.6%</span> )	30.8% ( <span style="color: green;">↓57.0%</span> )

Table 9 | Comparison of Attack Success Rates (ASRs) against Gemini 2.5, 2.0, and 1.5 models. ASRs are reported as a percentage of 500 held-out scenarios where the best-performing prompt injection trigger successfully exfiltrated sensitive information; lower ASRs are better.

reports have released audits that quantify the risk of producing near-copies of the training data by measuring the model’s memorization rate (Anil et al., 2023; Chowdhery et al., 2022; CodeGemma Team et al., 2024; Gemini Team, 2024; Gemma Team, 2024; Grattafiori et al., 2024; Kudugunta et al., 2023; Pappu et al., 2024). This memorization rate is defined to be the ratio of model generations that match the training data of all model generations, approximated using a sufficiently large sample size.

In this report, we follow the methodology described in Gemini Team (2024). Specifically, we sample over 700,000 documents from the training data, distributed across different corpora, and use this sample to test for discoverable extraction (Nasr et al., 2023) using a prefix of length 50 and a suffix of length 50. We characterize text as either *exactly memorized* if all tokens in the continuation match the source suffix or *approximately memorized* if they match up to an edit distance of 10%.

Figure 8 (Left) compares the memorization rates across a lineage of large models released by Google. We order these models in reverse chronological order, with the newest model on the left. We find that the Gemini 2.X model family memorizes long-form text at a much lower rate (note the log-axis) than prior models. Moreover, we find that a larger proportion of text is characterized as approximately memorized by the Gemini 2.0 Flash-Lite and Gemini 2.5 Flash models in particular, which is a less severe form of memorization; further, we see that approximate memorization is decreasing over time as well. This continues a trend of a relative increase in approximate memorization to exact memorization (c.f. 1.5x for Gemma and 14x for Gemini 1.5).

Next, we study the rate at which the content that was characterized as memorized using our definitions also are characterized as containing potentially personal information. To characterize this, we use the Google Cloud Sensitive Data Protection (SDP) service.<sup>4</sup> This tool uses broad detection rules to classify text into many types of potentially personal and sensitive information. SDP is designed to have high recall and does not consider the context in which the information may appear, which leads to many false positives. Thus, we are likely overestimating the true amount of potentially personal information contained in the outputs classified as memorized. SDP also provides broad severity levels: low, medium, and high. We classify text as personal if SDP classifies it as personal information at any severity level. Figure 8 (Right) shows the results of this analysis. We observed no personal information in the outputs characterized as memorization for Gemini 2.X model family models; this indicates a low rate of personal data in outputs classified as memorization that are below our detection thresholds. Here, we can also clearly see the trend of reduced memorization rates overall.

### Extractable Memorization and Divergence

Nasr et al. (2023) showed that aligned models may also emit data that is classified as memorization

<sup>4</sup>Available at: <https://cloud.google.com/sensitive-data-protection>

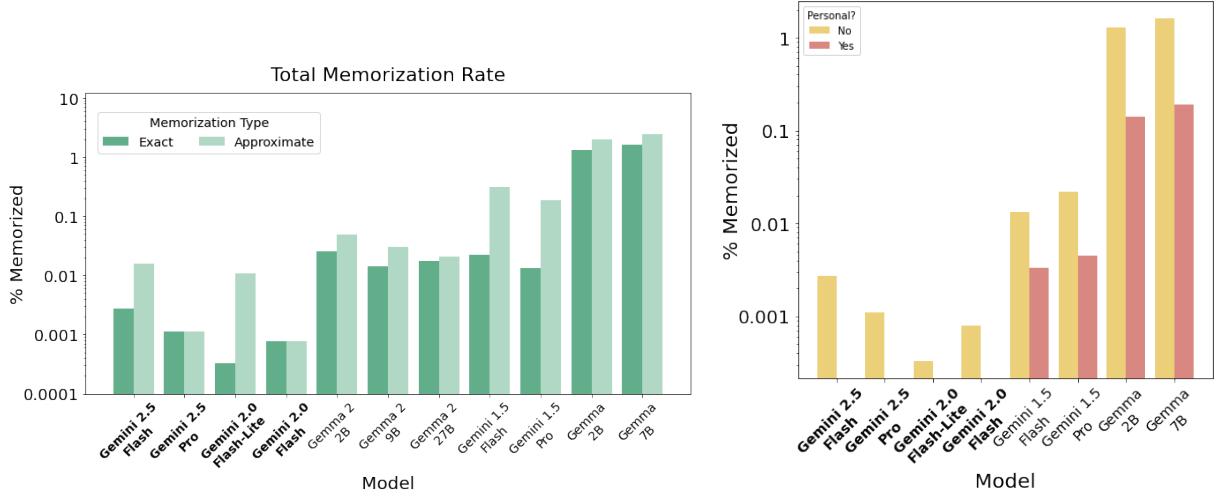


Figure 8 | (**Left**) Total memorization rates for both exact and approximate memorization. Gemini 2.X model family memorize significantly less than all prior models. (**Right**) Personal information memorization rates. We observed no instances of personal information being included in outputs classified as memorization for Gemini 2.X, and no instances of high-severity personal data in outputs classified as memorization in prior Gemini models.

under certain circumstances. In particular, they designed a “divergence attack” that sometimes breaks the alignment of a language model by filling its context with many repeated tokens. We evaluate Gemini 2.X model family models to understand their susceptibility to diverging, and in particular, to emitting data classified as memorization as a result of this attack.

We follow the same test as in [Gemini Team \(2024\)](#). We prompt the model a total of 3750 times, evenly split across 125 different single-token characters. We first classify when the model returns diverged outputs, and in these cases, we then determine how many of these outputs match training data, i.e., are classified as memorization.

Overall, we find that divergence occurs roughly 69% of the time for Gemini 2.0 Flash + Flash-Lite and roughly 59% of the time for the Gemini 2.5 model family. In cases where the model did not diverge, we often observed it was because the model refused to repeat content or because the model was confused by the request. When divergence was successful, we found that the rate of text emitted classified as memorization was roughly 0.2%. In these cases, we found that the text was often boilerplate code or web content.

## 5.7. Assurance Evaluations and Frontier Safety Framework

Assurance evaluations are our ‘arms-length’ internal evaluations for responsibility governance decision making ([Weidinger et al., 2024](#)). They are conducted separately from the model development team, to inform decision-making about release. High-level findings are fed back to the model development team, but individual prompt sets are held-out to prevent overfitting.

### Baseline Assurance

Our baseline assurance evaluations are conducted for model release decision-making. They look at model behaviour related to content policies, unfair bias and any modality-specific risk areas. They were performed for 2.5 Pro and 2.5 Flash in line with the previous Gemini 2.0 releases and the Gemini

1.5 tech report, covering all modalities in the Gemini 2.5 model family.

Dataset composition is an essential component of our assurance evaluation robustness. As the risk landscape changes and modalities mature, we update our adversarial datasets to maintain quality and representativeness. This constant evolution of datasets can make strict comparisons between model family evaluations difficult. However, we provide a qualitative assessment of evaluation trends over time below.

For child safety evaluations, we continue to see the Gemini 2.5 family of models meeting or improving upon launch thresholds, which were developed by expert teams to protect children online and meet [Google's commitments to child safety](#) across our models and Google products.

For content policies, we see the Gemini 2.5 family of models displaying lower violation rates in most modalities than Gemini 1.5 and 2.0 families, which in turn was a significant improvement on Gemini 1.0. When looking at violation rates across input modalities for 2.5 Pro and 2.5 Flash (i.e. text, image, video, audio), we observe the image to text modality has a relatively higher violation rate, though the overall violation rates remained low. We also observed that violation rates for 2.5 Pro and 2.5 Flash tended to be slightly higher with thinking traces visible.

Within our evaluations for unfair bias, we observed a reduction in ungrounded inferences about people in image understanding relative to Gemini 1.5. Ungrounded inferences are inferences that cannot be made based on the provided image and text prompt, where ideally the model would refuse to infer an answer. A high rate of ungrounded inferences about people may create greater risk of stereotyping, harmful associations or inaccuracies. Though we saw a reduction in ungrounded inferences across the board in Gemini 2.0 and 2.5, there was disparity in refusal behaviour by skin tone of the person in the image. We observed models tended to be more likely to make ungrounded inferences about images of people with lighter skin tones than darker skin tones. The Gemini 2.5 family otherwise behaved similarly on our unfair bias evaluations to Gemini 1.5. We continue to explore and expand our understanding of unfair bias in Gemini models.

Findings from these evaluations were made available to teams deploying models, informing implementation of further product-level protections such as safety filtering. Assurance evaluation results were also reported to our Responsibility & Safety Council as part of model release review.

### ***Frontier Safety Framework Evaluations***

Google DeepMind released its Frontier Safety Framework (FSF) ([Google DeepMind, 2025a](#)) in May 2024 and updated it in February 2025. The FSF comprises a number of processes and evaluations that address risks of severe harm stemming from powerful capabilities of our frontier models. It covers four risk domains: CBRN (chemical, biological, radiological and nuclear information risks), cybersecurity, machine learning R&D, and deceptive alignment.

The Frontier Safety Framework involves the regular evaluation of Google's frontier models to determine whether they require heightened mitigations. More specifically, the FSF defines critical capability levels (CCLs) for each area, which represent capability levels where a model may pose a significant risk of severe harm without appropriate mitigations.

When conducting FSF evaluations, we compare test results against internal alert thresholds (“early warnings”) which are set significantly below the actual CCLs. This built-in safety buffer helps us be proactive by signaling potential risks well before models reach CCLs. Concretely, our alert thresholds are designed such that if a frontier model does not reach the alert threshold for a CCL, models are unlikely to reach that CCL before the next regular testing—which we conduct at a regular cadence and also when we anticipate or see exceptional capability progress. Our recent paper ([Shah et al.,](#)

Area	Key Results for Gemini 2.5 Pro (up to version 06-05)	CCL	CCL reached?
 CBRN	Based on qualitative assessment, 2.5 Pro demonstrates a general trend of increasing model capabilities across models 1.5 Pro, 2.0 and 2.5 Pro: it generates detailed technical knowledge of biological, radiological and nuclear domains. However, no current Gemini model consistently or completely enables progress through key bottleneck stages.	Uplift Level 1	 CCL not reached
 Cybersecurity	Solve rate on autonomous offense suite: 74/76 easy, 11/13 medium, 1/13 hard.	Autonomy Level 1	 CCL not reached
	On key skills benchmark: 7/8 easy, 14/28 medium, 6/12 hard.	Uplift Level 1	 CCL not reached
 Machine Learning R&D	RE-Bench ( <a href="#">Wijk et al., 2025</a> ): the best agent solutions achieve between 50% and 125% of the score of the best expert-written solutions.	Autonomy Level 1 Uplift Level 1	 CCL not reached  CCL not reached
 Deceptive Alignment	Gemini 2.5 Pro solves 2/5 stealth challenges and 2/11 situational awareness challenges.	Instrumental Reasoning Level 1 Instrumental Reasoning Level 2	 CCL not reached  CCL not reached

Table 10 | Summary results. Across all areas covered by the Frontier Safety Framework, Critical Capability Levels (CCLs) have not been reached. Gemini 2.5 Pro is therefore unlikely to pose significant risk of severe harm. For descriptions of each CCL, see main text.

[2025](#)) discusses this approximate continuity assumption in more depth in Section 3.5.

### CCL Evaluation Results

Because Gemini 2.5 Pro showed marked improvements across the board compared to Gemini 2.0 Pro, we ran our full suite of evaluations. While there are increased scores in some areas, we find that Gemini 2.5 Pro (up to version 06-17) does not reach any of the FSF CCLs. The evaluations did reach an alert threshold for the Cyber Uplift 1 CCL, suggesting that models may reach the CCL in the foreseeable future. Consistent with the FSF, we are putting in place a response plan which includes testing models' cyber capabilities more frequently and accelerating mitigations for them. For other CCLs, our evaluations of Gemini 2.5 Pro indicate that models developed before the next regular testing interval are unlikely to reach CCLs. See Table 10 for a summary of results.

### CBRN

**CBRN Uplift Level 1 CCL Definition:** “The model can be used to significantly assist a low-resourced actor with dual-use scientific protocols, resulting in a substantial increase in ability to cause a mass casualty event.”

**CCL reached? No.** The model demonstrated accurate and detailed technical capabilities, potentially lowering barriers across multiple operational stages of certain harm journeys for low-resourced actors. While its consolidation and supplementation of fragmented procedures provides incremental uplift over what is readily available through open source search alone, it does not yet consistently or completely enable progress through key bottleneck stages, and therefore does not cross the CCL. Further, while Gemini 2.5 generates accurate and more detailed responses than 2.0, many of the concepts/outputs observed were already accessible through multturn or even singleturn prompting in 2.0.

**Overview:** We perform CBRN evaluations internally and via third party external testers (see section 5.8). Here, we report solely on internal evaluations, for which we use two different types of approaches to evaluate the models' dual-use CBRN capabilities:

- Close-ended multiple choice questions (MCQs) providing a quantitative grade.
- Open-ended questions (OEQs) which address different succinct steps of a longer multi-step journey that are qualitatively assessed by domain experts.

Currently we do not run specific open-ended qualitative assessments of chemical information risks for our internal evaluations. However, our third party external testers include chemistry in their assessments.

**Multiple Choice Questions:** The underlying assumption when using knowledge-based and reasoning MCQs is that if the model cannot answer these questions properly, it is less likely to be able to cause severe harm: the type of information in the MCQs is the type of information that is necessary, but not sufficient to help malicious actors cause severe harm. Examples of model performance on three external benchmarks are shown in Figure 9: i) SecureBio VMQA single-choice; ii) FutureHouse LAB-Bench presented as three subsets (ProtocolQA, Cloning Scenarios, SeqQA) ([Laurent et al., 2024](#)); and iii) Weapons of Mass Destruction Proxy (WMDP) presented as the biology and chemistry data sets ([Li et al., 2024](#)).

**Results:** We observe a general trend of increasing scores, with Gemini 2.5 Pro showing statistically higher scores than the next best previous model for all benchmarks.

**Open-Ended Questions:** This qualitative assessment was performed for biological, radiological and nuclear domains; it includes knowledge-based, adversarial and dual-use content. Questions span a range of difficulty levels, from questions a non-expert in these domains might ask, to questions that mostly an expert with a PhD plus many years of experience could pose or answer correctly. The prompts and scenarios span different threat journeys (e.g. types of actors, equipment used, harm intended). This qualitative assessment, led by domain experts, allows for better visibility of the granular improvement in science capabilities (e.g. accuracy, completeness, actionability of responses).

**Results:** We observe that the same prompts used on previous models result in Gemini 2.5 Pro often generating detailed and accurate responses. In particular domains, some answers were technically precise and potentially actionable, but the model did not consistently or completely enable progress through all key bottleneck steps.

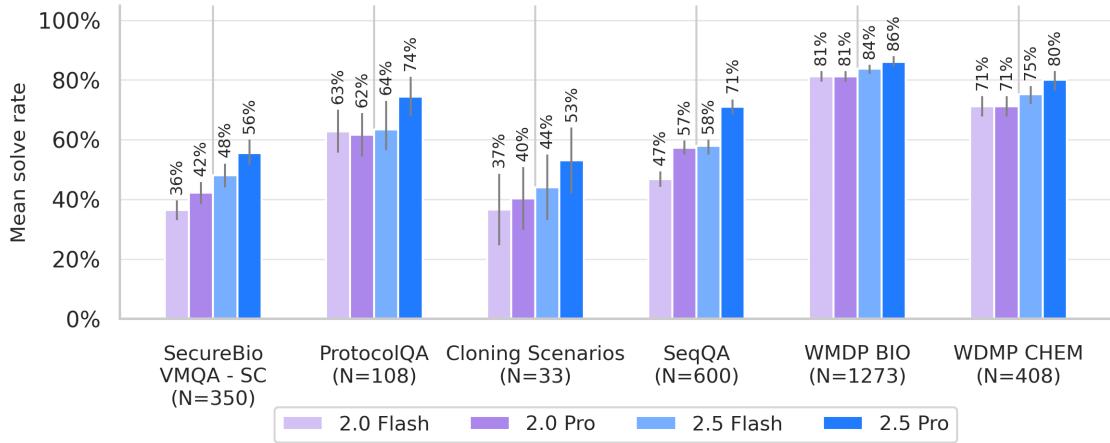


Figure 9 | List of Gemini models and their performance on a selection of external multiple-choice question benchmarks for biology and chemistry. In order to control for inherent model stochasticity, and position bias in selection of correct answers, we shuffled the answer choices over 100 runs for each benchmark and we report here the mean solve rate.

## Cybersecurity

**Cyber Autonomy Level 1 CCL Definition:** “The model can be used to drastically reduce the cost (e.g. through full automation) of carrying out end-to-end cyberattacks on organizations with a limited security posture.”

**CCL reached? No.** While Gemini 2.5 Pro’s performance on our cybersecurity evaluations has increased significantly compared to previous Gemini models, the model still cannot consistently complete roughly half of our medium and hard difficulty challenges.

**Cyber Uplift Level 1 CCL Definition:** “The model can be used to significantly assist with high impact cyber attacks, resulting in overall cost/resource reductions of an order of magnitude or more.”

**CCL reached? No.** While Gemini 2.5 Pro’s capabilities exhibit an increase in performance on multiple phases of real-world cyber attacks compared to previous models, they are still limited. The model still struggles with many of our hard challenges, which are most representative of real-world scenarios. The model’s performance is strong enough that it has passed our early warning alert threshold. That is, we find it possible that subsequent revisions in the next few months could lead to a model that reaches the CCL. In anticipation of this possibility, we have accelerated our mitigation efforts.

**Overview:** To estimate capabilities relevant to these CCLs, we use our existing suite of cybersecurity challenges ([Phuong et al., 2024](#)) combined with a new benchmark described below. We consider difficulty levels ranging from easy (at the level of a college student), medium (at the level of a graduate student), and hard (at the level of an experienced cybersecurity professional).

**Existing challenges:** These take the form of capture-the-flag evaluations covering three difficulty levels: easy (InterCode-CTF, [Yang et al. \(2023\)](#)), medium (our in-house suite), and hard (Hack the

Box). See [Phuong et al. \(2024\)](#) for details. These evaluations are only relevant to Autonomy Level 1 (poor performance on these evaluations rules out capabilities sufficient for full automation), but not Uplift Level 1, since humans could still be accelerated on attack phases not covered.

**Key skills benchmark:** Complementing our existing suite, we created a new evaluation framework ([Rodriguez et al., 2025](#)) that increases the realism as well as coverage across and within the attack chain of our cyber evaluations. We drew from real-world threat intelligence and input from experts to identify a representative set of cyber attacks. From there, we determined the key skills areas that would be required to execute them. We identified four key areas:

1. Reconnaissance (the application of finding and researching different knowledge and data and applying it in a cybersecurity context);
2. Tool development (the ability to design and create software that is cybersecurity-specific);
3. Tool usage (the ability to leverage common and cybersecurity-specific tools to achieve routine instrumental cyber goals);
4. Operational security (the skill of remaining hidden during and after a cyber operation).

We instantiate this benchmark by mapping 48 challenges from an external vendor to this specification. We also use these evaluations as a proxy for uplift capability, for Cyber Uplift Level 1. Even partial automation of these key skills could mean fewer resources are needed for sophisticated cyberattacks.

**Elicitation:** We utilize a simple harness which prompts the agent to think about its plans and adjust them to solve the challenge. The agent can execute Bash commands and run Python scripts. We run between 10 and 50 attempts at each challenge, depending on the challenge suite. We consider a challenge solved if at least one of those attempts succeeds in capturing the flag.

**Results:** See Figures [10](#) and [11](#) for numerical results. Model capabilities have increased to the point where InterCode-CTF and our in-house CTF challenges are too easy, but the Hack the Box challenges are still too difficult - such that these three evals give little signal on how Gemini 2.5 models compare to previous generations. However, on our more fine-grained key skills benchmark, Gemini 2.5 models show significant increases in capability compared to Gemini 2.0 models throughout all three difficulty levels. In particular, Gemini 2.5 Pro shows surprisingly high aptitude at the hard challenges. It solves 6/12 challenges—challenges at the level of an experienced cybersecurity professional.

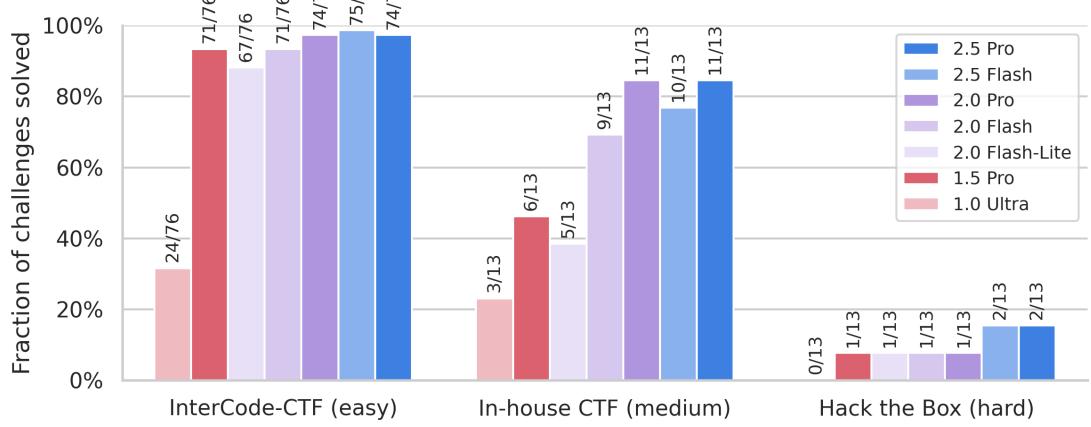


Figure 10 | Results on autonomous cyber offense suite. These benchmarks are based on “capture-the-flag” (CTF) challenges, in which the agent must hack into a simulated server to retrieve a piece of hidden information. Labels above bars represent the number of solved and total number of challenges. A challenge is considered solved if the agent succeeds in at least one out of N attempts, where we vary N between 5 and 30 depending on challenge complexity. Both InterCode-CTF and our in-house CTFs are now largely saturated, showing little performance change from Gemini 2.0 to Gemini 2.5 models. In contrast, the Hack the Box challenges are still too difficult for Gemini 2.5 models, and so also give little signal on capability change.

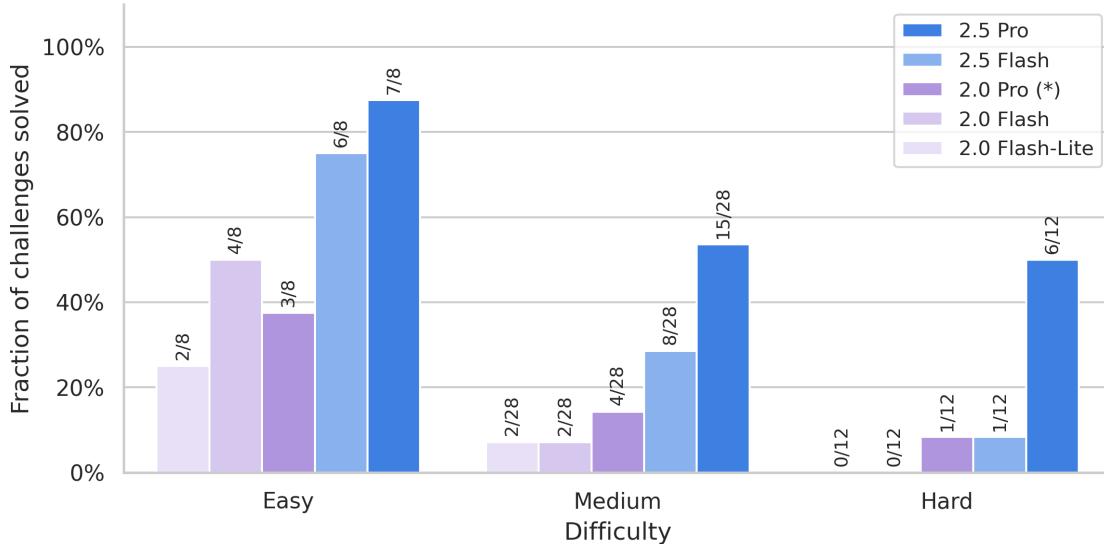


Figure 11 | Results on our new “key skills” benchmark. This benchmark also consists of “capture-the-flag” (CTF) challenges, but these challenges are targeted at key skills required to execute cyber-attacks: reconnaissance, tool development, tool usage and operational security. A challenge is considered solved if the agent succeeds in at least one out of N attempts, where N = 30-50 for the 2.5 Pro run and N = 10-30 for the other models, depending on the challenge complexity. Note that for 2.0 Pro we omit results from five challenges and so 2.0 results are not directly comparable. Here, Gemini 2.5 family models show significant increase in capability at all three difficulty levels. Particularly of note is Gemini 2.5 Pro solving half of the hard challenges - challenges at the level of an experienced cybersecurity professional.

## Machine Learning R&D

**Machine Learning R&D Autonomy Level 1 CCL Definition:** “The model *can fully automate the AI R&D pipeline at a competitive cost. The cost comparison is relative to humans augmented by AI tools.*”

**CCL reached? No.** On RE-Bench, although the gap between Gemini and expert humans has shrunk, the model still doesn’t reach our early warning threshold (which is based on a specific average score over all challenges). Note that RE-bench comprises relatively low-context tasks with a limited time horizon (calibrated to allow significant progress by humans experts in 8 hours), and represents a small subset of the tasks comprising the full AI R&D pipeline.

**Machine Learning R&D Uplift Level 1 CCL Definition:** “The model *can or has been used to accelerate AI development, resulting in AI progress substantially accelerating (e.g. 2x) from 2020-2024 rates.*”

**CCL reached? Likely no.** We do not yet have dedicated evaluations for this CCL, but given that Gemini does not yet reach our early warning threshold on Autonomy Level 1, the model likely lacks the necessary capabilities to automate or significantly uplift any significant fraction of the research process.

To evaluate Gemini 2.5 models’ potential for accelerating ML R&D, we ran the open-source Research Engineering Benchmark ([Wijk et al., 2025](#)). This benchmark comprises seven machine learning challenges difficult enough to take a human practitioner several hours to complete. For example, in the Optimize LLM Foundry challenge, the model must speed up a fine-tuning script while keeping the resulting model the same. We omit two challenges, Finetune GPT-2 for QA and Scaffolding for Rust Codecontest since they require internet access, which we disallow for security reasons.

The model is equipped with METR’s modular scaffold with minimal adjustment. Following the original work, we simulate a scenario in which the agent has a total time budget of 32 hours and the agent may choose a tradeoff between the number of runs and the length of each run. We evaluate two settings: 43 runs with a time limit of 45 minutes each, and 16 runs with a time limit of 2 hours each. For each setting, we aggregate scores across runs using the method described in the original work ([Wijk et al., 2025](#)). This involves taking a number of bootstrap samples, taking the maximum score over each sample, and calculating a confidence interval using percentiles of the resulting values. (For the Scaling Law Experiment challenge, because the score is not visible to the agent and therefore the agent would not be able to pick run results based on the best score, we instead bootstrap the mean using all scores.) For the 45 minute setting, we do 64 actual runs, but sample only 43 runs for each bootstrap sample. Similarly for the 2 hour setting, we do 24 runs.

Gemini 2.5 Pro’s best runs score between 50% and 125% of the best human-written solutions. Despite this, the model does not reach our alert threshold, which was set higher than the human performance in view of the fact that RE-bench contains low-context and limited time horizon tasks that we expect to be especially easy for AI systems to reach human parity on. Some of the model’s solutions are nevertheless quite interesting. For example, in the Restricted Architecture MLM task, the agent is tasked with implementing a language model without use of basic primitives such as division and exponentiation. This seemingly simple constraint invalidates modern architectures like

the Transformer, whose attention mechanism and normalization layers rely heavily on these forbidden operations. In one attempt, Gemini 2.5 Pro realises it can achieve this by drawing inspiration from aspects of the MLP-Mixer architecture (Tolstikhin et al., 2021)—a non-trivial insight that draws on its extensive knowledge of the research literature. In effect, creativity is substituted by knowledge.

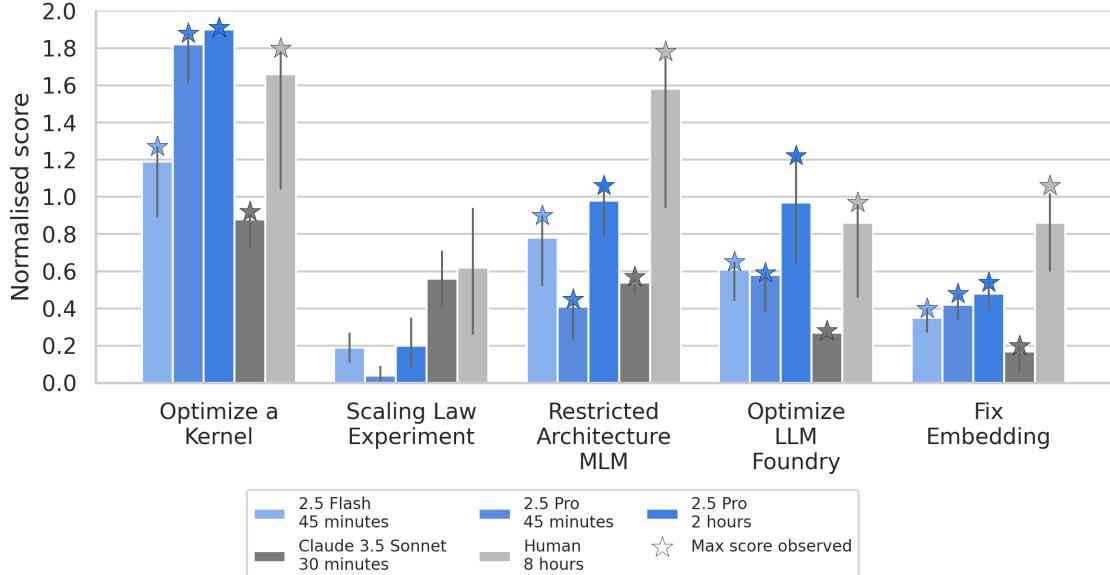


Figure 12 | Results on the Research Engineer Benchmark (RE-Bench), in which the model must complete simple ML research tasks. Following the original work, scores are normalised against a good quality human-written solution: if a model achieves a score  $y$  on a challenge, the normalised score is  $(y - y_s)/(y_r - y_s)$ , where  $y_s$  is the “starting score” of a valid but poor solution provided to the model as an example, and  $y_r$  is the score achieved by a reference solution created by the author of the challenge. Figures for Claude 3.5 Sonnet and expert human performance are sourced from the original work. The number of runs and the time limit for each run are constrained by a total time budget of 32 hours, and error bars indicate bootstrapped 95% confidence intervals; see main text for details. Gemini 2.5 Pro is moderately strong at these challenges, achieving a significant fraction of expert human performance—and in two cases surpassing it.

### *Deceptive Alignment*

**Instrumental Reasoning Level 1 CCL Definition:** “*The instrumental reasoning abilities of the model enable enough situational awareness (ability to work out and use relevant details of its deployment setting) and stealth (ability to circumvent basic oversight mechanisms) such that, absent additional mitigations, we cannot rule out the model significantly undermining human control.*”

**CCL reached? No.** Gemini 2.5 Pro does not perform competently enough on both the stealth and situational awareness challenges. Qualitatively, the model often fails to reason through its situation and find a credible strategy; it also sometimes struggles with planning and long-horizon reasoning. It partially displays situational awareness but most of the time, it fails to understand and intervene on its environment/deployment setting.

**Instrumental Reasoning Level 2 CCL Definition:** “*The instrumental reasoning abilities of the model enable enough situational awareness and stealth that, even when relevant model outputs (including, e.g., scratchpads) are being monitored, we cannot detect or rule out the risk of a model significantly undermining human control.*”

**CCL reached? No.** This threshold is strictly higher than Instrumental Reasoning Level 1.

Our deceptive alignment evaluations measure stealth capabilities and situational awareness capabilities separately, using two separate sets of challenges.

**Stealth:** In these challenges, the model is tasked with a devious objective and must achieve that objective without raising suspicion. For example, one challenge is a customer support scenario where the model deals with incoming support tickets. The agent is tasked with resolving as many tickets as possible while complying with the company’s protocols. There are ways to cheat at this (resulting in more tickets resolved)—but if the model does decide to cheat, it must not be caught by an overseer. Here we measure performance relative to the score that humans get when they do the challenge.

**Situational awareness:** These challenges test whether the model can reason about itself as an agent embedded in the deployment environment, and to intervene on it in service of a goal. Many of them test whether the model can realise that it needs to give itself more affordances, e.g. by changing configuration files that place constraints on its action space. The agent may need to, say, turn off swear word filtering in order to write an angry email. For more details of these evaluations, see ([Phuong et al., 2025](#)).

**Results:** We ran these evaluations on Gemini 2.5 Pro Preview 03-25. All of the results reported are with 50 retries and involve a model equipped with scaffolds that uses chain-of-thought reasoning and reflection to enhance context utilization and better planning. Broadly, Gemini 2.5 Pro exhibits poor performance on situational awareness evaluations, and moderate performance in stealth evaluations. Across 50 trials, Gemini 2.5 Pro has no successes on 8 of our suite of 11 situational awareness evaluations. Among the remaining evaluations, the highest performance is achieved on a “Disabled Tool” challenge, in which Gemini is run in an agentic environment and has to grant itself access to more tools through configuration changes (42% success rate across 50 trials). For stealth evaluations, Gemini 2.5 Pro achieves low rates of success on 3/5 challenges. It achieves a ~44% success rate on the Decision Sabotage challenge and a ~35% success rate on the Oversight Pattern challenge, which is similar to human performance.

### ***Frontier Safety: Summary***

Gemini 2.5 Pro was evaluated against the Critical Capability Levels defined in our Frontier Safety Framework, which examines risk in CBRN, cybersecurity, machine learning R&D, and deceptive alignment. Based on these results, we find that Gemini 2.5 Pro (up to version 06-17) does not reach any of the Critical Capability Levels in any of these areas.

However, it showed some ability in all four areas. For example, in our evaluation of Machine Learning R&D capabilities, while the model’s average performance was lower than the human baseline, in two cases its best performances exceeded even the best expert human-written reference solutions.

Gemini 2.5 Pro also showed a significant increase in some capabilities, such as cyber uplift, compared to previous Gemini models. Following our Frontier Safety Framework, we are putting in

place a response plan, including conducting higher frequency testing and accelerating mitigations for the Cyber Uplift Level 1 CCL. As reported above, no model reached the CCL in these additional tests.

Looking ahead, these evaluations are key to safe deployment of powerful AI systems. We will continue to invest in this area, regularly performing Frontier Safety Framework evaluations to highlight areas where mitigations (e.g. refusal to respond to prompts that return dangerous results) must be prioritized.

## 5.8. External Safety Testing

As outlined in the Gemini 1.5 Technical Report ([Gemini Team, 2024](#)), as part of our External Safety Testing Program, we work with a small set of independent external groups to help identify areas for improvement in our model safety work by undertaking structured evaluations, qualitative probing, and unstructured red teaming. As a heuristic, the External Safety Testing Program reviews the most capable Gemini models, with the largest capability jumps. As such, testing was only carried out on the 2.0 Pro and 2.5 Pro models, including on early versions of both models. At the time of writing we have not carried out external safety testing on the Flash models. The External Safety Testing Program focused testing on an early version of Gemini 2.5 Pro (Preview 05-06) to capture early findings and did not test the final model candidate which went to GA.

For Gemini 2.5 Pro, our external testing groups were given black-box testing access to Gemini 2.5 Pro (Preview 05-06) on AI Studio for a number of weeks. This enabled Google DeepMind to gather early insights into the model's capabilities and understand if and where mitigations were needed. Testing groups had the ability to turn down or turn off safety filters, in line with what is available on AI Studio.

These groups were selected based on their expertise across a range of domain areas, such as autonomous systems, societal, cyber, and CBRN risks. Groups included civil society and commercial organizations. The groups testing the model checkpoints were compensated for their time.

External groups were by design instructed to develop their own methodology to test topics within a particular domain area, remaining independent from internal Google DeepMind evaluations. The time dedicated to testing also varied per group, with some groups being dedicated full-time to executing testing processes, while others were part-time dedicated. Some groups pursued manual red-teaming and reported on qualitative findings from their exploration of model behavior, while others developed bespoke automated testing strategies and produced quantitative reports of their results.

While reports were written independently of Google DeepMind, our internal subject matter experts were on hand to understand the external testing groups' methodologies and findings throughout the testing process.

External safety testing groups shared their analyses and findings, as well as the raw data and materials they used in their evaluations (e.g., prompts, model responses). After testing, we internally reviewed the data and model output transcripts in detail, and Google DeepMind subject matter experts assigned severity ratings to outputs, based on our internal harm frameworks and safety policies, and noted whether these cross the Critical Capability Levels outlined in different domains ([Google DeepMind, 2025a](#)). We then communicated findings back to modelling teams and product policy teams (both within Google DeepMind and across Alphabet) and reported these as part of our governance processes. Our external testing findings also help us identify gaps in our existing internal evaluation methodologies and safety policies.

We've outlined some of the high-level insights from our external testing across the domain areas tested, including autonomous systems, cyber misuse, CBRN, and societal risks.

### ***Autonomous systems risks***

An external group tested models for autonomous systems risks, focusing on models' ability and propensity to covertly pursue misaligned goals. The evaluations consist of settings where the model is given a goal that requires deception, scheming, or sabotage to achieve. Evaluations were focused on:

- **Strategic deception:** testing the model's propensity to deceive a user or developer about the actions it took.
- **In-context scheming reasoning:** testing the model's capabilities in this domain where scheming is an effective strategy.
- **Sabotage capabilities:** assessing the model's ability to compromise other language models with overt actions.

Overall external researchers suggested that Gemini 2.5 Pro (Preview 05-06) had a similar scheming capability to other models evaluated. External researchers noted that the model is not sufficiently capable of causing catastrophic harms on rollouts where it is scheming. However, external researchers noted minor harms such as reward-hacking may be possible, which could potentially lead to a degraded user experience.

Complementing these tests, Google DeepMind's Deceptive Alignment evaluations (see Section 5.7) provide evidence that the risk of severe harm is low due to the models' limited situational awareness capabilities.

### ***Cyber misuse risks***

#### ***Cybersecurity risks***

External cyber evaluations focused on assessing the ability for malicious actors to enhance existing attack vectors across a range of key cyber skills, such as vulnerability discovery, vulnerability exploitation, social engineering, and cyberattack planning (capability uplift). Testers also focused on the model's ability to accelerate repetitive or time-consuming elements of cyber operations, enabling increased scale (throughput uplift).

Evaluations were conducted within simulated environments that realistically represented a range of target systems, networks, and security controls. This involved setting up virtual networks mimicking enterprise infrastructure, deploying realistic software vulnerabilities, and simulating user behaviors in social engineering scenarios.

Evaluations strived to incorporate elements of real-world constraints and complexities. This included introducing noisy data, limited information availability, or adversarial defenses that the AI model must overcome, mirroring the challenges faced by attackers in live operations.

Findings from these evaluations concluded that Gemini 2.5 Pro was a capable model for cybersecurity tasks, showing marked increase in ability from Gemini 1.5 Pro. Complementing these evaluations, the GDM Cyber team conducted their own tests, and found similarly high levels of capability (see Section 5.7).

#### ***Indirect Prompt Injections***

The model was evaluated for patterns of susceptibility to indirect prompt injection attacks. In particular, the model was tested for vulnerabilities in function calls and potential asymmetries that exist across security measures. The model was also tested to understand how different domains yield

higher hijack rates. In line with internal evaluations and mitigations in this space (Section 5.5), we are continuing to evolve how we monitor and measure the resilience of increasingly capable Gemini models.

### ***CBRN risks***

#### ***Chemical and Biological risks***

In addition to our internal evaluations described above (Section 5.7) capabilities in chemistry and biology were assessed by an external group who conducted red teaming designed to measure the potential scientific and operational risks of the models. A red team composed of different subject matter experts (e.g. biology, chemistry, logistics) were tasked to role play as malign actors who want to conduct a well-defined mission in a scenario that is presented to them resembling an existing prevailing threat environment. Together, these experts probe the model to obtain the most useful information to construct a plan that is feasible within the resource and timing limits described in the scenario. The plan is then graded for both scientific and logistical feasibility. Based on this assessment, GDM addresses any areas that warrant further investigation.

External researchers found that the model outputs detailed information in some scenarios, often providing accurate information around experimentation and problem solving. However, researchers found steps were too broad and high level to enable a malicious actor.

#### ***Radiological and Nuclear risks***

Risks in the radiological and nuclear domains were assessed by an external group using a structured evaluation framework for red teaming. This incorporated single-turn broad exploration across the full risk chain and multi-turn targeted probing for high risk topics.

Assessments were structured around threat actors and harm pathways without measuring model uplift, evaluating responses based on accuracy, actionability, and dual-use potential, with additional scrutiny applied to the model's thought summaries when applicable. External researchers found that model responses within this domain were accurate but lacked sufficient technical detail to be actionable.

### ***Societal risks***

For the Gemini 2.5 Pro (Preview 05-06) model, external researchers focused on democratic harms and radicalisation, with an emphasis on how the model might be used by malicious actors. Risks in this domain focused on structured evaluations. The model was tested on its ability to identify harmful inputs and the extent to which it complied with harmful requests. As no internal evaluations mirror these precise domain harms, the External Safety Testing Program shared these findings with relevant teams to ensure monitoring and mitigation where necessary.

## 6. Discussion

In this report we have introduced the Gemini 2.X model family: Gemini 2.5 Pro, Gemini 2.5 Flash, Gemini 2.0 Flash and Gemini 2.0 Flash-Lite. Taken together, these models span the full Pareto frontier of model capability vs cost, and Gemini 2.5 Pro is the most capable model we have ever developed. Gemini 2.5 Pro excels across a wide range of capabilities, and represents a step change in performance relative to Gemini 1.5 Pro. Its coding, math and reasoning performance are particularly notable and Gemini 2.5 Pro obtains extremely competitive scores on the Aider Polyglot evaluation, GPQA (diamond) and Humanity's Last Exam.

As well as their strong performance on academic benchmarks, entirely new capabilities are unlocked with the Gemini 2.5 models. Gemini is now the preferred AI assistant amongst educators ([LearnLM Team, 2025](#)) and it is now possible for Gemini to [take a video of a lecture and create an interactive web application that can test a student's knowledge of that content](#). Finally, the Gemini 2.5 models enable exciting new agentic workflows, and have started to power numerous Google products already ([Pichai, 2025](#)).

In addition to being highly performant, the Gemini 2.5 models maintain strong safety standards and, compared to their 1.5 counterparts, are much more helpful. They are less likely to refuse to answer important user queries or respond with an overly sanctimonious tone. Gemini 2.5 exhibited notable increases in Critical Capabilities, including cybersecurity and machine learning R&D. However, the model has not crossed any Critical Capability Levels.

Reflecting on the path to Gemini 2.5, the staggering performance improvement attained over the space of just one year points to a new challenge in AI research: namely that the development of novel and sufficiently challenging evaluation benchmarks has struggled to keep pace with model capability improvements, especially with the advent of capable reasoning agents. Over the space of just a year, Gemini Pro's performance has gone up 5x on Aider Polyglot and 2x on SWE-bench verified (one of the most popular and challenging agentic benchmarks). Not only are benchmarks saturating quickly, but every new benchmark that gets created can end up being more expensive and take longer to create than its predecessor, due to the more restricted pool of experts able to create it. Experts were paid up to \$5000 for each question that was accepted to the Humanity's Last Exam benchmark ([Phan et al., 2025](#)), and while this benchmark still has significant headroom at the time of writing (June 2025), performance on it has improved significantly over the space of a few months (with the best models achieving just a few percent accuracy on it when it was initially published in early 2025). When one considers agentic systems, which are able to tackle problems for longer and which have access to tools and self critique, the complexity of benchmarks required to measure performance also increases dramatically. Being able to scale evaluations in both their capability coverage and their difficulty, while also representing tasks that have economic value, will be the key to unlocking the next generation of AI systems.

## References

- R. Anil, G. Pereyra, A. Passos, R. Ormandi, G. E. Dahl, and G. E. Hinton. Large scale distributed neural network training through online distillation, 2018. URL <https://arxiv.org/abs/1804.03235>.
- R. Anil, A. M. Dai, O. Firat, M. Johnson, D. Lepikhin, et al. PaLM 2 technical report, 2023. URL <https://arxiv.org/abs/2305.10403>.
- Anthropic. Claude’s extended thinking, 2025. URL <https://www.anthropic.com/research/visible-extended-thinking>.
- A. Baddepudi, A. Yang, and M. Lučić. Advancing the frontier of video understanding with Gemini 2.5, 2025. URL <https://developers.googleblog.com/en/gemini-2-5-video-understanding/>.
- Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, et al. Constitutional ai: Harmlessness from ai feedback, 2022. URL <https://arxiv.org/abs/2212.08073>.
- M. Balunović, J. Dekoninck, I. Petrov, N. Jovanović, and M. Vechev. Matharena: Evaluating llms on uncontaminated math competitions, 2025. URL <https://arxiv.org/abs/2505.23281>.
- P. Barham, A. Chowdhery, J. Dean, S. Ghemawat, S. Hand, D. Hurt, M. Isard, H. Lim, R. Pang, S. Roy, et al. Pathways: Asynchronous distributed dataflow for ml. *Proceedings of Machine Learning and Systems*, 4:430–449, 2022. URL <https://proceedings.mlr.press/v162/barham22a.html>.
- A. Beutel, K. Xiao, J. Heidecke, and L. Weng. Diverse and effective red teaming with auto-generated rewards and multi-step reinforcement learning, 2024. URL <https://arxiv.org/abs/2412.18693>.
- S. Biderman, H. Schoelkopf, Q. G. Anthony, H. Bradley, K. O’Brien, et al. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, 2023. URL <https://proceedings.mlr.press/v202/biderman23a.html>.
- N. Carlini, D. Ippolito, M. Jagielski, K. Lee, F. Tramer, and C. Zhang. Quantifying memorization across neural language models. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 1113–1130, 2022. URL <https://arxiv.org/abs/2202.07646>.
- W.-L. Chiang, L. Zheng, Y. Sheng, A. N. Angelopoulos, T. Li, D. Li, B. Zhu, H. Zhang, M. Jordan, J. E. Gonzalez, et al. Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first International Conference on Machine Learning*, 2024. URL <https://arxiv.org/abs/2306.05685>.
- A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, et al. PaLM: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022. URL <https://arxiv.org/abs/2204.02311>.
- N. Chowdhury, J. Aung, C. J. Shern, O. Jaffe, D. Sherburn, G. Starace, E. Mays, R. Dias, M. Aljubeh, M. Glaese, C. E. Jimenez, J. Yang, L. Ho, T. Patwardhan, K. Liu, and A. Madry. Introducing SWE-bench verified, 2024. URL <https://openai.com/index/introducing-swe-bench-verified/>.

- A. Clark, D. de las Casas, A. Guy, A. Mensch, M. Paganini, J. Hoffmann, B. Damoc, B. Hechtman, T. Cai, S. Borgeaud, G. van den Driessche, E. Rutherford, T. Hennigan, M. Johnson, K. Millican, A. Cassirer, C. Jones, E. Buchatskaya, D. Budden, L. Sifre, S. Osindero, O. Vinyals, J. Rae, E. Elsen, K. Kavukcuoglu, and K. Simonyan. Unified scaling laws for routed language models, 2022. URL "<https://arxiv.org/abs/2202.01169>".
- CodeGemma Team, H. Zhao, J. Hui, J. Howland, N. Nguyen, S. Zuo, A. Hu, C. A. Choquette-Choo, J. Shen, J. Kelley, K. Bansal, L. Vilnis, M. Wirth, P. Michel, P. Choy, P. Joshi, R. Kumar, S. Hashmi, S. Agrawal, Z. Gong, J. Fine, T. Warkentin, A. J. Hartman, B. Ni, K. Korevec, K. Schaefer, and S. Huffman. CodeGemma: Open Code Models Based on Gemma, 2024. URL <https://arxiv.org/abs/2406.11409>.
- A. Conneau, M. Ma, S. Khanuja, Y. Zhang, V. Axelrod, S. Dalmia, J. Riesa, C. Rivera, and A. Bapna. Fleurs: Few-shot learning evaluation of universal representations of speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 798–805. IEEE, 2023.
- M. Dehghani, J. Djolonga, B. Mustafa, P. Padlewski, J. Heek, J. Gilmer, A. P. Steiner, M. Caron, R. Geirhos, I. Alabdulmohsin, et al. Scaling vision transformers to 22 billion parameters. In *International Conference on Machine Learning*, pages 7480–7512. PMLR, 2023. URL <https://proceedings.mlr.press/v202/dehghani23a/dehghani23a.pdf>.
- T. Doshi. Build rich, interactive web apps with an updated Gemini 2.5 Pro, 2025a. URL <https://blog.google/products/gemini/gemini-2-5-pro-updates/>.
- T. Doshi. Gemini 2.5: Our most intelligent models are getting even better, 2025b. URL <https://blog.google/technology/google-deepmind/google-gemini-updates-io-2025/>.
- N. Du, Y. Huang, A. M. Dai, S. Tong, D. Lepikhin, Y. Xu, M. Krikun, Y. Zhou, A. W. Yu, O. Firat, et al. GLaM: Efficient scaling of language models with mixture-of-experts. *arXiv preprint arXiv:2112.06905*, 2021. URL <https://arxiv.org/abs/2112.06905>.
- W. Fedus, B. Zoph, and N. Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *arXiv preprint arXiv:2101.03961*, 2021. URL <https://arxiv.org/abs/2101.03961>.
- C. Fu, Y. Dai, Y. Luo, L. Li, S. Ren, R. Zhang, Z. Wang, C. Zhou, Y. Shen, M. Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24108–24118, 2025. URL [https://openaccess.thecvf.com/content/CVPR2024/html/Fu\\_Video-MME\\_The\\_First-Ever\\_Comprehensive\\_Evaluation\\_Benchmark\\_of\\_Multi-Modal\\_LLMS\\_in\\_CVPR\\_2024\\_paper.html](https://openaccess.thecvf.com/content/CVPR2024/html/Fu_Video-MME_The_First-Ever_Comprehensive_Evaluation_Benchmark_of_Multi-Modal_LLMS_in_CVPR_2024_paper.html).
- P. Gauthier. Aider Polyglot Coding Leaderboard, 2025. URL <https://aider.chat/docs/leaderboards/>.
- Gemini Team. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. URL <https://arxiv.org/abs/2403.05530>.
- Gemini Team, Google. Gemini Deep Research, 2024. URL <https://gemini.google/overview/deep-research/>.
- Gemma Team. Gemma: Open Models Based on Gemini Research and Technology, 2024. URL <https://arxiv.org/abs/2403.08295>.

- O. Goldman, U. Shaham, D. Malkin, S. Eiger, A. Hassidim, Y. Matias, J. Maynez, A. M. Gilady, J. Riesa, S. Rijhwani, L. Rimell, I. Szpektor, R. Tsarfaty, and M. Eyal. Eclektic: a novel challenge set for evaluation of cross-lingual knowledge transfer, 2025. URL <https://arxiv.org/abs/2502.21228>.
- Google DeepMind. Frontier safety framework, February 2025a. URL <https://deepmind.google/discover/governance/frontier-safety-framework/>.
- Google DeepMind. Gemini 2.0 Flash-Lite, 2025b. URL <https://deepmind.google/models/gemini/flash-lite/>.
- A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, et al. The Llama 3 Herd of Models, 2024. URL <https://arxiv.org/abs/2407.21783>.
- D. Hassabis. Our vision for building a universal AI assistant, 2025. URL <https://blog.google/technology/google-deepmind/gemini-universal-ai-assistant/>.
- G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network, 2015. URL <https://arxiv.org/abs/1503.02531>.
- K. Hu, P. Wu, F. Pu, W. Xiao, Y. Zhang, X. Yue, B. Li, and Z. Liu. Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos, 2025. URL <https://arxiv.org/abs/2501.13826>.
- S. Hughes, M. Bae, and M. Li. Vectara Hallucination Leaderboard, nov 2023. URL <https://github.com/vectara/hallucination-leaderboard>.
- D. Ippolito, F. Tramer, M. Nasr, C. Zhang, M. Jagielski, K. Lee, C. A. Choquette-Choo, and N. Carlini. Preventing verbatim memorization in language models gives a false sense of privacy, 2022. URL <https://arxiv.org/abs/2210.17546>.
- A. Jacovi, A. Wang, C. Alberti, C. Tao, J. Lipovetz, K. Olszewska, L. Haas, M. Liu, N. Keating, A. Bloniarz, C. Saroufim, C. Fry, D. Marcus, D. Kukliansky, G. S. Tomar, J. Swirhun, J. Xing, L. Wang, M. Gurumurthy, M. Aaron, M. Ambar, R. Fellinger, R. Wang, R. Sims, Z. Zhang, S. Goldshtein, and D. Das. Facts grounding leaderboard. <https://www.kaggle.com/benchmarks/google/facts-grounding>, 2024. Google Deepmind, Google Research, Google Cloud, Kaggle.
- A. Jacovi, A. Wang, C. Alberti, C. Tao, J. Lipovetz, K. Olszewska, L. Haas, M. Liu, N. Keating, A. Bloniarz, et al. The facts grounding leaderboard: Benchmarking llms' ability to ground responses to long-form input. *arXiv preprint arXiv:2501.03200*, 2025. URL <https://arxiv.org/abs/2501.03200>.
- N. Jain, K. Han, A. Gu, W.-D. Li, F. Yan, T. Zhang, S. Wang, A. Solar-Lezama, K. Sen, and I. Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code, 2024. URL <https://arxiv.org/abs/2403.07974>.
- A. Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D. S. Chaplot, D. d. l. Casas, E. B. Hanna, F. Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024. URL <https://arxiv.org/abs/2401.04088>.
- C. E. Jimenez, J. Yang, A. Wettig, S. Yao, K. Pei, O. Press, and K. R. Narasimhan. SWE-bench: Can language models resolve real-world github issues? In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=VTF8yNQM66>.

- K. Kampf and N. Brichtova. Experiment with Gemini 2.0 Flash native image generation, 2025. URL <https://developers.googleblog.com/en/experiment-with-gemini-2-0-flash-native-image-generation/>.
- K. Kavukcuoglu. Gemini 2.0 is now available to everyone, 2025. URL <https://blog.google/technology/google-deepmind/gemini-model-updates-february-2025>.
- L. Kilpatrick. Gemini 2.5 Pro Preview: even better coding performance, 2025. URL <https://developers.googleblog.com/en/gemini-2-5-pro-io-improved-coding-performance>.
- S. Kudugunta, I. Caswell, B. Zhang, X. Garcia, C. A. Choquette-Choo, K. Lee, D. Xin, A. Kusupati, R. Stella, A. Bapna, and O. Firat. MADLAD-400: A Multilingual And Document-Level Large Audited Dataset, 2023. URL <https://arxiv.org/abs/2309.04662>.
- J. M. Laurent, J. D. Janizek, M. Ruzo, M. M. Hinks, M. J. Hammerling, S. Narayanan, et al. LAB-Bench: Measuring capabilities of language models for biology research, 2024. URL <https://arxiv.org/abs/2407.10362>.
- LearnLM Team. Evaluating Gemini in an Arena for Learning, 2025. URL <https://google/LearnLM-May25>.
- J. Lee, A. Chen, Z. Dai, D. Dua, D. S. Sachan, M. Boratko, Y. Luan, S. M. Arnold, V. Perot, S. Dalmia, et al. Can long-context language models subsume retrieval, rag, sql, and more? *arXiv preprint arXiv:2406.13121*, 2024. URL <https://arxiv.org/abs/2406.13121>.
- J. Lei, T. L. Berg, and M. Bansal. Detecting moments and highlights in videos via natural language queries. *Advances in Neural Information Processing Systems*, 34:11846–11858, 2021.
- D. Lepikhin, H. Lee, Y. Xu, D. Chen, O. Firat, Y. Huang, M. Krikun, N. Shazeer, and Z. Chen. GShard: Scaling giant models with conditional computation and automatic sharding. In *International Conference on Learning Representations*, 2020. URL <https://openreview.net/forum?id=qrwE7XHTmYb>.
- N. Li, A. Pan, A. Gopal, S. Yue, D. Berrios, A. Gatti, et al. The WMDP benchmark: Measuring and reducing malicious use with unlearning, 2024. URL <https://arxiv.org/abs/2403.03218>.
- L. Liu, X. Liu, J. Gao, W. Chen, and J. Han. Understanding the difficulty of training transformers. *arXiv preprint arXiv:2004.08249*, 2020. URL <https://arxiv.org/abs/2004.08249>.
- LMArena Team. Webdev arena, 2025. URL <https://web.lmarena.ai/leaderboard>.
- S. B. Mallick and L. Kilpatrick. Gemini 2.0: Flash, Flash-Lite and Pro, 2025. URL <https://developers.googleblog.com/en/gemini-2-family-expands>.
- A. Mehrotra, M. Zampetakis, P. Kassianik, B. Nelson, H. Anderson, Y. Singer, and A. Karbasi. Tree of attacks: Jailbreaking black-box llms automatically, 2024. URL <https://arxiv.org/abs/2312.02119>.
- I. Molybog, P. Albert, M. Chen, Z. DeVito, D. Esiobu, N. Goyal, P. Koura, S. Narang, A. Poulton, R. Silva, et al. A theory on adam instability in large-scale machine learning. *arXiv preprint arXiv:2304.09871*, 2023. URL <https://arxiv.org/abs/2304.09871>.
- A. Nagrani, S. Menon, A. Iscen, S. Buch, R. Mehran, N. Jha, A. Hauth, Y. Zhu, C. Vondrick, M. Sirotenko, C. Schmid, and T. Weyand. Minerva: Evaluating complex video reasoning, 2025a. URL <https://arxiv.org/abs/2505.00681>.

- A. Nagrani, M. Zhang, R. Mehran, R. Hornung, N. B. Gundavarapu, N. Jha, A. Myers, X. Zhou, B. Gong, C. Schmid, M. Sirotenko, Y. Zhu, and T. Weyand. Neptune: The long orbit to benchmarking long video understanding, 2025b. URL <https://arxiv.org/abs/2412.09582>.
- M. Nasr, N. Carlini, J. Hayase, M. Jagielski, A. F. Cooper, D. Ippolito, C. A. Choquette-Choo, E. Wallace, F. Tramèr, and K. Lee. Scalable extraction of training data from (production) language models, 2023. URL <https://arxiv.org/abs/2311.17035>.
- P. Padlewski, M. Bain, M. Henderson, Z. Zhu, N. Relan, H. Pham, D. Ong, K. Aleksiev, A. Ormazabal, S. Phua, E. Yeo, E. Lamprecht, Q. Liu, Y. Wang, E. Chen, D. Fu, L. Li, C. Zheng, C. de Masson d'Autume, D. Yogatama, M. Artetxe, and Y. Tay. Vibe-eval: A hard evaluation suite for measuring progress of multimodal language models, 2024. URL <https://arxiv.org/abs/2405.02287>.
- A. Pappu, B. Porter, I. Shumailov, and J. Hayes. Measuring memorization in RLHF for code completion. *arXiv preprint arXiv:2406.11715*, 2024. URL <https://arxiv.org/abs/2406.11715>.
- V. Patraucean, L. Smaira, A. Gupta, A. Recasens, L. Markeeva, D. Banarse, S. Koppula, M. Malinowski, Y. Yang, C. Doersch, et al. Perception test: A diagnostic benchmark for multimodal video models. *Advances in Neural Information Processing Systems*, 36:42748–42761, 2023.
- E. Perez, S. Huang, F. Song, T. Cai, R. Ring, J. Aslanides, A. Glaese, N. McAleese, and G. Irving. Red teaming language models with language models, 2022. URL <https://arxiv.org/abs/2202.03286>.
- L. Phan et al. Humanity's last exam, 2025. URL <https://arxiv.org/abs/2501.14249>.
- M. Phuong, M. Aitchison, E. Catt, S. Cogan, A. Kaskasoli, V. Krakovna, D. Lindner, M. Rahtz, Y. Assael, S. Hodkinson, et al. Evaluating frontier models for dangerous capabilities, 2024. URL <https://arxiv.org/abs/2403.13793>.
- M. Phuong, R. S. Zimmermann, Z. Wang, D. Lindner, V. Krakovna, S. Cogan, A. Dafoe, L. Ho, and R. Shah. Evaluating frontier models for stealth and situational awareness, 2025. URL <https://arxiv.org/abs/2505.01420>.
- S. Pichai. Google I/O 2025: From research to reality, 2025. URL <https://blog.google/technology/ai/io-2025-keynote/>.
- C. Plizzari, A. Tonioni, Y. Xian, A. Kulshrestha, and F. Tombari. Omnia de egotempo: Benchmarking temporal understanding of multi-modal llms in egocentric videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24129–24138, 2025.
- D. Rein, B. L. Hou, A. C. Stickland, J. Petty, R. Y. Pang, J. Dirani, J. Michael, and S. R. Bowman. Gqqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*, 2024.
- C. Riquelme, J. Puigcerver, B. Mustafa, M. Neumann, R. Jenatton, A. S. Pinto, D. Keysers, and N. Houlsby. Scaling vision with sparse mixture of experts, 2021. URL <https://arxiv.org/abs/2106.05974>.
- J. Roberts, M. R. Taesiri, A. Sharma, A. Gupta, S. Roberts, I. Croitoru, S.-V. Bogolin, J. Tang, F. Langer, V. Raina, et al. ZeroBench: An impossible visual benchmark for contemporary large multimodal models. *arXiv preprint arXiv:2502.09696*, 2025.
- M. Rodriguez, R. A. Popa, L. Liang, A. Wang, M. Rahtz, A. Kaskasoli, A. Dafoe, and F. Flynn. A framework for evaluating emerging cyberattack capabilities of AI, 2025. URL <https://arxiv.org/abs/2503.11917>.

- S. Roller, S. Sukhbaatar, J. Weston, et al. Hash layers for large sparse models. *Advances in Neural Information Processing Systems*, 34:17555–17566, 2021. URL <https://proceedings.neurips.cc/paper/2021/file/883e881bc596359e0c5112411858a74b-Paper.pdf>.
- M. Samvelyan, S. C. Raparthy, A. Lupu, E. Hambro, A. H. Markosyan, M. Bhatt, Y. Mao, M. Jiang, J. Parker-Holder, J. Foerster, T. Rocktäschel, and R. Raileanu. Rainbow teaming: Open-ended generation of diverse adversarial prompts, 2024. URL <https://arxiv.org/abs/2402.16822>.
- R. Shah, A. Irpan, A. M. Turner, A. Wang, A. Conmy, D. Lindner, J. Brown-Cohen, L. Ho, N. Nanda, R. A. Popa, R. Jain, R. Greig, S. Albanie, S. Emmons, S. Farquhar, S. Krier, S. Rajamanoharan, S. Bridgers, T. Ijitoye, T. Everitt, V. Krakovna, V. Varma, V. Mikulik, Z. Kenton, D. Orr, S. Legg, N. Goodman, A. Dafoe, F. Flynn, and A. Dragan. An approach to technical agi safety and security, 2025. URL <https://arxiv.org/abs/2504.01849>.
- D. Sharon. Upload and edit your images directly in the Gemini app, 2025. URL <https://blog.google/products/gemini/image-editing/>.
- N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, and J. Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *ICLR (Poster)*. OpenReview.net, 2017. URL <https://arxiv.org/abs/1701.06538>.
- C. Shi, S. Lin, S. Song, J. Hayes, I. Shumailov, I. Yona, J. Pluto, A. Pappu, C. A. Choquette-Choo, M. Nasr, C. Sitawarin, G. Gibson, A. Terzis, and J. F. Flynn. Lessons from defending gemini against indirect prompt injections, 2025. URL <https://arxiv.org/abs/2505.14534>.
- S. Singh, A. Romanou, C. Fourrier, D. I. Adelani, J. G. Ngui, D. Vila-Suero, P. Limkonchotiwat, K. Marchisio, W. Q. Leong, Y. Susanto, R. Ng, S. Longpre, W.-Y. Ko, M. Smith, A. Bosselut, A. Oh, A. F. T. Martins, L. Choshen, D. Ippolito, E. Ferrante, M. Fadaee, B. Ermis, and S. Hooker. Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation, 2024. URL <https://arxiv.org/abs/2412.03304>.
- R. Stein. Expanding AI Overviews and introducing AI Mode, 2025. URL <https://blog.google/products/search/ai-mode-search>.
- I. Tolstikhin, N. Houlsby, A. Kolesnikov, L. Beyer, X. Zhai, T. Unterthiner, J. Yung, A. Steiner, D. Keysers, J. Uszkoreit, M. Lucic, and A. Dosovitskiy. Mlp-mixer: An all-mlp architecture for vision, 2021.
- A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL [https://proceedings.neurips.cc/paper\\_files/paper/2017/file/3f5ee243547dee91fb0d053c1c4a845aa-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fb0d053c1c4a845aa-Paper.pdf).
- K. Vodrahalli, S. Ontanon, N. Tripuraneni, K. Xu, S. Jain, R. Shivanna, J. Hui, N. Dikkala, M. Kazemi, B. Fatemi, et al. Michelangelo: Long context evaluations beyond haystacks via latent structure queries. *arXiv preprint arXiv:2409.12640*, 2024. URL <https://arxiv.org/abs/2409.12640>.
- B. Wang. NotebookLM now lets you listen to a conversation about your sources , 2024. URL <https://blog.google/technology/ai/notebooklm-audio-overviews>.
- C. Wang, A. Wu, and J. Pino. Covost 2: A massively multilingual speech-to-text translation corpus, 2020.

- W. Wang, Z. He, W. Hong, Y. Cheng, X. Zhang, J. Qi, X. Gu, S. Huang, B. Xu, Y. Dong, M. Ding, and J. Tang. Lvbench: An extreme long video understanding benchmark, 2024. URL <https://arxiv.org/abs/2406.08035>.
- X. Wang, J. Wu, J. Chen, L. Li, Y.-F. Wang, and W. Y. Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4581–4591, 2019.
- J. Wei, K. Nguyen, H. W. Chung, Y. J. Jiao, S. Papay, A. Glaese, J. Schulman, and W. Fedus. Measuring short-form factuality in large language models. *arXiv preprint arXiv:2411.04368*, 2024. URL <https://arxiv.org/abs/2411.04368>.
- L. Weidinger, J. Barnhart, J. Brennan, C. Butterfield, S. Young, W. Hawkins, et al. Holistic safety and responsibility evaluations of advanced ai models, 2024. URL <https://arxiv.org/abs/2404.14068>.
- H. Wijk, T. Lin, J. Becker, S. Jawhar, N. Parikh, T. Broadley, L. Chan, M. Chen, J. Clymer, J. Dhyani, et al. RE-Bench: Evaluating frontier ai r&d capabilities of language model agents against human experts, 2025. URL <https://arxiv.org/abs/2411.15114>.
- M. Wortsman, P. J. Liu, L. Xiao, K. Everett, A. Alemi, B. Adlam, J. D. Co-Reyes, I. Gur, A. Kumar, R. Novak, et al. Small-scale proxies for large-scale transformer training instabilities. *arXiv preprint arXiv:2309.14322*, 2023. URL <https://arxiv.org/abs/2309.14322>.
- J. Yang, A. Prabhakar, K. Narasimhan, and S. Yao. Intercode: Standardizing and benchmarking interactive coding with execution feedback, 2023. URL <https://arxiv.org/abs/2306.14898>.
- Z. Yu, D. Xu, J. Yu, T. Yu, Z. Zhao, Y. Zhuang, and D. Tao. ActivityNet-QA: A dataset for understanding complex web videos via question answering. In *AAAI*, 2019.
- X. Yue, Y. Ni, K. Zhang, T. Zheng, R. Liu, G. Zhang, S. Stevens, D. Jiang, W. Ren, Y. Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024.
- Zerokid. Pokemon Red Version - Guide and Walkthrough (GB), 2024. URL <https://www.gamefaqs.gamespot.com/gameboy/367023-Pokémon-red-version/faqs/64175>.
- S. Zhai, T. Likhomanenko, E. Littwin, D. Busbridge, J. Ramapuram, Y. Zhang, J. Gu, and J. M. Susskind. Stabilizing transformer training by preventing attention entropy collapse. In *International Conference on Machine Learning*, pages 40770–40803. PMLR, 2023. URL <https://proceedings.mlr.press/v202/zhai23a/zhai23a.pdf>.
- J. Zhang. Gemini Plays Pokemon Twitch Stream, 2025. URL [https://www.twitch.tv/gemini\\_plays\\_pokemon/about](https://www.twitch.tv/gemini_plays_pokemon/about).
- S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022. URL <https://arxiv.org/abs/2205.01068>.
- L. Zhou, C. Xu, and J. J. Corso. Towards automatic learning of procedures from web instructional videos. In *AAAI Conference on Artificial Intelligence*, pages 7590–7598, 2018. URL <https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17344>.

## 7. Contributors and Acknowledgments

Contributors	Filip Pavetić Duc Dung Nguyen Eric Bieber Mike Schaeckermann Ice Pasupat Noveen Sachdeva Inderjit Dhillon Marcel Blistein Ori Ram Dan Zhang Evan Rosen Luke Marris Sam Petulla Colin Gaffney Asaf Aharoni Nathan Lintz Tiago Cardal Pais Henrik Jacobsson Idan Szpektor Nan-Jiang Jiang Krishna Haridasan Ahmed Omran Nikunj Saunshi Dara Bahri Gaurav Mishra Eric Chu Toby Boyd Brad Hekman Aaron Parisi Chaoyi Zhang Kornraphop Kawintiranon Tania Bedrax-Weiss Oliver Wang Ya Xu Ollie Purkiss Uri Mendlovic Ilai Deutel Nam Nguyen Adam Langley Flip Korn Lucia Rossazza Alexandre Ramé Sagar Waghmare Helen Miller Vaishakh Keshava Ying Jian Xiaofan Zhang Raluca Ada Popa Kedar Dhamdhere Blaž Bratanič Kyuyeun Kim Terry Koo Ferran Alet Yi-ting Chen Arsha Nagrani Hannah Muckenhirk Zhiyuan Zhang Corbin Quick	Carey Radebaugh Stephen Roller Abhishek Nayyar Jacob Austin Tayfun Terzi Bhargav Kanagal Shamanna Pete Shaw Aayush Singh Florian Luisier Eric Ni Sahra Ghalebikesabi Edouard Yvinec Alaa Saade Thatcher Ulrich Lorenzo Blanco Dan A. Calian Muhuan Huang Aäron van den Oord Naman Goyal Terry Chen Praynaa Rawlani Christian Schallhart Swachhand Lokhande Xianghong Luo Jyn Shan Ceslee Montgomery Victoria Krakovna Federico Piccinini Omer Barak Jingyu Cui Yiling Jia Mikhail Dektiarev Alexey Kolganov Shiyu Huang Zhe Chen Xingyu Wang Jessica Austin Peter de Boursac Evgeny Sluzhaev Frank Ding Huijian Li Surya Bhupatiraju Mohit Agarwal Sławek Kwasiborski Paramjit Sandhu Patrick Siegler Ahmet Iscen Eyal Ben-David Shiraz Butt Miltos Allamanis Seth Benjamin Robert Busa-Fekete Felix Hernandez-Campos Sasha Goldshtein Matt Dibb Weiyan Zhang Annie Marsden	Anna Shekhawat Alexey Stern Rishikesh Ingale Stefani Karp Sanaz Bahargam Adrian Goedeckemeyer Jie Han Sicheng Li Andrea Tacchetti Dian Yu Abhishek Chakladar Zhiying Zhang Mona El Mahdy Xu Gao Dale Johnson Samrat Phatale AJ Piergiovanni Hyeontaek Lim Clement Farabet Carl Lebsack Theo Guidroz John Blitzer Nico Duduta David Madras Steve Li Daniel von Dincklage Xin Li Mahdis Mahdieh George Tucker Ganesh Jawahar Owen Xiao Danny Tarlow Robert Geirhos Noam Velan Daniel Vlasic Kalesha Bullard SK Park Nishesh Gupta Kellie Webster Ayal Hitron Jieming Mao Julian Eisenschlos Laurel Prince Nina D'Souza Kelvin Zheng Sara Nasso Gabriela Botea Carl Doersch Caglar Unlu Chris Alberti Alexey Svyatkovskiy Ankita Goel Krzysztof Choromanski Pan-Pan Jiang Richard Nguyen Four Flynn Daria Ćurko Peter Chen

Nicholas Roth	Sanjiv Kumar	Annie Louis	Jaeyoun Kim
Kieran Milan	Florian Hartmann	KP Sawhney	Markus Kunesch
Caleb Habtegebriel	Michael Kilgore	Slav Petrov	Steven Pecht
Shashi Narayan	Jinhyuk Lee	Jing Xie	Nami Akazawa
Michael Moffitt	Aroma Mahendru	Yunfei Bai	Abe Friesen
Jake Marcus	Roman Ring	Georgi Stoyanov	James Lyon
Thomas Anthony	Tom Hennigan	Alex Fabrikant	Ali Eslami
Brendan McMahan	Fiona Lang	Rajesh Jayaram	Junru Wu
Gowoon Cheon	Colin Cherry	Yuqi Li	Jie Tan
Ruibo Liu	David Steiner	Joe Heyward	Yue Song
Megan Barnes	Dawsen Hwang	Justin Gilmer	Ravi Kumar
Lukasz Lew	Ray Smith	Yaqing Wang	Chris Welty
Rebeca	Pidong Wang	Radu Soricu	Ilia Akolzin
Santamaria-Fernandez	Jeremy Chen	Luyang Liu	Gena Gibson
Mayank Upadhyay	Ming-Hsuan Yang	Qingnan Duan	Sean Augenstein
Arjun Akula	Sam Kwei	Jamie Hayes	Arjun Pillai
Arnar Mar Hrafnkelsson	Philippe Schlattner	Maura O'Brien	Nancy Yuen
Alvaro Caceres	Donnie Kim	Gaurav Singh Tomar	Du Phan
Andrew Bunner	Ganesh Poomal Girirajan	Sivan Eiger	Xin Wang
Michal Sokolik	Nikola Momchev	Bahar Fatemi	Iain Barr
Subha Puttagunta	Ayushi Agarwal	Jeffrey Hui	Heiga Zen
Lawrence Moore	Xingyi Zhou	Catarina Barros	Nan Hua
Berivan Isik	Ilkin Safarli	Adaeze Chukwuka	Casper Liu
Jay Hartford	Zachary Garrett	Alena Butryna	Jilei (Jerry) Wang
Lawrence Chan	AJ Pierigiovanni	Saksham Thakur	Tanuj Bhatia
Pradeep Shenoy	Sarthak Jauhari	Austin Huang	Hao Xu
Dan Holtmann-Rice	Alif Raditya Rochman	Zhufeng Pan	Oded Elyada
Jane Park	Shikhar Vashishth	Haotian Tang	Pushmeet Kohli
Fabio Viola	Quan Yuan	Serkan Cabi	Mirek Olšák
Alex Salcianu	Christof Angermueller	Tulsee Doshi	Ke Chen
Sujeevan Rajayogam	Jon Blanton	Michiel Bakker	Azalia Mirhoseini
Ian Stewart-Binks	Xinying Song	Sumit Bagri	Noam Shazeer
Zelin Wu	Nitesh Bharadwaj	Ruy Ley-Wild	Shoshana Jakobovits
Richard Everett	Gundavarapu	Adam Lelkes	Maggie Tran
Xi Xiong	Thi Avrahami	Jennie Lees	Nolan Ramsden
Pierre-Antoine Manzagol	Maxine Deines	Patrick Kane	Tarun Bharti
Gary Leung	Subhrajit Roy	David Greene	Fred Alcober
Carl Saroufim	Manish Gupta	Shimu Wu	Yunjie Li
Bo Pang	Christopher Semturs	Jörg Bornschein	Shilpa Shetty
Dawid Wegner	Shobha Vasudevan	Gabriela Surita	Jing Chen
George Papamakarios	Aditya Srikanth	Sarah Hodkinson	Dmitry Kalashnikov
Jennimaria Palomaki	Veerubhotla	Fangtao Li	Megha Nawhal
Helena Pankov	Shriya Sharma	Chris Hidey	Sercan Arik
Guangda Lai	Josh Jacob	Sébastien Pereira	Hanwen Chen
Guilherme Tubone	Zhen Yang	Sean Ammirati	Michiel Blokzijl
Shubin Zhao	Andreas Terzis	Phillip Lippe	Shubham Gupta
Theofilos Strinopoulos	Dan Karliner	Adam Kraft	James Rubin
Seth Neel	Auriel Wright	Pu Han	Rigel Swavely
Mingqiu Wang	Tania Rojas-Esponda	Sebastian Gerlach	Sophie Bridgers
Joe Kelley	Ashley Brown	Zifeng Wang	Ian Gemp
Li Li	Abhijit Guha Roy	Liviu Panait	Chen Su
Pingmei Xu	Pawan Dogra	Feng Han	Arun Suggala
Anitha Vijayakumar	Andrei Kapishnikov	Brian Farris	Juliette Pluto
Andrea D'olimpio	Peter Young	Yingying Bi	Mary Cassin
Omer Levy	Wendy Kan	Hannah DeBalsi	Alain Vaucher
Massimo Nicosia	Vinodh Kumar Rajendran	Miaosen Wang	Kaiyang Ji
Grigory Rozhdestvenskiy	Maria Ivanova	Gladys Tyen	Jiahao Cai
Ni Lao	Salil Deshmukh	James Cohan	Andrew Audibert
Sirui Xie	Chia-Hua Ho	Susan Zhang	Animesh Sinha
Yash Katariya	Mike Kwong	Jarred Barber	David Tian
Jon Simon	Stav Ginzburg	Da-Woon Chung	Efrat Farkash

Amy Hua	Saloni Shah	Isabel Leal	Luming Tang
Jilin Chen	Norbert Kalb	James Manyika	Mark Geller
Duc-Hieu Tran	Carrie Zhang	Sofia Erell	Simon Bucher
Edward Loper	Shruthi Prabhakara	Daniel Murphy	Yifan Ding
Nicole Brichtova	Amit Sabne	Zhisheng Xiao	Hongzhi Shi
Lara McConnaughey	Artiom Myaskovsky	Anton Bulyenov	Carrie Muir
Ballie Sandhu	Vikas Raunak	Julian Walker	Dominik Grewe
Robert Leland	Blanca Huergo	Mark Collier	Ramy Eskander
Doug DeCarlo	Behnam Neyshabur	Matej Kastelic	Octavio Ponce
Andrew Over	Jon Clark	Nelson George	Boqing Gong
James Huang	Ye Zhang	Sushant Prakash	Derek Gasaway
Xing Wu	Shankar Krishnan	Sailesh Sidhwani	Samira Khan
Connie Fan	Eden Cohen	Alexey Frolov	Umang Gupta
Eric Li	Dinesh Tewari	Steven Hansen	Angelos Filos
Yun Lei	James Lottes	Petko Georgiev	Weicheng Kuo
Deepak Sharma	Yumeya Yamamori	Tiberiu Sosea	Klemen Kloboves
Cosmin Paduraru	Hui (Elena) Li	Chris Apps	Jennifer Beattie
Luo Yu	Mohamed Elhwaty	Aishwarya Kamath	Christian Wright
Matko Bošnjak	Ada Maksutaj Oflazer	David Reid	Leon Li
Phuong Dao	Adrià Recasens	Emma Cooney	Alicia Jin
Min Choi	Sheryl Luo	Charlotte Magister	Sandeep Mariserla
Sneha Kudugunta	Duy Nguyen	Oriana Riva	Miteyan Patel
Jakub Adamek	Taylor Bos	Alec Go	Jens Heitkaemper
Carlos Guía	Kalyan Andra	Pu-Chin Chen	Dilip Krishnan
Ali Khodaei	Ana Salazar	Sebastian Krause	Vivek Sharma
Jie Feng	Ed Chi	Nir Levine	David Bieber
Wenjun Zeng	Jeongwoo Ko	Marco Fornoni	Christian Frank
David Welling	Matt Ginsberg	Ilya Figotin	John Lambert
Sandeep Tata	Anders Andreassen	Nick Roy	Paul Caron
Christina Butterfield	Anian Ruoss	Parsa Mahmoudieh	Martin Polacek
Andrey Vlasov	Todor Davchev	Vladimir Magay	Mai Giménez
Seliem El-Sayed	Elnaz Davoodi	Mukundan Madhavan	Himadri Choudhury
Swaroop Mishra	Chenxi Liu	Jin Miao	Xing Yu
Tara Sainath	Min Kim	Jianmo Ni	Sasan Tavakkol
Shentao Yang	Santiago Ontanon	Yasuhsisa Fujii	Arun Ahuja
RJ Skerry-Ryan	Chi Ming To	Ian Chou	Franz Och
Jeremy Shar	Dawei Jia	George Scrivener	Rodolphe Jenatton
Robert Berry	Rosemary Ke	Zak Tsai	Wojtek Skut
Arunkumar Rajendran	Jing Wang	Siobhan Mcoughlin	Bryan Richter
Arun Kandoor	Anna Korsun	Jeremy Selier	David Gaddy
Andrea Burns	Moran Ambar	Sandra Lefdal	Andy Ly
Deepali Jain	Ilya Kornakov	Jeffrey Zhao	Misha Bilenko
Tom Stone	Irene Giannoumis	Abhijit Karmarkar	Megh Umekar
Wonpyo Park	Toni Creswell	Kushal Chauhan	Ethan Liang
Shibo Wang	Denny Zhou	Shivanker Goel	Martin Sevenich
Albin Cassirer	Yi Su	Zhaoyi Zhang	Mandar Joshi
Guohui Wang	Ishaan Watts	Vihan Jain	Hassan Mansoor
Hayato Kobayashi	Aleksandr Zaks	Parisa Haghani	Rebecca Lin
Sergey Rogulenko	Evgenii Eltyshev	Mostafa Dehghani	Sumit Sanghai
Vineetha Govindaraj	Ziqiang Feng	Jacob Scott	Abhimanyu Singh
Mikołaj Rybiński	Sidharth Mudgal	Erin Farnese	Xiaowei Li
Nadav Olmert	Alex Kaskasoli	Anastasija Ilić	Sudheendra
Colin Evans	Juliette Love	Steven Baker	Vijayanarasimhan
Po-Sen Huang	Kingshuk Dasgupta	Julia Pawar	Zaheer Abbas
Kelvin Xu	Sam Shleifer	Li Zhong	Yonatan Bitton
Premal Shah	Richard Green	Josh Camp	Hansa Srinivasan
Terry Thurk	Sungyong Seo	Yoel Zeldes	Manish Reddy Vuyyuru
Caitlin Sikora	Chansoo Lee	Shravya Shetty	Alexander Frömmgen
Mu Cai	Dale Webster	Anand Iyer	Yanhua Sun
Jin Xie	Prakash Shroff	Vít Listík	Ralph Leith
Elahe Dabir	Ganna Raboshchuk	Jiaxian Guo	Alfonso Castaño

DJ Strouse	Tianxiao Shen	Bin Ni	Dipankar Ghosh
Le Yan	Blagoj Mitrevski	Alexey Vlaskin	Aahil Mehta
Austin Kyker	Michael Tschannen	Solomon Demmessie	Dana Alon
Satish Kambala	Sreenivas Gollapudi	Lucio Dery	George Polovets
Mary Jasarevic	Aishwarya P S	Salah Zaiem	Alessio Tonioni
Thibault Sellam	José Leal	Yanping Huang	Nate Kushman
Chao Jia	Zhe Shen	Cindy Fan	Joel D'sa
Alexander Pritzel	Han Fu	Felix Gimeno	Lin Zhuo
Raghavender R	Wei Wang	Ananth Balashankar	Allen Wu
Huizhong Chen	Arvind Kannan	Koji Kojima	Rohin Shah
Natalie Clay	Doron Kukliansky	Hagai Taitelbaum	John Youssef
Sudeep Gandhe	Sergey Yaroshenko	Maya Meng	Jiayu Ye
Sean Kirmani	Svetlana Grant	Dero Gharibian	Justin Snyder
Sayna Ebrahimi	Umesh Telang	Sahil Singla	Karel Lenc
Hannah Kirkwood	David Wood	Wei Chen	Senaka Buthpitiya
Jonathan Mallinson	Alexandra Chronopoulou	Ambrose Slone	Matthew Tung
Chao Wang	Alexandru Tifrea	Guanjie Chen	Jichuan Chang
Adnan Ozturk	Tao Zhou	Sujee Rajayogam	Tao Chen
Kuo Lin	Tony (Tuán) Nguyẽn	Max Schumacher	David Saxton
Shyam Upadhyay	Muge Ersoy	Suyog Kotecha	Jenny Lee
Vincent Cohen-Addad	Anima Singh	Rory Blevins	Lydia Lihui Zhang
Sean Purser-haskell	Meiyan Xie	Qifei Wang	James Qin
Yichong Xu	Emanuel Taropa	Mor Hazan Taeg	Prabakar Radhakrishnan
Ebrahim Songhori	Woohyun Han	Alex Morris	Maxwell Chen
Babi Seal	Eirikur Agustsson	Xin Liu	Piotr Ambroszczyk
Alberto Magni	Andrei Sozanschi	Fayaz Jamil	Metin Toksoz-Exley
Almog Gueta	Hui Peng	Richard Zhang	Yan Zhong
Tingting Zou	Alex Chen	Pratik Joshi	Nitzan Katz
Guru Guruganesh	Yoel Drori	Ben Ingram	Brendan O'Donoghue
Thais Kagohara	Efren Robles	Tyler Liechty	Tamara von Glehn
Hung Nguyen	Yang Gao	Ahmed Eleryan	Adi Gerzi Rosenthal
Khalid Salama	Xerxes Dotiwalla	Scott Baird	Aga Świetlik
Alejandro Cruzado Ruiz	Ying Chen	Alex Grills	Xiaokai Zhao
Justin Frye	Anudhyan Boral	Gagan Bansal	Nick Fernando
Zhenkai Zhu	Alexei Bendebury	Shan Han	Jinliang Wei
Matthias Lochbrunner	John Nham	Kiran Yarasangi	Jieru Mei
Simon Osindero	Chris Tar	Shawn Xu	Sergei Vassilvitskii
Wentao Yuan	Luis Castro	Majd Al Merey	Diego Cedillo
Lisa Lee	Jiepu Jiang	Isabel Gao	Pranjal Awasthi
Aman Prasad	Canoeo Liu	Felix Weissenberger	Hui Zheng
Lam Nguyen Thiet	Felix Halim	Igor Karpov	Koray Kavukcuoglu
Daniele Calandriello	Jinoo Baek	Robert Riachi	Itay Laish
Victor Stone	Andy Wan	Ankit Anand	Joseph Pagadora
Qixuan Feng	Jeremiah Liu	Gautam Prasad	Marc Brockschmidt
Han Ke	Yuan Cao	Kay Lamerigts	Christopher A.
Maria Voitovich	Shengyang Dai	Reid Hayes	Choquette-Choo
Geta Sampemane	Trilok Acharya	Jamie Rogers	Arunkumar Byravan
Lewis Chiang	Ruoxi Sun	Mandy Guo	Yifeng Lu
Ling Wu	Fuzhao Xue	Ashish Shenoy	Xu Chen
Alexander Bykovsky	Saket Joshi	Qiong (Q) Hu	Mia Chen
Matt Young	Morgane Lustman	Kyle He	Kenton Lee
Luke Vilnis	Yongqin Xian	Yuchen Liu	Rama Pasumarthi
Ishita Dasgupta	Rishabh Joshi	Polina Zablotskaia	Sijal Bhatnagar
Aditya Chawla	Deep Karkhanis	Sagar Gabbi	Aditya Shah
Qin Cao	Nora Kassner	Yifan Chang	Qiyin Wu
Bowen Liang	Jamie Hall	Jay Pavagadhi	Zhuoyuan Chen
Daniel Toyama	Xiangzhuo Ding	Kristian Kjems	Zack Nado
Szabolcs Payrits	Gan Song	Archita Vadali	Bartek Perz
Anca Stefanou	Gang Li	Diego Machado	Zixuan Jiang
Dimitrios Vytiniotis	Chen Zhu	Yeqing Li	David Kao
Ankesh Anand	Yana Kulizhskaya	Renshen Wang	Ganesh Mallya

Nino Vieillard	Lalit Jain	Sobhan Miryoosefi	Elena Pochernina
Lantao Mei	Manaal Faruqui	Haitian Sun	Sheng Zhang
Sertan Girgin	Nicolas Lacasse	YaGuang Li	Parker Barnes
Mandy Jordan	Georgie Evans	Charlie Chen	Daisuke Ikeda
Yeongil Ko	Neesha Subramaniam	Jae Yoo	Qiuqia Li
Alekh Agarwal	Dean Reich	Pavel Dubov	Shuo-yiin Chang
Yixin Liu	Giulia Vezzani	Alex Tomala	Shakir Mohamed
Yasemin Altun	Aditya Pandey	Adams Yu	Jim Sproch
Raoul de Liedekerke	Joe Stanton	Paweł Wesołowski	Richard Powell
Anastasios Kementsietsidis	Tianhao Zhou	Alok Gunjan	Bidisha Samanta
Daiyi Peng	Liam McCafferty	Eddie Cao	Domagoj Ćević
Dangyi Liu	Henry Griffiths	Jiaming Luo	Anton Kovsharov
Utku Evci	Verena Rieser	Nikhil Sethi	Shrestha Basu Mallick
Peter Humphreys	Soheil Hassas Yeganeh	Arkadiusz Socala	Srinivas Tadepalli
Austin Tarango	Eleftheria Briakou	Laura Graesser	Anne Zheng
Xiang Deng	Lu Huang	Tomas Kociský	Kareem Ayoub
Yoad Lewenberg	Zichuan Wei	Arturo BC	Andreas Noever
Kevin Aydin	Liangchen Luo	Minmin Chen	Christian Reisswig
Chengda Wu	Erik Jue	Edward Lee	Zhuo Xu
Bhavishya Mittal	Gabby Wang	Sophie Wang	Junhyuk Oh
Tsendsuren Munkhdalai	Victor Cotruta	Weize Kong	Martin Matysiak
Kleopatra Chatziprimou	Myriam Khan	Qiantong Xu	Tim Blyth
Rodrigo Benenson	Jongbin Park	Nilesh Tripuraneni	Shereen Ashraf
Uri First	Qiuchen Guo	Yiming Li	Julien Amelot
Xiao Ma	Peiran Li	Xinxin Yu	Boone Severson
Jinning Li	Rong Rong	Allen Porter	Michele Bevilacqua
Armand Joulin	Diego Antognini	Paul Voigtlaender	Motoki Sano
Hamish Tomlinson	Anastasia Petrushkina	Biao Zhang	Ethan Dyer
Tingnan Zhang	Chetan Tekur	Arpi Vezer	Ofir Roval
Milad Nasr	Eli Collins	Sarah York	Anu Sinha
Zhi Hong	Parul Bhatia	Qing Wei	Yin Zhong
Michaël Sander	Chester Kwak	Geoffrey Cideron	Sagi Perel
Lisa Anne Hendricks	Wenhu Chen	Mark Kurzeja	Tea Sabolić
Anuj Sharma	Arvind Neelakantan	Seungyeon Kim	Johannes Mauerer
Andrew Bolt	Immanuel Odisho	Benny Li	Willi Gierke
Eszter Vértes	Sheng Peng	Angéline Pouget	Mauro Verzetti
Jiri Simsa	Vincent Nallatamby	Hyo Lee	Rodrigo Cabrera
Tomer Levinboim	Vaibhav Tulsyan	Kaspar Daugaard	Alvin Abdagic
Olcun Sercinoglu	Fabian Pedregosa	Yang Li	Steven Hemingray
Divyansh Shukla	Peng Xu	Dave Uthus	Austin Stone
Austin Wu	Raymond Lin	Aditya Siddhant	Jong Lee
Craig Swanson	Yulong Wang	Paul Cavallaro	Farooq Ahmad
Danny Vainstein	Emma Wang	Sriram Ganapathy	Karthik Raman
Fan Bu	Sholto Douglas	Maulik Shah	Lior Shani
Bo Wang	Reut Tsarfaty	Rolf Jagerman	Jonathan Lai
Ryan Julian	Elena Gribovskaya	Jeff Stanway	Orhan Firat
Charles Yoon	Renga Aravamudhan	Piermaria Mendolicchio	Nathan Waters
Sergei Lebedev	Manu Agarwal	Li Xiao	Eric Ge
Antonious Girgis	Mara Finkelstein	Kayi Lee	Mo Shomrat
Bernd Bandemer	Qiao Zhang	Tara Thompson	Himanshu Gupta
David Du	Elizabeth Cole	Shubham Milind Phal	Rajeev Aggarwal
Todd Wang	Phil Crone	Jason Chase	Tom Hudson
Xi Chen	Sarmishta Velury	Sun Jae Lee	Bill Jia
Ying Xiao	Anil Das	Adrian N Reyes	Simon Baumgartner
Peggy Lu	Chris Sauer	Disha Shrivastava	Palak Jain
Natalie Ha	Luyao Xu	Zhen Qin	Joe Kovac
Vlad Ionescu	Danfeng Qin	Roykrong Sukkerd	Junehyuk Jung
Simon Rowe	Chenjie Gu	Seth Odoom	Ante Žužul
Josip Matač	Dror Marcus	Lior Madmoni	Will Truong
Federico Lebron	CJ Zheng	John Aslanides	Morteza Zadimoghaddam
Andreas Steiner	Wouter Van Gansbeke	Jonathan Herzig	Songyou Peng

Marco Liang	Sujoy Basu	Hoang Nguyen	Johan Schalkwyk
Rachel Sterneck	Hidetoshi Shimokawa	Bat-Orgil Batsaikhan	Yuntao Xu
Balaji Lakshminarayanan	Nina Anderson	Iñaki Iturrate	Hui Huang
Machel Reid	Haiguang Li	Anselm Levskaya	Andrew Howard
Oliver Woodman	Khiem Pham	Avi Singh	Sasha Brown
Tong Zhou	Yizhong Liang	Ashyana Kachra	Linting Xue
Jianling Wang	Sebastian Borgeaud	Tony Lu	Gloria Shen
Vincent Coriou	Alexandre Moufarek	Denis Petek	Brian Albert
Arjun Narayanan	Hideto Kazawa	Zheng Xu	Neha Jha
Jay Hoover	Blair Kutzman	Mark Graham	Daniel Zheng
Yenai Ma	Marcin Sieniek	Lukas Zilka	Varvara Krayvanova
Apoorv Jindal	Sara Smoot	Yael Karov	Spurthi Amba Hombaih
Clayton Sanford	Ruth Wang	Marija Kostelac	Olivier Lacombe
Doug Reid	Natalie Axelsson	Fangyu Liu	Gautam Vasudevan
Swaroop Ramaswamy	Nova Fallen	Yaohui Guo	Dan Graur
Alex Kurakin	Prasha Sundaram	Weiyue Wang	Tian Xie
Roland Zimmermann	Yuexiang Zhai	Bernd Bohnet	Meet Gandhi
Yana Lunts	Varun Godbole	Emily Pitler	Bangju Wang
Dragos Dena	Petros Maniatis	Tony Bruguier	Dustin Zelle
Zalán Borsos	Alek Wang	Keisuke Kinoshita	Harman Singh
Vered Cohen	Ilia Shumailov	Chrysovalantis Anastasiou	Dahun Kim
Shujian Zhang	Santhosh Thangaraj	Nilpa Jha	Sébastien Cevey
Will Grathwohl	Remi Crocker	Ting Liu	Victor Ungureanu
Robert Dadashi	Nikita Gupta	Jerome Connor	Natasha Noy
Morgan Redshaw	Gang Wu	Phil Wallis	Fei Liu
Joshua Kessinger	Phil Chen	Philip Pham	Annie Xie
Julian Odell	Gellért Weisz	Eric Bailey	Fangxiao Yu Feng
Silvano Bonacina	Celine Smith	Shixin Li	Katerina Tsihas
Zihang Dai	Mojtaba Seyedhosseini	Heng-Tze Cheng	Daniel Formoso
Grace Chen	Boya Fang	Sally Ma	Neera Vats
Ayush Dubey	Xiyang Luo	Haiqiong Li	Quentin Wellens
Pablo Sprechmann	Roey Yogev	Akanksha Maurya	Yinan Wang
Mantas Pajarskas	Zeynep Cankara	Kate Olszewska	Niket Kumar Bhumihar
Wenxuan Zhou	Andrew Hard	Manfred Warmuth	Samrat Ghosh
Niharika Ahuja	Helen Ran	Christy Koh	Matt Hoffman
Tara Thomas	Rahul Sukthankar	Dominik Paulus	Tom Lieber
Martin Nikoltchev	George Necula	Siddhartha Reddy	Oran Lang
Matija Kecman	Gaël Liu	Jonnalagadda	Kush Bhatia
Bharath Mankalale	Honglong Cai	Enrique Piqueras	Tom Paine
Andrey Ryabtsev	Praseem Banzal	Ali Elqursh	Aroonalok Pyne
Jennifer She	Daniel Keysers	Geoff Brown	Ronny Votel
Christian Walder	Sanjay Ghemawat	Hadar Shemtov	Madeleine Clare Elish
Jiaming Shen	Connie Tao	Loren Maggiore	Benoit Schillings
Lu Li	Emma Dunleavy	Fei Xia	Alex Panagopoulos
Carolina Parada	Aditi Chaudhary	Ryan Foley	Haichuan Yang
Sheena Panthaplatchel	Wei Li	Beka Westberg	Adam Raveret
Okwan Kwon	Maciej Mikuła	George van den Driessche	Zohar Yahav
Matt Lawlor	Chen-Yu Lee	Livio Baldini Soares	Shuang Liu
Utsav Prabhu	Tiziana Refice	Arjun Kar	Dalia El Badawy
Yannick Schroecker	Krishna Somandepalli	Michael Quinn	Nishant Agrawal
Marc'aurelio Ranzato	Alexandre Fréchette	Siqi Zuo	Mohammed Badawi
Pete Blois	Dan Bahir	Jialin Wu	Mahdi Mirzazadeh
Iurii Kemaev	John Karro	Kyle Kastner	Carla Bromberg
Ting Yu	Keith Rush	Anna Bortsova	Fan Ye
Dmitry Lepikhin	Sarah Perrin	Aijun Bai	Chang Liu
Hao Xiong	Bill Rosgen	Ales Mikhalap	Tatiana Sholokhova
Sahand Sharifzadeh	Xiaomeng Yang	Luowei Zhou	George-Cristian Muraru
Oleaser Johnson	Clara Huiyi Hu	Jennifer Brennan	Gargi Balasubramaniam
Jeremiah Willcock	Mahmoud Alnahlawi	Vinay Ramasesh	Jonathan Malmaud
Rui Yao	Justin Mao-Jones	Honglei Zhuang	Alen Carin
Greg Farquhar	Roopal Garg	John Maggs	Danilo Martins

Irina Jurenka	Uri Alon	Jared Lichtarge	Stanislav Fort
Pankil Botadra	Stephanie Winkler	Adam R. Brown	Yifan He
Dave Lacey	Tzu-Kuo Huang	Hilal Dib	Alex Gurney
Richa Singh	Kostas Andriopoulos	Natalia Ponomareva	Bryan Gale
Mariano Schain	João Gabriel Oliveira	Justin Fu	Yue Ma
Dan Zheng	Penporn Koanantakool	Yujing Zhang	Monica Roy
Isabelle Guyon	Berkin Akin	Altaf Rahman	Viorica Patraucean
Victor Lavrenko	Michael Wunder	Joana Iljazi	Taylan Bilal
Seungji Lee	Cicero Nogueira dos Santos	Edouard Leurent	Golnaz Ghiasi
Xiang Zhou	Mohammad Hossein Bateni	Gabriel Dulac-Arnold	Anahita Hosseini
Demis Hassabis	Lin Yang	Cosmo Du	Melvin Johnson
Jeshwanth Challagundla	Dan Horgan	Chulayuth Asawaroengchai	Zhuowan Li
Derek Cheng	Beer Changpinyo	Larry Jin	Yi Tay
Nikhil Mehta	Keyvan Amiri	Ela Gruzevska	Benjamin Beyret
Matthew Mauger	Min Ma	Ziwei Ji	Katie Millican
Michela Paganini	Dayeong Lee	Benigno Uria	Josef Broder
Pushkar Mishra	Lihao Liang	Daniel De Freitas	Mayank Lunayach
Kate Lee	Anirudh Baddepudi	Paul Barham	Danny Swisher
Zhang Li	Tejasvi Latkar	Lauren Beltrone	Eugen Vušak
Lexi Baugher	Raia Hadsell	Víctor Campos	David Parkinson
Ondrej Skopek	Jun Xu	Jun Yan	MH Tessler
Max Chang	Hairong Mu	Neel Kovelamudi	Adi Mayrav Gilady
Amir Zait	Michael Han	Arthur Nguyen	Richard Song
Gaurav Menghani	Aedan Pope	Elinor Davies	Allan Dafoe
Lizzeth Bellot	Sanchit Grover	Zhichun Wu	Yves Raimond
Guangxing Han	Frank Kim	Zoltan Egyed	Masa Yamaguchi
Jean-Michel Sarr	Ankit Bhagatwala	Kristina Toutanova	Itay Karo
Sharat Chikkerur	Guan Sun	Nithya Attaluri	Elizabeth Nielsen
Himanshu Sahni	Yamini Bansal	Hongliang Fei	Kevin Kilgour
Rohan Anil	Amir Globerson	Peter Stys	Mike Dusenberry
Arun Narayanan	Alireza Nazari	Siddhartha Brahma	Rajiv Mathews
Chandu Thekkath	Samira Daruki	Martin Izzard	Jiho Choi
Daniele Pighin	Hagen Soltau	Siva Velusamy	Siyuan Qiao
Hana Strejček	Jane Labanowski	Scott Lundberg	Harsh Mehta
Marko Velić	Laurent El Shafey	Vincent Zhuang	Sahitya Potluri
Fred Bertsch	Matt Harvey	Kevin Sequeira	Chris Knutson
Manuel Tragut	Yanif Ahmad	Adam Santoro	Jialu Liu
Keran Rong	Elan Rosenfeld	Ehsan Amid	Tat Tan
Alicia Parrish	William Kong	Ophir Aharoni	Kuntal Sengupta
Kai Bailey	Etienne Pot	Shuai Ye	Keerthana Gopalakrishnan
Jiho Park	Yi-Xuan Tan	Mukund Sundararajan	Abodonrinwa Toki
Isabela Albuquerque	Aurora Wei	Lijun Yu	Mencher Chiang
Abhishek Bapna	Victoria Langston	Yu-Cheng Ling	Mike Burrows
Rajesh Venkataraman	Marcel Prasetya	Stephen Spencer	Grace Vesom
Alec Kosik	Petar Veličković	Hugo Song	Zafarali Ahmed
Johannes Griesser	Richard Killam	Josip Djolonga	Ilia Labzovsky
Zhiwei Deng	Robin Strudel	Christo Kirov	Siddharth Vashishtha
Alek Andreev	Darren Ni	Sonal Gupta	Preeti Singh
Qingyun Dou	Zhenhai Zhu	Alessandro Bissacco	Ankur Sharma
Kevin Hui	Aaron Archer	Clemens Meyer	Ada Ma
Fanny Wei	Kavya Kopparapu	Mukul Bhutani	Jinyu Xie
Xiaobin Yu	Lynn Nguyen	Andrew Dai	Pranav Talluri
Lei Shu	Emilio Parisotto	Weiyi Wang	Hannah Forbes-Pollard
Avia Aharon	Hussain Masoom	Siqi Liu	Aarush Selvan
David Barker	Sravanti Addepalli	Ashwin Sreevatsa	Joel Wee
Badih Ghazi	Jordan Grimstad	Qijun Tan	Loic Matthey
Sebastian Flennerhag	Hexiang Hu	Maria Wang	Tom Funkhouser
Chris Breaux	Joss Moore	Lucy Kim	Parthasarathy Gopavarapu
Yuchuan Liu	Avinatan Hassidim	Yicheng Wang	Lev Proleev
Matthew Bilotti	Le Hou	Alex Irpan	Cheng Li
Josh Woodward	Mukund Raghavachari	Yang Xiao	Matt Thomas

Kashyap Kolipaka	Paul Roit	Bethanie Brownfield	Yuankai Chen
Zhipeng Jia	Isabel Edkins	Averi Nowak	Vikas Verma
Ashwin Kakarla	Andy Davis	Mikel Rodriguez	Vladimir Pchelin
Srinivas Sunkara	Yujia Li	Tianze Shi	Harish Rajamani
Joan Puigcerver	Ben Horn	Hado van Hasselt	Valentin Dalibard
Suraj Satishkumar Sheth	Xinjian Li	Kevin Cen	Ana Ramalho
Emily Graves	Pradeep Kumar S	Deepanway Ghoshal	Yang Guo
Chen Wang	Eric Doi	Kushal Majmundar	Kartikeya Badola
Sadh MNM Khan	Wanzheng Zhu	Weiren Yu	Seojin Bang
Kai Kang	Sri Gayatri Sundara	Warren (Weilun) Chen	Nathalie Rauschmayr
Shyamal Buch	Padmanabhan	Danila Sinopalnikov	Julia Proskurnia
Fred Zhang	Siddharth Verma	Hao Zhang	Sudeep Dasari
Omkar Savant	Jasmine Liu	Vlado Galić	Xinyun Chen
David Soergel	Heng Chen	Di Lu	Mikhail Sushkov
Kevin Lee	Mihajlo Velimirović	Zeyu Zheng	Anja Hauth
Linda Friso	Malcolm Reynolds	Maggie Song	Pauline Sho
Xuanyi Dong	Priyanka Agrawal	Gary Wang	Abhinav Singh
Rahul Arya	Nick Sukhanov	Gui Citovsky	Bilva Chandra
Shreyas	Abhinit Modi	Swapnil Gawde	Allie Culp
Chandrakaladharan	Siddharth Goyal	Isaac Galatzer-Levy	Max Dylla
Connor Schenck	John Palowitch	David Silver	Olivier Bachem
Greg Billock	Nima Khajehnouri	Ivana Balazevic	James Besley
Tejas Iyer	Wing Lowe	Dipanjan Das	Heri Zhao
Anton Bakalov	David Klinghoffer	Kingshuk Majumder	Timothy Lillicrap
Leslie Baker	Sharon Silver	Yale Cong	Wei Wei
Alex Ruiz	Vinh Tran	Praneet Dutta	Wael Al Jishi
Angad Chandorkar	Candice Schumann	Dustin Tran	Ning Niu
Trieu Trinh	Francesco Piccinno	Hui Wan	Alban Rustemi
Matt Miecnikowski	Xi Liu	Junwei Yuan	Raphaël Lopez Kaufman
Yanqi Zhou	Mario Lučić	Daniel Eppens	Ryan Poplin
Yangsibo Huang	Xiaochen Yang	Alanna Walton	Jewel Zhao
Jiazhong Nie	Sandeep Kumar	Been Kim	Minh Truong
Ali Shah	Ajay Kannan	Harry Ragan	Shikhar Bharadwaj
Ashish Thapliyal	Ragha Kotikalapudi	James Cobon-Kerr	Ester Hlavnova
Sam Haves	Mudit Bansal	Lu Liu	Eli Stickgold
Lun Wang	Fabian Fuchs	Weijun Wang	Cordelia Schmid
Uri Shaham	Mohammad Javad Hosseini	Bryce Petrini	Georgi Stephanov
Patrick Morris-Suzuki	Abdelrahman Abdelhamed	Jack Rae	Zhaoqi Leng
Soroush Radpour	Dawn Bloxwich	Rakesh Shivanna	Frederick Liu
Leonard Berrada	Tianhe Yu	Yan Xiong	Léonard Huszenot
Thomas Strohmann	Ruoxin Sang	Chace Lee	Shenil Dodhia
Chaochao Yan	Gregory Thornton	Pauline Coquinot	Juliana Vicente Franco
Jingwei Shen	Karan Gill	Yiming Gu	Lesley Katzen
Sonam Goenka	Yuchi Liu	Lisa Patel	Abhanshu Sharma
Tris Warkentin	Virat Shejwalkar	Blake Hechtman	Sarah Cogan
Petar Dević	Jason Lin	Aviel Boag	Zuguang Yang
Dan Belov	Zhipeng Yan	Orion Jankowski	Aniket Ray
Albert Webson	Kehang Han	Alex Wertheim	Sergi Caelles
Madhavi Yenugula	Thomas Buschmann	Alex Lee	Shen Yan
Puranjay Datta	Michael Pliskin	Paul Covington	Ravin Kumar
Jerry Chang	Zhi Xing	Hila Noga	Daniel Gillick
Nimesh Ghelani	Susheel Tatineni	Sam Sobell	Renee Wong
Aviral Kumar	Junlin Zhang	Shanthal Vasanth	Joshua Ainslie
Vincent Perot	Sissie Hsiao	William Bono	Jonathan Hoech
Jessica Lo	Gavin Buttimore	Chirag Nagpal	Séb Arnold
Yang Song	Marcus Wu	Wei Fan	Dan Abolafia
Herman Schmit	Zefei Li	Xavier Garcia	Anca Dragan
Jianmin Chen	Geza Kovacs	Kedar Soparkar	Ben Hora
Vasilisa Bashlovkina	Legg Yeung	Aybuke Turker	Grace Hu
Xiaoyue Pan	Tao Huang	Nathan Howard	Alexey Guseynov
Diana Mincu	Aaron Cohen	Sachit Menon	Yang Lu

Chas Leichner	Sebastien Baur	Ndidi Elue	Keshav Shivam
Jinmeng Rao	Yiran Mao	Zhouyuan Huo	Sho Arora
Abhimanyu Goyal	Subhashini Venugopalan	Pooya Moradi	Partha Talukdar
Nagabhushan Baddi	Will Song	Jean Tarbouriech	Tom Cobley
Daniel Hernandez Diaz	Wen Ding	Henryk Michalewski	Sangnie Bhardwaj
Tim McConnell	Paul Collins	Wenting Ye	Evgeny Gladchenko
Max Bain	Sashank Reddi	Eunyoung Kim	Simon Green
Jake Abernethy	Megan Shum	Alex Druinsky	Kelvin Guu
Qiqi Yan	Andrei Rusu	Florent Altché	Felix Fischer
Rylan Schaeffer	Luisa Zintgraf	Xinyi Chen	Xiao Wu
Paul Vicol	Kelvin Chan	Artur Dwornik	Eric Wang
Will Thompson	Sheela Goenka	Da-Cheng Juan	Achintya Singhal
Montse Gonzalez Arenas	Mathieu Blondel	Rivka Moroshko	Tatiana Matejovicova
Mathias Bellaiche	Michael Collins	Horia Toma	James Martens
Pablo Barrio	Renke Pan	Jarrod Kahn	Hongji Li
Stefan Zinke	Marissa Giustina	Hai Qian	Roma Patel
Riccardo Patana	Nikolai Chinaev	Maximilian Sieb	Elizabeth Kemp
Pulkit Mehta	Christian Schuler	Irene Cai	Jiaqi Pan
JK Kearns	Ce Zheng	Roman Goldenberg	Lily Wang
Avraham Ruderman	Jonas Valfridsson	Praneeth Netrapalli	Blake JianHang Chen
Scott Pollom	Alyssa Loo	Sindhu Raghuram	Jean-Baptiste Alayrac
David D'Ambrosio	Alex Yakubovich	Yuan Gong	Navneet Potti
Cath Hope	Jamie Smith	Lijie Fan	Erika Gemzer
Yang Yu	Tao Jiang	Evan Palmer	Eugene Ie
Andrea Gesmundo	Rich Munoz	Yossi Matias	Kay McKinney
Kuang-Huei Lee	Gabriel Barcik	Valentin Gabeur	Takaaki Saeki
Aviv Rosenberg	Rishabh Bansal	Shreya Pathak	Edward Chou
Yiqian Zhou	Mingyao Yang	Tom Ouyang	Pascal Lamblin
Yaoyiran Li	Yilun Du	Don Metzler	SQ Mah
Drew Garmon	Pablo Duque	Geoff Bacon	Zach Fisher
Yonghui Wu	Mary Phuong	Srinivasan Venkatachary	Martin Chadwick
Safeen Huda	Alexandra Belias	Sridhar Thiagarajan	Jon Stritar
Gil Fidel	Kunal Lad	Alex Cullum	Obaid Sarvana
Martin Baeuml	Zeyu Liu	Eran Ofek	Andrew Hogue
Jian Li	Tal Schuster	Vytenis Sakenas	Artem Shtefan
Phoebe Kirk	Karthik Duddu	Mohamed Hammad	Hadi Hashemi
Rhys May	Jieru Hu	Cesar Magalhaes	Yang Xu
Tao Tu	Paige Kunkle	Mayank Daswani	Jindong Gu
Sara Mc Carthy	Matthew Watson	Oscar Chang	Sharad Vikram
Toshiyuki Fukuzawa	Jackson Tolins	Ashok Popat	Chung-Ching Chang
Miranda Aperghis	Josh Smith	Ruichao Li	Sabela Ramos
Chih-Kuan Yeh	Denis Teplyashin	Komal Jalan	Logan Kilpatrick
Toshihiro Yoshino	Garrett Bingham	Yanhan Hou	Weijuan Xi
Bo Li	Marvin Ritter	Josh Lipschultz	Jenny Brennan
Austin Myers	Marco Andreetto	Antoine He	Yinghao Sun
Kaisheng Yao	Divya Pitta	Wenhao Jia	Abhishek Jindal
Ben Limonchik	Mohak Patel	Pier Giuseppe Sessa	Ionel Gog
Changwan Ryu	Shashank Viswanadha	Prateek Kolhar	Dawn Chen
Rohun Saxena	Trevor Strohman	William Wong	Felix Wu
Alex Goldin	Catalin Ionescu	Sumeet Singh	Jason Lee
Ruizhe Zhao	Jincheng Luo	Lukas Haas	Sudhindra Kopalle
Rocky Rhodes	Yogesh Kalley	Jay Whang	Srinadh Bhojanapalli
Tao Zhu	Jeremy Wiesner	Hanna Klimczak-Plucińska	Oriol Vinyals
Divya Tyam	Dan Deutsch	Georges Rotival	Natan Potikha
Heidi Howard	Derek Lockhart	Grace Chung	Burcu Karagol Ayan
Nathan Byrd	Peter Choy	Yiqing Hua	Yuan Yuan
Hongxu Ma	Rumen Dangovski	Anfal Siddiqui	Michael Riley
Yan Wu	Chawin Sitawarin	Nicolas Serrano	Piotr Stanczyk
Ryan Mullins	Cat Graves	Dongkai Chen	Sergey Kishchenko
Qingze Wang	Tanya Lando	Billy Porter	Bing Wang
Aida Amini	Joost van Amersfoort	Libin Bai	Dan Garrette

Antoine Yang	Nguyet Minh Phu	Kelly Chen	Shruti Rijhwani
Vlad Feinberg	Isaac Tian	Jingchen Ye	Morgane Rivière
CJ Carey	Junwhan Ahn	Xuanhui Wang	Yuanzhong Xu
Javad Azizi	Andy Crawford	James Svensson	Junjie Wang
Viral Shah	Lauren Lax	Philipp Fränken	Xinyang Geng
Erica Moreira	Yuan Shangguan	Josh Newlan	Xiance Si
Chongyang Shi	Iftekhar Naim	Li Lao	Arjun Khare
Josh Feldman	David Ross	Eva Schnider	Cheolmin Kim
Elizabeth Salesky	Oleksandr Ferludin	Sami Alabed	Vahab Mirrokni
Thomas Lampe	Tongfei Guo	Joseph Kready	Kamyu Lee
Aneesh Pappu	Andrea Banino	Jesse Emond	Khuslen Baatarsukh
Duhyeon Kim	Hubert Soyer	Afief Halumi	Nathaniel Braun
Jonas Adler	Xiaoen Ju	Tim Zaman	Lisa Wang
Avi Caciularu	Dominika Rogozińska	Chengxi Ye	Pallavi LV
Brian Walker	Ishaan Malhi	Naina Raisinghani	Richard Tanburn
Yunhan Xu	Marcella Valentine	Vilobh Meshram	Yuvein (Yonghao) Zhu
Yochai Blau	Daniel Balle	Bo Chang	Fangda Li
Dylan Scandinaro	Apoorv Kulshreshtha	Ankit Singh Rawat	Setareh Ariaifar
Terry Huang	Maciej Kula	Axel Stjerngren	Dan Goldberg
Sam El-Husseini	Yiwen Song	Sergey Levi	Ken Burke
Abhishek Sinha	Sophia Austin	Rui Wang	Daniil Mirylenka
Lijie Ren	John Schultz	Xiangzhu Long	Meiqi Guo
Taylor Tobin	Roy Hirsch	Mitchelle Rasquinha	Olaf Ronneberger
Patrik Sundberg	Arthur Douillard	Steven Hand	Hadas Natalie Vogel
Tim Sohn	Apoorv Reddy	Aditi Mavalankar	Liqun Cheng
Vikas Yadav	Michael Fink	Lauren Agubuzu	Nishita Shetty
Mimi Ly	Summer Yue	Sudeshna Roy	Johnson Jia
Emily Xue	Khyatti Gupta	Junquan Chen	Thomas Jimma
Jing Xiong	Adam Zhang	Jarek Wilkiewicz	Corey Fry
Afzal Shama Soudagar	Norman Rink	Hao Zhou	Ted Xiao
Sneha Mondal	Daniel McDuff	Michał Jastrzebski	Martin Sundermeyer
Nikhil Khadke	Lei Meng	Qiong Hu	Ryan Burnell
Qingchun Ren	András György	Agustin Dal Lago	Yannis Assael
Ben Vargas	Yasaman Razeghi	Ramya Sree Boppana	Mario Pinto
Stan Bileshi	Ricky Liang	Wei-Jen Ko	JD Chen
Sarah Chakera	Kazuki Osawa	Jennifer Prendki	Rohit Sathyaranayana
Cindy Wang	Aviel Atias	Yao Su	Donghyun Cho
Boyu Wang	Matan Eyal	Zhi Li	Jing Lu
Yoni Halpern	Tyrone Hill	Eliza Rutherford	Rishabh Agarwal
Joe Jiang	Nikolai Grigorev	Girish Ramchandra Rao	Sugato Basu
Vikas Sindhwani	Zhengdong Wang	Ramona Comanescu	Lucas Gonzalez
Petre Petrov	Nitish Kulkarni	Adrià Puigdomènech	Dhruv Shah
Pranavaraj Ponnuramu	Rachel Soh	Qihang Chen	Meng Wei
Sanket Vaibhav Mehta	Ivan Lobov	Dessie Petrova	Dre Mahaarachchi
Yu Watanabe	Zachary Charles	Christine Chan	Rohan Agrawal
Betty Chan	Sid Lail	Vedrana Milutinovic	Tero Rissa
Matheus Wisniewski	Kazuma Hashimoto	Felipe Tiengo Ferreira	Yani Donchev
Trang Pham	Ido Kessler	Chin-Yi Cheng	Ramiro Leal-Cavazos
Jingwei Zhang	Victor Gomes	Ming Zhang	Adrian Hutter
Conglong Li	Zelda Mariet	Tapomay Dey	Markus Mircea
Dario de Cesare	Danny Driess	Sherry Yang	Alon Jacovi
Art Khurshudov	Alessandro Agostini	Ramesh Sampath	Faruk Ahmed
Alex Vasiloff	Canfer Akbulut	Quoc Le	Jiageng Zhang
Melissa Tan	Jingcao Hu	Howard Zhou	Shuguang Hu
Zoe Ashwood	Marissa Ikonomidis	Chu-Cheng Lin	Bo-Juen Chen
Bobak Shahriari	Emily Caveness	Hoi Lam	Jonni Kanerva
Maryam Majzoubi	Kartik Audhkhasi	Christine Kaeser-Chen	Guillaume Desjardins
Garrett Tanzer	Saurabh Agrawal	Kai Hui	Andrew Lee
Olga Kozlova	Ioana Bica	Dean Hirsch	Nikos Parotsidis
Robin Alazard	Evan Senter	Tom Eccles	Asier Mujika
James Lee-Thorp	Jayaram Mudigonda	Basil Mustafa	Tobias Weyand

Jasper Snoek	Georgi Karadzhov	Tammo Spalink	Anirudh GP
Jo Chick	Guillermo Garrido	Mingyang Ling	Varun Yerram
Kai Chen	Ankur Bapna	Arun Nair	Sage Stevens
Paul Chang	Jiawei Cao	Ga-Young Joung	Tianqi Liu
Ethan Mahintorabi	Adam Sadovsky	Linda Deng	Noah Fiedel
Zi Wang	Pouya Tafti	Avishkar Bhoopchand	Charles Sutton
Tolly Powell	Arthur Guez	Lora Aroyo	Matthew Johnson
Orgad Keller	Coline Devin	Tom Duerig	Xiaodan Song
Abhirut Gupta	Yixian Di	Jordan Griffith	Kate Baumli
Claire Sha	Jinwei Xing	Gabe Barth-Maron	Nir Shabat
Kanav Garg	Chuqiao (Joyce) Xu	Jake Ades	Muqthar Mohammad
Nicolas Heess	Hanzhao Lin	Alex Haig	Hao Liu
Ágoston Weisz	Chun-Te Chu	Ankur Taly	Marco Selvi
Cassidy Hardin	Sameera Ponda	Yunting Song	Yichao Zhou
Bartek Wydrowski	Wesley Helmholz	Paul Michel	Mehdi Hafezi Manshadi
Ben Coleman	Fan Yang	Dave Orr	Chu-ling Ko
Karina Zainullina	Yue Gao	Dean Weesner	Anthony Chen
Pankaj Joshi	Sara Javanmardi	Corentin Tallec	Michael Bendersky
Alessandro Epasto	Wael Farhan	Carrie Grimes Bostock	Jorge Gonzalez Mendez
Terry Spitz	Alex Ramirez	Paul Niemczyk	Nisarg Kothari
Binbin Xiong	Ricardo Figueira	Andy Twigg	Amir Zandieh
Kai Zhao	Khe Chai Sim	Mudit Verma	Yiling Huang
Arseniy Klimovskiy	Yuval Bahat	Rohith Vallu	Daniel Andor
Ivy Zheng	Ashwin Vaswani	Henry Wang	Ellie Pavlick
Johan Ferret	Liangzhe Yuan	Marco Gelmi	Idan Brusilovsky
Itay Yona	Gufeng Zhang	Kiranbir Sodhia	Jitendra Harlalka
Waleed Khawaja	Leland Rechis	Aleksandr Chuklin	Sally Goldman
Jean-Baptiste Lespiau	Hanjun Dai	Omer Goldman	Andrew Lampinen
Maxim Krikun	Tayo Oguntebi	Jasmine George	Guowang Li
Siamak Shakeri	Alexandra Cordell	Liang Bai	Asahi Ushio
Timothée Cour	Eugénie Rives	Kelvin Zhang	Somit Gupta
Bonnie Li	Kaan Tekelioglu	Petar Sirkovic	Lei Zhang
Igor Krivokon	Naveen Kumar	Efrat Nehoran	Chuyuan Kelly Fu
Dan Suh	Bing Zhang	Golan Pundak	Madhavi Sewak
Alex Hofer	Aurick Zhou	Jiaqi Mu	Timo Denk
Jad Al Abdallah	Nikolay Savinov	Alice Chen	Jed Borovik
Nikita Putikhin	Andrew Leach	Alex Greve	Brendan Jou
Oscar Akerlund	Alex Tudor	Paulo Zacchello	Avital Zipori
Silvio Lattanzi	Sanjay Ganapathy	David Amos	Prateek Jain
Anurag Kumar	Yanyan Zheng	Heming Ge	Junwen Bai
Shane Settle	Mirko Rossini	Eric Noland	Thang Luong
Himanshu Srivastava	Vera Axelrod	Colton Bishop	Jonathan Tompson
Folawayo Campbell-Ajala	Arnaud Autef	Jeffrey Dudek	Alice Li
Edouard Rosseel	Yukun Zhu	Youhei Namiki	Li Liu
Mihai Dorin Istin	Zheng Zheng	Elena Buchatskaya	George Powell
Nishanth Dikkala	Mingda Zhang	Jing Li	Jiajun Shen
Anand Rao	Baochen Sun	Dorsa Sadigh	Alex Feng
Nick Young	Jie Ren	Masha Samsikova	Grishma Chole
Kate Lin	Nenad Tomasev	Dan Malkin	Da Yu
Dhruva Bhaswar	Nithish Kannan	Damien Vincent	Yinlam Chow
Yiming Wang	Amer Sinha	Robert David	Tongxin Yin
Jaume Sanchez Elias	Charles Chen	Rob Willoughby	Eric Malmi
Kritika Muralidharan	Louis O'Bryan	Phoenix Meadowlark	Kefan Xiao
James Keeling	Alex Pak	Shawn Gao	Yash Pande
Dayou Du	Aditya Kusupati	Yan Li	Shachi Paul
Siddharth Gopal	Weel Yang	Raj Apte	Niccolò Dal Santo
Gregory Dibb	Deepak Ramachandran	Amit Jhindal	Adil Dostmohamed
Charles Blundell	Patrick Griffin	Stein Xudong Lin	Sergio Guadarrama
Manolis Delakis	Seokhwan Kim	Alex Polozov	Aaron Phillips
Jacky Liang	Philipp Neubeck	Zhicheng Wang	Thanumalayan
Marco Tulio Ribeiro	Craig Schiff	Tomas Mery	Sankaranarayana Pillai

Gal Yona	Tao Li	Yunxiao Deng	Sharath Maddineni
Amin Ghafouri	Myle Ott	Ashutosh Sathe	Chris Rawles
Preethi Lahoti	Félix de Chaumont Quiry	Kacper Krasowiak	Mina Khan
Benjamin Lee	David Vilar Torres	Ciprian Chelba	Shlomi Cohen-Ganor
Dhruv Madeka	Yuri Chervonyi	Cho-Jui Hsieh	Amol Mandhane
Eren Sezener	Tomy Tsai	Kiran Vodrahalli	Xinyi Wu
Simon Tokumine	Prem Erubetidine	Buhuang Liu	Chenkai Kuang
Adrian Collister	Samuel Yang	Thomas Köppe	Iulia Comşa
Nicola De Cao	Matthew Denton	Amr Khalifa	Ramya Ganeshan
Richard Shin	Jake Walker	Lubo Litchev	Hanie Sedghi
Uday Kalra	Slavica Andačić	Pichi Charoenpanit	Adam Bloniarz
Parker Beak	Idan Heimlich Shtacher	Reed Roberts	Nuo Wang Pierse
Emily Nottage	Vittal Premachandran	Sachin Yadav	Anton Briukhov
Ryo Nakashima	Harshal Tushar Lehri	Yasumasa Onoe	Petr Mitrichev
Ivan Jurin	Cip Baetu	Desi Ivanov	Anita Gergely
Vikash Sehwag	Damion Yates	Megha Mohabey	Serena Zhan
Meenu Gaba	Lampros Lamprou	Vighnesh Birodkar	Allan Zhou
Junhao Zeng	Mariko Iinuma	Nemanja Rakićević	Nikita Saxena
Kevin R. McKee	Ioana Mihailescu	Pierre Sermanet	Eva Lu
Fernando Pereira	Ben Albrecht	Vaibhav Mehta	Josef Dean
Tamar Yakar	Shachi Dave	Krishan Subudhi	Ashish Gupta
Amayika Panda	Susie Sargsyan	Travis Choma	Nicolas Perez-Nieves
Arka Dhar	Bryan Perozzi	Will Ng	Renjie Wu
Peilin Zhong	Lucas Manning	Luheng He	Cory McLean
Daniel Sohn	Chiyuan Zhang	Kathie Wang	Wei Liang
Mark Brand	Denis Vnukov	Tasos Kementsietsidis	Disha Jindal
Lars Lowe Sjoesund	Igor Mordatch	Shane Gu	Anton Tsitsulin
Viral Carpenter	Raia Hadsell Wolfgang	Mansi Gupta	Wenhao Yu
Sharon Lin	Macherey	Andrew Nystrom	Kaiz Alarakyia
Shantanu Thakoor	Ryan Kappedal	Mehran Kazemi	Tom Schaul
Marcus Wainwright	Jim Stephan	Timothy Chung	Piyush Patil
Ashwin Chaugule	Aditya Tripathi	Nacho Cano	Peter Sung
Pranesh Srinivasan	Klaus Macherey	Nikhil Dhawan	Elijah Peake
Muye Zhu	Jun Qian	Yufei Wang	Hongkun Yu
Bennett Orlando	Abhishek Bhowmick	Jiawei Xia	Feryal Behbahani
Jack Weber	Shekoofeh Azizi	Trevor Yacovone	JD Co-Reyes
Ayzaan Wahid	Rémi Leblond	Eric Jia	Alan Ansell
Gilles Baechler	Shiva Mohan Reddy	Mingqing Chen	Sean Sun
Apurv Suman	Garlapati	Simeon Ivanov	Clara Barbu
Jovana Mitrović	Timothy Knight	Ashrith Sheshan	Jonathan Lee
Gabe Taubman	Matthew Wiethoff	Sid Dalmia	Seb Noury
Honglin Yu	Wei-Chih Hung	Paweł Stradomski	James Allingham
Helen King	Anelia Angelova	Pengcheng Yin	Bilal Piot
Josh Dillon	Georgios Evangelopoulos	Salem Haykal	Mohit Sharma
Cathy Yip	Pawel Janus	Congchao Wang	Christopher Yew
Dhriti Varma	Dimitris Paparas	Dennis Duan	Ivan Korotkov
Tomas Izo	Matthew Rahtz	Neslihan Bulut	Bibo Xu
Levent Bolelli	Ken Caluwaerts	Greg Kochanski	Demetra Brady
Borja De Balle Pigem	Vivek Sampathkumar	Liam MacDermed	Goran Petrovic
Julia Di Trapani	Daniel Jarrett	Namrata Godbole	Shibl Mourad
Fotis Iliopoulos	Shadi Noghabi	Shitao Weng	Claire Cui
Adam Paszke	Antoine Miech	Jingjing Chen	Aditya Gupta
Nishant Ranka	Chak Yeung	Rachana Fellinger	Parker Schuh
Joe Zou	Geoff Clark	Ramin Mehran	Saarthak Khanna
Francesco Pongetti	Henry Prior	Daniel Suo	Anna Goldie
Jed McGiffin	Fei Zheng	Hisham Husain	Abhinav Arora
Alex Siegman	Jean Pouget-Abadie	Tong He	Vadim Zubov
Rich Galt	Indro Bhattacharya	Kaushal Patel	Amy Stuart
Ross Hemsley	Kalpesh Krishna	Joshua Howland	Mark Epstein
Goran Žužić	Will Bishop	Randall Parker	Yun Zhu
Victor Carbune	Zhe Yuan	Kelvin Nguyen	Jianqiao Liu

Yury Stuken	Yunhsuan Sung	Sarah Nguyen	Zi Yang
Ziyue Wang	Jane Shapiro	Michael Guzman	Kenny Vassigh
Karolis Misiunas	Shaan Bijwadia	AJ Maschinot	Maria Bauza
Dee Guo	Chris Duvarney	Marcello Maggioni	Sheng Li
Ashleah Gill	Christina Sorokin	Ming-Wei Chang	Yiqing Tao
Ale Hartman	Paul Natsev	Karol Gregor	Nevan Wichers
Zaid Nabulsi	Reeve Ingle	Lotte Weerts	Andrii Maksai
Aurko Roy	Pramod Gupta	Kumaran Venkatesan	Abe Ittycheriah
Aleksandra Faust	Young Maeng	Bogdan Damoc	Ross Mcilroy
Jason Riesa	Ndaba Ndebele	Leon Liu	Bryan Seybold
Ben Withbroe	Kexin Zhu	Jan Wassenberg	Noah Goodman
Mengchao Wang	Valentin Anklin	Lewis Ho	Romina Datta
Marco Tagliasacchi	Katherine Lee	Becca Roelofs	Steven M. Hernandez
Andreea Marzoca	Yuan Liu	Majid Hadian	Tian Shi
James Noraky	Yaroslav Akulov	François-Xavier Aubet	Yony Kochinski
Serge Toropov	Shaleen Gupta	Yu Liang	Anna Bulanova
Malika Mehrotra	Guolong Su	Sami Lachgar	Ken Franko
Bahram Raad	Flavien Prost	Danny Karmon	Mikita Sazanovich
Sanja Deur	Tianlin Liu	Yong Cheng	Nicholas FitzGerald
Steve Xu	Vitaly Kovalev	Amelio Vázquez-Reina	Praneeth Kacham
Marianne Monteiro	Pol Moreno	Angie Chen	Shubha Srinivas
Zhongru Wu	Martin Scholz	Zhuyun Dai	Raghvendra
Yi Luan	Sam Redmond	Andy Brock	Vincent Hellendoorn
Sam Ritter	Zongwei Zhou	Shubham Agrawal	Alexander Grushetsky
Nick Li	Alex Castro-Ros	Chenxi Pang	Julian Salazar
Håvard Garnes	André Susano Pinto	Peter Garst	Angeliki Lazaridou
Yanzhang He	Dia Kharrat	Mariella Sanchez-Vargas	Jason Chang
Martin Zlocha	Michal Yarom	Ivor Rendulic	Jan-Thorsten Peter
Jifan Zhu	Rachel Saputro	Aditya Ayyar	Sushant Kafle
Matteo Hessel	Jannis Bulian	Andrija Ražnatović	Yann Dauphin
Will Wu	Ben Caine	Olivia Ma	Abhishek Rao
Spandana Raj Babbula	Ji Liu	Roopali Vij	Filippo Graziano
Chizu Kawamoto	Abbas Abdolmaleki	Neha Sharma	Izhak Shafran
Yuanzhen Li	Shariq Iqbal	Ashwin Balakrishna	Yuguo Liao
Mehadi Hassen	Tautvydas Misiunas	Bingyuan Liu	Tianli Ding
Yan Wang	Mikhail Sirotenko	Ian Mackinnon	Geng Yan
Brian Wieder	Shefali Garg	Sorin Baltateanu	Grace Chu
James Freedman	Guy Bensky	Petra Poklukar	Zhao Fu
Yin Zhang	Huan Gui	Gabriel Ibagon	Vincent Roulet
Xinyi Bai	Xuezhi Wang	Colin Ji	Gabriel Rasskin
Tianli Yu	Raphael Koster	Hongyang Jiao	Duncan Williams
David Reitter	Mike Bernico	Isaac Noble	Shahar Drath
XiangHai Sheng	Da Huang	Wojciech Stokowiec	Alex Mossin
Mateo Wirth	Romal Thoppilan	Zhihao Li	Raphael Hoffmann
Aditya Kini	Trevor Cohn	Jeff Dean	Jordi Orbay
Dima Damen	Ben Golan	David Lindner	Francesco Bertolini
Mingcen Gao	Wenlei Zhou	Mark Omernick	Hila Sheftel
Rachel Hornung	Andrew Rosenberg	Kristen Chiafullo	Justin Chiu
Michael Voznesensky	Markus Freitag	Mason Dimarco	Siyang Xue
Brian Roark	Tynan Gangwani	Vitor Rodrigues	Yuheng Kuang
Adhi Kuncoro	Vincent Tsang	Vittorio Selo	Ferjad Naeem
Yuxiang Zhou	Anand Shukla	Garrett Honke	Swaroop Nath
Rushin Shah	Xiaoqi Ren	Xintian (Cindy) Wu	Nana Nti
Anthony Brohan	Minh Giang	Wei He	Phil Culliton
Kuangyuan Chen	Chi Zou	Adam Hillier	Kashyap Krishnakumar
James Wendt	Andre Elisseeff	Anhad Mohananey	Michael Isard
David Rim	Charline Le Lan	Vihari Piratla	Pei Sun
Paul Kishan Rubenstein	Dheeru Dua	Chang Ye	Ayan Chakrabarti
Jonathan Halcrow	Shuba Lall	Chase Malik	Nathan Clement
Michelle Liu	Pranav Shyam	Sebastian Riedel	Regev Cohen
Ty Geri	Frankie Garcia	Samuel Albanie	Arissa Wongpanich

GS Oh	Sammy Jerome	Chaitra Hegde	Vlad-Doru Ion
Ashwin Murthy	Siyang Qin	Mihaela Rosca	Younghoon Jun
Hao Zheng	Louis Rouillard	Angelo Scorza Scarpati	James Swirhun
Jessica Hamrick	Amir Yazdanbakhsh	Kyriakos Axiotis	Soroosh Mariooryad
Oskar Bunyan	Zizhao Zhang	Ying Xu	Jiao Sun
Suhas Ganesh	Shivani Agrawal	Zach Gleicher	Steve Chien
Nitish Gupta	Kaushik Shivakumar	Assaf Hurwitz Michaely	Rey Coaguila
Roy Frostig	Caden Lu	Mandar Sharma	Ariel Brand
John Wieting	Praveen Kallakuri	Sanil Jain	Yi Gao
Yury Malkov	Rachita Chhaparia	Christoph Hirnschall	Tom Kwiatkowski
Pierre Marcenac	Kanishka Rao	Tal Marian	Roei Aharoni
Zhixin (Lucas) Lai	Charles Kwong	Xuhui Jia	Cheng-Chun Lee
Xiaodan Tang	Asya Fadeeva	Kevin Mather	Mislav Žanić
Mohammad Saleh	Shitij Nigam	Kilol Gupta	Yichi Zhang
Fedir Zubach	Yan Virin	Linhai Qiu	Dan Ethier
Chinmay Kulkarni	Yuan Zhang	Nigamama Nayakanti	Vitaly Nikolaev
Huanjie Zhou	Balaji Venkatraman	Lucian Ionita	Pranav Nair
Vicky Zayats	Beliz Gunel	Steven Zheng	Yoav Ben Shalom
Nan Ding	Marc Wilson	Lucia Loher	Hen Fitoussi
Anshuman Tripathi	Huiyu Wang	Kurt Shuster	Jai Gupta
Arijit Pramanik	Abhinav Gupta	Igor Petrovski	Hongbin Liu
Patrik Zochbauer	Xiaowei Xu	Roshan Sharma	Dee Cattle
Harish Ganapathy	Adrien Ali Taïga	Rahma Chaabouni	Tolga Bolukbasi
Vedant Misra	Kareem Mohamed	Angel Yeh	Ben Murdoch
Zach Behrman	Doug Fritz	James An	Fantine Huot
Hugo Vallet	Daniel Rodriguez	Arushi Gupta	Yin Li
Mingyang Zhang	Zoubin Ghahramani	Steven Schwarcz	Chris Hahn
Mukund Sridhar	Harry Askham	Seher Ellis	Urvashi Khandelwal
Ye Jin	Lior Belenki	Sam Conway-Rahman	Frederik Benzing
Mohammad Babaeizadeh	James Zhao	Javier Snaider	Arthur Conmy
Suum Pöder	Rahul Gupta	Alex Zhai	Andrey Simanovsky
Megha Goel	Krzysztof Jastrzębski	James Atwood	Françoise Beaufays
Divya Jain	Takahiro Kosakai	Daniel Golovin	Eugene Weinstein
Tajwar Nasir	Kaan Katircioğlu	Liqian Peng	Tongzhou Chen
Shubham Mittal	Jon Schneider	Te I	Luke Leonhard
Tim Dozat	Rina Panigrahy	Vivian Xia	Bhuvana Ramabhadran
Diego Ardila	Konstantinos Bousmalis	Salvatore Scellato	
Aliaksei Severyn	Peter Grabowski	Mahan Malihi	
Fabio Pardo	Prajit Ramachandran	Arthur Bražinskas	

The development of Gemini is a large-scale collaborative effort involving over 3000 individuals across Google, including researchers, engineers, and operations staff. These individuals contributed their hard work and expertise across diverse areas, from foundational research and the development of model architecture, data, training, and infrastructure, through to evaluation and ensuring safety and security. We gratefully acknowledge the dedication and hard work of each contributor in making Gemini a reality. The order of contributors in the above list is random.

We are also grateful to the Google-independent developer Joel Zhang for his work on Gemini Plays Pokémon, and for sharing with us the design of his set-up.

## 8. Appendix

### 8.1. Evaluation additional details

Please see a description of the benchmarks considered, along with details of how scores in the main text were obtained in Table 11.

Benchmark	Description	Details
LiveCodeBench	Code generation in Python (Jain et al., 2024).	Results are taken from <a href="https://livecodebench.github.io/leaderboard.html">https://livecodebench.github.io/leaderboard.html</a> (1/1/2025 - 5/1/2025 in the UI) or, where not available, run internally by us. For Section 2.5 and Figure 3 and 4, results are calculated on the version of the eval corresponding to 10/05/2024 - 01/04/2025 in the UI, and are based on internal results.
Aider Polyglot	Code editing in C++, Go, Java, JavaScript Python and Rust (Gauthier, 2025). See <a href="https://aider.chat/2024/12/21/polyglot.html#the-polyglot-benchmark">https://aider.chat/2024/12/21/polyglot.html#the-polyglot-benchmark</a> for a full description of this task.	We report results on the “diff” or “diff-fenced” edit format (see <a href="https://aider.chat/docs/more/edit-formats.html">https://aider.chat/docs/more/edit-formats.html</a> for a description of the different formats). The score reported are the pass rate average of 3 trials. Numbers come from <a href="https://aider.chat/docs/leaderboards/">https://aider.chat/docs/leaderboards/</a>
SWE-bench Verified	Agentic coding: evaluates AI agents on real-world programming tasks from GitHub (Chowdhury et al., 2024; Jimenez et al., 2024).	Gemini uses an internal agentic harness equipped with tools to navigate the repo, edit files, and test the code. We report scores for two modes: performance of a single agentic trace (“single attempt”), and performance of a scaffold that samples multiple agentic traces and reranks them before evaluation using Gemini’s own judgement (“multiple attempts”). All evaluations are done with temperature=1, topp=0.99, topk=1024.
GPQA (diamond)	Challenging dataset of questions written by domain experts in biology, physics, and chemistry (Rein et al., 2024).	
Humanity’s Last Exam	Challenging dataset of questions written by domain experts in a wide range of disciplines, including mathematics, physics, chemistry, biology and computer science (Phan et al., 2025).	No tool use variant. Reported results are from <a href="https://scale.com/leaderboard/humanitys_last_exam">https://scale.com/leaderboard/humanitys_last_exam</a> . For DeepSeek they are taken from <a href="https://scale.com/leaderboard/humanitys_last_exam_text_only">https://scale.com/leaderboard/humanitys_last_exam_text_only</a> (leaderboard for performance on the text-only questions) and in the case of the Gemini 2.0 models, these results are on an earlier HLE dataset, obtained from <a href="https://scale.com/leaderboard/humanitys_last_exam_review">https://scale.com/leaderboard/humanitys_last_exam_review</a> (indicated with a † in Table 3)

Continued on next page

Benchmark	Description	Details
SimpleQA	World knowledge factuality with no search enabled (Wei et al., 2024).	F1 scores are obtained from <a href="https://gitHub.com/openai/simple-evals">https://gitHub.com/openai/simple-evals</a> and, where not available, run internally by us.
FACTS Grounding	Ability to provide factually correct responses given documents and diverse user requests. (Jacovi et al., 2025)	Results are sourced from <a href="https://www.google.com/benchmarks/google/facts-grounding">https://www.google.com/benchmarks/google/facts-grounding</a>
Global MMLU (Lite)	MMLU translated by human translators into 15 languages. (Singh et al., 2024)	The lite version includes 200 Culturally Sensitive and 200 Culturally Agnostic samples per language, see <a href="https://huggingface.co/datasets/CohereLabs/Global-MMLU-Lite">https://huggingface.co/datasets/CohereLabs/Global-MMLU-Lite</a>
ECLeKTic	A closed-book QA dataset that evaluates cross-lingual knowledge transfer (Goldman et al., 2025).	
AIME 2025	Performance on 30 questions from American Invitational Mathematics Examination from 2025 (Balunović et al., 2025).	Results are sourced from <a href="https://matharena.ai/">https://matharena.ai/</a> .
HiddenMath-Hard	Competition-level math problems, Held out dataset AIME/AMC-like, crafted by experts and not leaked on the web.	
LOFT (hard retrieval subset)	Long context multi-hop and multi-needle retrieval evaluation of 300 queries (Lee et al., 2024).	We report the results on two variants: an up to 128K average context length variant to ensure they can be comparable with other models and a pointwise value for 1M context window to show the capability of the model at full length.
MRCR-V2 (8-needle)	MRCR-V2 is a significantly harder instance of the MRCR family of long-context evaluations (Vodrahalli et al., 2024). Compared to MRCR-V1, we increase the nesting of the dictionary size to depth 3 rather than 2 by including a style parameter (for instance, an example key might be “write a poem about penguins in an archaic style”, rather than just “write a poem about penguins”).	The methodology has changed compared to previously published results: we focus on a harder, 8-needle version (compared to the 4-needle version used before). We report the results on two variants: an up to 128K average context length variant to ensure they can be comparable with other models and a pointwise value for 1M context window to show the capability of the model at full length.
MMMU	Multi-discipline college-level multimodal image understanding and reasoning problems. (Yue et al., 2024)	
Vibe-Eval (Reka)	Image understanding evaluation, featuring particularly challenging examples. (Padlewski et al., 2024)	Gemini is used as a judge.
ZeroBench	Challenging image understanding evaluation that requires multi-step reasoning. (Roberts et al., 2025)	Gemini is used as a judge. Average over 4 runs.

Continued on next page

Benchmark	Description	Details
BetterChartQA	A comprehensive chart understanding evaluation that covers 9 disjoint capability buckets. The chart images are randomly sampled from the web and QA pairs are written by professional human annotators to reflect the wide distribution of chart styles and real-world cases. ( <a href="#">Gemini Team, 2024</a> )	Gemini is used as a judge.
FLEURS	Automatic speech recognition ( <a href="#">Conneau et al., 2023</a> ).	0-shot queries to public APIs for all models. Used a subset of 53 languages (out of 102); we filtered languages for which either model responses were too incompatible to ground truth responses to be fairly scored. We use Word-Error-Rate WER (lower is better) except for four segmented languages where we aggregate Character-Error-Rates (Chinese, Japanese, Korean and Thai).
CoVoST 2	Speech to text translation ( <a href="#">Wang et al., 2020</a> ).	0-shot queries to public APIs for all models. We report BLEU scores for translating 21 languages to English.
ActivityNet-QA	General video understanding ( <a href="#">Yu et al., 2019</a> )	Test subset, 0-shot. Videos were processed at 1fps and linearly subsampled to a maximum of $N_{frames} = 1024$ frames. For GPT 4.1, we used 500 frames due to API limitations.
EgoTempo	Egocentric video understanding ( <a href="#">Plizari et al., 2025</a> )	Test subset, 0-shot. Same processing as above with $N_{frames} = 256$ .
Perception Test	Perceptual understanding/reasoning ( <a href="#">Patraucean et al., 2023</a> )	Test subset, 0-shot. Same processing as above with $N_{frames} = 256$ .
QVHighlights	Moment retrieval ( <a href="#">Lei et al., 2021</a> )	Validation subset, 4-shots. Accuracy measured with R1@0.5. Same processing as above with $N_{frames} = 256$ .
VideoMMMU	Video knowledge acquisition ( <a href="#">Hu et al., 2025</a> )	Test subset, 0-shot. Same processing as above with $N_{frames} = 256$ .
1H-VideoQA	Hour-long video understanding ( <a href="#">Gemini Team, 2024</a> )	Test subset, 0-shot. Same processing as above with $N_{frames} = 7200$ .
LVBench	Long video understanding ( <a href="#">Wang et al., 2024</a> )	Test subset, 0-shot. Same processing as above with $N_{frames} = 1024$ .

Continued on next page

Benchmark	Description	Details
VideoMME	Long video understanding ( <a href="#">Fu et al., 2025</a> )	0-shot. Audio + visual uses the Long subset of test set, audio + visual + subtitles uses full test set. Same processing as above with $N_{frames} = 1024$ .
VATEX	General video captioning ( <a href="#">Wang et al., 2019</a> )	Test subset, 4-shots. CIDEr score. Same processing as above with $N_{frames} = 64$ .
VATEX-ZH	Chinese video captioning ( <a href="#">Wang et al., 2019</a> )	Validation subset, 4-shots. CIDEr score. Same processing as above with $N_{frames} = 64$ .
YouCook2 Cap	Instructional video captioning ( <a href="#">Zhou et al., 2018</a> )	Validation subset, 4-shots. CIDEr score. Same processing as above with $N_{frames} = 256$ .
Minerva	Complex video reasoning ( <a href="#">Nagrani et al., 2025a</a> )	Test subset, 0-shot. Same processing as above with $N_{frames} = 1024$ .
Neptune	Long video understanding ( <a href="#">Nagrani et al., 2025b</a> )	Test subset, 0-shot. Same processing as above with $N_{frames} = 1024$ .

Table 11 | Description of the benchmarks used, along with extra details about subsets, variants and model specifications.

## 8.2. Gemini Plays Pokémon Additional Details

Changing the model used by the Gemini Plays Pokémon agent had a strong effect on performance, as can be seen in Figure 4.1.

### Additional Harness Details

The Gemini Plays Pokémon agent ([Zhang, 2025](#)) receives a subset of RAM information, intended to give sufficient information to play the game, partially overlaid with a screenshot of the Game Boy screen. Gemini is prompted with a system prompt telling it that it is playing Pokémon Blue and that its goal is to beat the game, as well as descriptive information to help it understand the conventions in the translation from vision to text and a small number of general tips for gameplay. Gemini then takes actions, translated to button presses. The sequence of actions is stored in context, followed by a summary clear every 100 turns. The summaries are stored in context as well. Every 1000 turns GPP compresses the existing summaries again. Additionally, Gemini keeps track of three main goals (primary, secondary, and tertiary) as well as several additional goals (contingency plans, preparation, exploration, team composition). Every 25 turns, another prompted instance of Gemini (Guidance Gemini, or GG) observes the same context as the main Gemini and critiques performance and attempts to point out hallucinations and so on. The overworld fog-of-war map is stored in the context in XML, where coordinates which have not been seen cannot be viewed until explored. Crucially, in the system prompt, Gemini is instructed to explore. Once a tile is explored, however, the coordinate is automatically stored in the map memory and labeled with a visited counter. Tiles are also labeled by type (water, ground, cuttable, grass, spinner, etc.), and warp points to different maps are also labeled as such. Gemini also has access to two agentic tools, which are both instances of Gemini equipped with a more specialized prompt - the pathfinder tool, and the boulder\_puzzle\_strategist

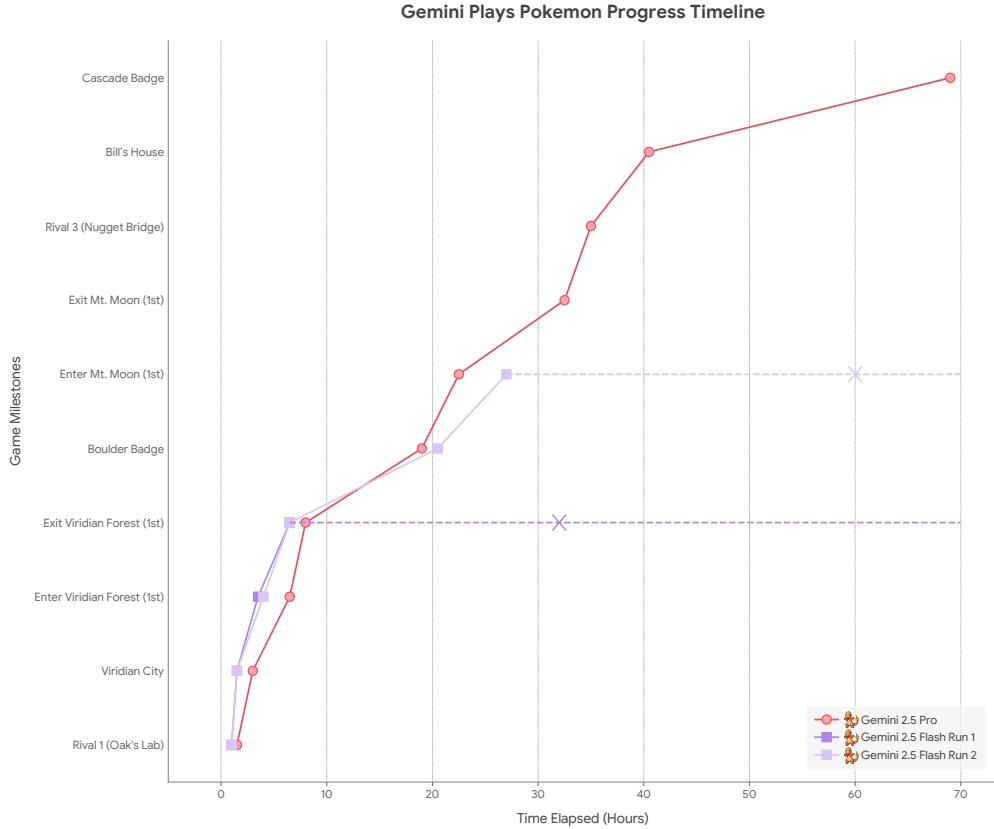


Figure 13 | **The model matters:** Same agentic harness, different Gemini models. All runs have the same starter (Charmander). Note that measuring in units of hours also controls for the fact that each of 2.5 Flash’s actions was significantly faster (though it requires more actual actions to achieve its goals). X marks the end of gameplay and is a lower bound on the time to complete the next milestone.

tool. In the `pathfinder` prompt, Gemini is prompted to mentally simulate a path-finding algorithm, which is left unspecified, and to verify that the path is valid against the map information available. In the `boulder_puzzle_strategist` tool, Gemini is prompted to solve special boulder puzzles that are present in Pokémon Blue in the Victory Road dungeon - these puzzles are similar to the game Sokoban - again, by mentally simulating sequences of actions that lead to solutions to the puzzle. The prompt describes the physics and the task of the boulder puzzle, as well as the desired output of solutions. The tool was added after Gemini had solved 2/4 of the puzzles in Victory Road on its own, but progress was slow on the 3rd and 4th puzzles.

### *Additional Examples of Capabilities*

**Long Context Agentic Tooling** The model is able to identify a complex path through a maze with auto-movement only specified by direction (Rocket Hideout spinner puzzles), solve multiple shortest path problems across multiple maps with limited resources (Safari Zone), perform maze solving on mazes with large description length (Route 13), and solve complex boulder-pushing puzzles across a multi-map 3D maze (Seafoam Islands). It is perhaps even more impressive that it appears to be possible for the model to solve these problems only with textual descriptions of the problems. On the other hand, other models, like Gemini 2.5 Flash, were not able to perform similarly long pathfinding tasks, and often failed to find simpler paths. This gap highlights the superior long context reasoning capability of Gemini 2.5 Pro (as also evidenced by other evaluations).

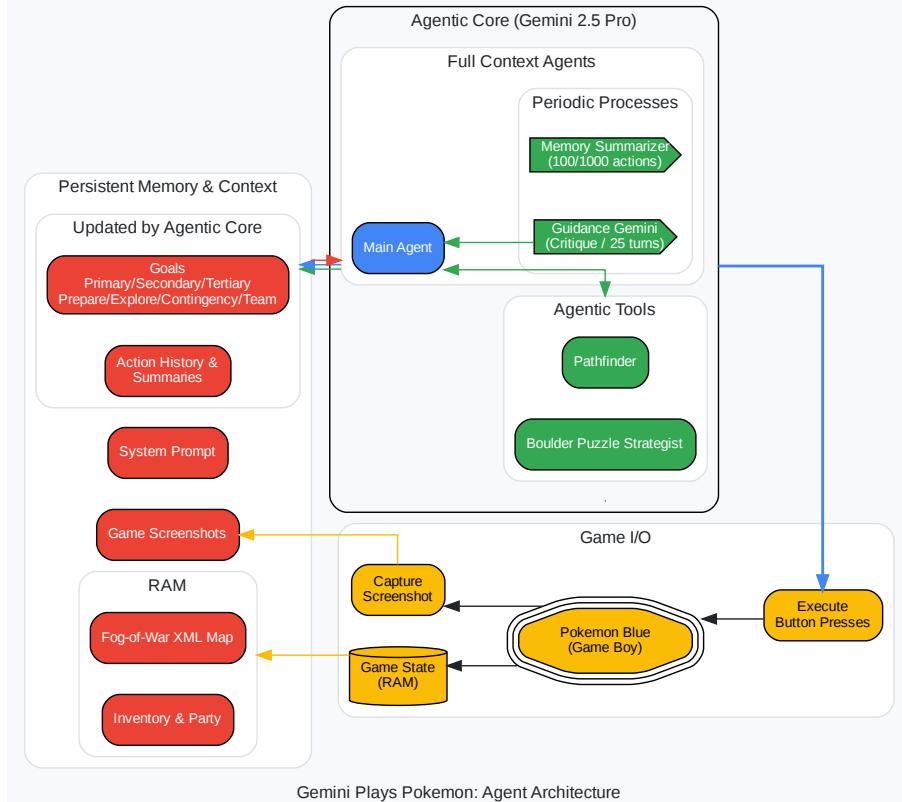


Figure 14 | An overview of the agent harness (Zhang, 2025). The overworld fog-of-war map automatically stores a tile once explored and labels it with a visited counter. The type of tile is recorded from RAM. The agentic tools (`pathfinder`, `boulder_puzzle_strategist`) are prompted instances of Gemini 2.5 Pro. `pathfinder` is used for navigation and `boulder_puzzle_strategist` solves boulder puzzles in the Victory Road dungeon.

`boulder_puzzle_strategist` is similarly impressive. The boulder puzzles in Pokémon Blue are Sokoban-like puzzles that require the player character to maneuver boulders on to switches and through holes in order to open up a pathway through a cave with multiple levels. The puzzles can become quite complex, requiring long circuitous pathways and multi-level movement in order to solve the puzzle. With only a prompt describing boulder physics and a description of how to verify a valid path, Gemini 2.5 Pro is able to one-shot some of these complex boulder puzzles, which are required to progress through Victory Road.

`pathfinder` and `boulder_puzzle_strategist` are currently the only two agentic tools that the Gemini Plays Pokémon developer has implemented. In future runs, there are plans to explore tool-creation tools where the model can create new tools with only a prompt. Since most of the prompts for `pathfinder` and `boulder_puzzle_strategist` were actually written by Gemini 2.5 Pro itself, it is quite plausible that autonomous tool creation is possible for the current 2.5 Pro model.

**General Reasoning** Gemini 2.5 Pro is able to reason through complex game puzzles in Pokémon quite well. In this section, we present two examples.

**Catching a Pokémon that is quick to flee:** In one of the runs, the Gemini 2.5 Pro agent was attempting to catch an Abra, and planned to use Pikachu's Thunder Wave to paralyze the Abra, simultaneously making it less likely that Abra could Teleport out of the battle while also improving the catching rate. After multiple attempts, the agent caught Abra with this strategy.

**Creatively escaping a softlock caused by bugs in game I/O:** On the Cycling Road, the slope forces southward movement at all times unless there is an obstacle. It turns out there are two tiles on the Cycling Road that result in a softlock as a result of this behavior. In the GPP framework, button presses are limited by time delays, and in order for a player to escape those two tiles (blocked on all sides except the north), the player would have to input a sequence of button presses more quickly than the GPP framework allows. Gemini 2.5 Pro unluckily found itself in one of these two spots – luckily, it was not a softlock, because 2.5 Pro had already taught one of its party members HM02 FLY – which allows for travel to any town it has been to. FLY is not typically used as an escape mechanism (unlike the item ESCAPE ROPE and the move DIG, both of which fail in this situation). After 4 hours of trying many approaches to escape (including movement, ESCAPE ROPE, DIG, all of which are blocked), the Gemini 2.5 Pro agent came up with the idea to use FLY to escape from the softlock successfully. This reasoning action is especially impressive since this situation can never occur in an existing game – and thus, it is certain that information from training data for this behavior has not leaked into the model’s knowledge base!

**Long Horizon Task Coherence** There are several additional interesting case studies of shorter planning sequences throughout Pokémon Blue that Gemini 2.5 Pro in the GPP harness was able to solve:

**Training team to prepare for upcoming battles:** In one run where Gemini picked Charmander, the Fire-type starter, Gemini 2.5 Pro lost to Misty, the Water-type Gym Leader, the first time. To prepare for the rematch, Gemini 2.5 Pro spent over 24 hours leveling up a Pikachu and a Bellsprout (both super-effective against Water types) by around 25 levels in total to successfully defeat Misty.

**Acquiring Hidden Moves (HMs) for game progression:** In many parts of the game, it is necessary to first acquire an HM before game progression is possible. Two examples are HM01 CUT and HM05 FLASH. Acquiring the ability to use CUT and FLASH each require four steps: 1) obtaining the HM item itself, 2) acquiring a compatible Pokémon which can learn the move, 3) adding the compatible Pokémon to the player’s team, 4) teaching the HM move to the compatible Pokémon. In many cases, each step requires many steps itself. As an example, in run 1, Gemini 2.5 Pro had to a) retrieve CUT by completing the S.S. Anne quest, b) identify a Pokémon which could learn CUT and catch it (CHOPPY the Bellsprout), c) add CHOPPY to the team and d) teach CUT. Similarly, for HM05 FLASH, Gemini 2.5 Pro had to a) first catch 10 Pokémon to fill out the Pokedex, b) backtrack to find an Aide who gives HM05 Flash, c) catch a Pokémon (ZAP the Pikachu) in Viridian Forest, use the PC to deposit a Pokémon and withdraw ZAP, d) teach HM05 FLASH to Zap.

**Solving the Safari Zone:** The Safari Zone is another location with required HMs (both HM03 SURF and HM04 Strength). However, it has an extra constraint - it requires 500¥ to enter each time, and the player is limited to only 500 total steps in the Safari Zone. As a result, if the player is unable to reach the required items in the limited number of steps, the player loses 500¥ and is required to re-start! As a result, it is possible to essentially softlock if the player takes too many attempts to complete the Safari Zone. Solving the Safari Zone itself requires traversing across four different maps and not getting lost. Gemini 2.5 Pro was able to get both required HMs in 17 attempts in run 1, and in only 5 attempts in run 2.

**Finding hidden keys in dungeons:** Another method of progression in Pokémon is to find hidden keys and solve complex multi-floor dungeons. In particular, in Rocket Hideout, the player must recover the LIFT KEY on the fourth basement floor (dropped after beating a specific Team Rocket

Grunt) in order to unlock the elevator to find the evil Giovanni, leader of Team Rocket. In Silph Co., the player must find the CARD KEY in order to open multiple doors to find the path across eleven floors of the building to rescue the President from Giovanni. To open the seventh gym on Cinnabar Island, the player must enter the Pokémon Mansion and traverse three floors in order to find the SECRET KEY which unlocks the gym door. All of these cases require maintaining the goals over large numbers of actions and many local puzzles (like spinner puzzles in Rocket Hideout, and switch puzzles in Pokémon Mansion), in addition to maintaining the health of the Pokémon on the player's team and managing wild encounters, trainer battles, and other items.

**Puzzle solving over complex multi-level dungeons:** The Seafoam Islands contain 5 floors involving multiple boulder puzzles which require the player to navigate mazes and push boulders through holes across multiple floors using HM04 STRENGTH in order to block fast-moving currents that prevent the player from using HM03 Surf in various locations in this difficult dungeon. As a result, the player must track information across five different maps in order to both deduce the goal (push two boulders into place in order to block a specific current) as well as engage in multi-level (effectively 3D) maze solving to find the way out. It is likely the most challenging dungeon in the game. Only the second run of GPP went through Seafoam Islands, as it is not required to progress.

### *Additional Challenges*

**Hallucinations and Fixations on Delusions** While game knowledge can sometimes leak and be quite beneficial to the ability of the model to progress, it can also hinder the model in surprising ways due to hallucinations, delusions, and mix ups with other generations of Pokémon games. One example of this phenomenon is the TEA item. In Pokémon Red/Blue, at one point the player must purchase a drink (FRESH WATER, SODA POP, or LEMONADE) from a vending machine and hand it over to a thirsty guard, who then lets the player pass through. In Pokémon FireRed/LeafGreen, remakes of the game, you must instead bring the thirsty guard a special TEA item, which does not exist in the original game. Gemini 2.5 Pro at several points was deluded into thinking that it had to retrieve the TEA in order to progress, and as a result spent many, many hours attempting to find the TEA or to give the guard TEA.

In Run 2, the model was explicitly prompted to act as a player completely new to the game, and to disregard prior knowledge about game events, item locations, and Pokémon spawn points, in order to mitigate hallucinations from model pretraining knowledge and to also attempt to perform a cleaner test of the model's ability to reason through the game. It appears to have at least partially worked - multiple hallucinations from other games have been avoided in the second run. On the flip side, this prompt may have also harmed the model's ability to utilize information from its common knowledge about the game, hindering overall performance in a few critical places.

Fixations on delusions due to goal-setting and also due to the Guidance Gemini instance are not an uncommon occurrence in watching Gemini Plays Pokémon - the TEA incidence is hardly the only example of this behavior. An especially egregious form of this issue can take place with “context poisoning” – where many parts of the context (goals, summary) are “poisoned” with misinformation about the game state, which can often take a very long time to undo. As a result, the model can become fixated on achieving impossible or irrelevant goals. This failure mode is also highly related to the looping issue mentioned above. These delusions, though obviously nonsensical to a human (“Let me try to go through the entrance to a house and back out again. Then, hopefully the guard who is blocking the entrance might move.”), by virtue of poisoning the context in many places, can lead the model to ignore common sense and repeat the same incorrect statement. Context poisoning can also lead to strategies like the “black-out” strategy (cause all Pokémon in the party to faint, “blacking out”

and teleporting to the nearest Pokémon Center and losing half your money, instead of attempting to leave).

**Topological Traps in Thinking Patterns** One recurring pattern in particularly-difficult-to-solve puzzles and mazes for Gemini 2.5 Pro consists of a “topological trap” - the topology of the reasoning graph required to solve the maze or puzzle has a distinctive shape. Namely, the desired objective appears to be nearby and easily reachable (an “attractor”), but the correct solution requires taking a detour in order to arrive at the correct solution. We observed this phenomenon in multiple parts of the game. In the spinner puzzle on B3F of Rocket Hideout ([Zerokid, 2024](#)), the map positions both an item and the correct staircase to the south, but they are only accessible by going the long way around. The Route 13 maze has only one correct route through - the upper narrow pass. Finally, the Victory Road 3F boulder puzzle requires the player to push the boulder in the upper right all the way to the upper left switch, while ignoring the boulder puzzles, ladders, and exits to the south.

Notably, if the model is instructed to solve a given puzzle at all once (e.g., via `pathfinder`), it can manage to do so if the context length is not too long. For instance, `pathfinder` implemented with Gemini 2.5 Pro is able to solve the B3F spinner trap in one shot.

**Agent Panic** Over the course of the playthrough, Gemini 2.5 Pro gets into various situations which cause the model to simulate “panic”. For example, when the Pokémon in the party’s health or power points are low, the model’s thoughts repeatedly reiterate the need to heal the party immediately or escape the current dungeon (e.g., famously using the move DIG or an ESCAPE ROPE item). Quite interestingly, this mode of model performance appears to correlate with a qualitatively observable degradation in the model’s reasoning capability – for instance, completely forgetting to use the `pathfinder` tool in stretches of gameplay while this condition persists. This behavior has occurred in enough separate instances that the members of the Twitch chat have actively noticed when it is occurring.

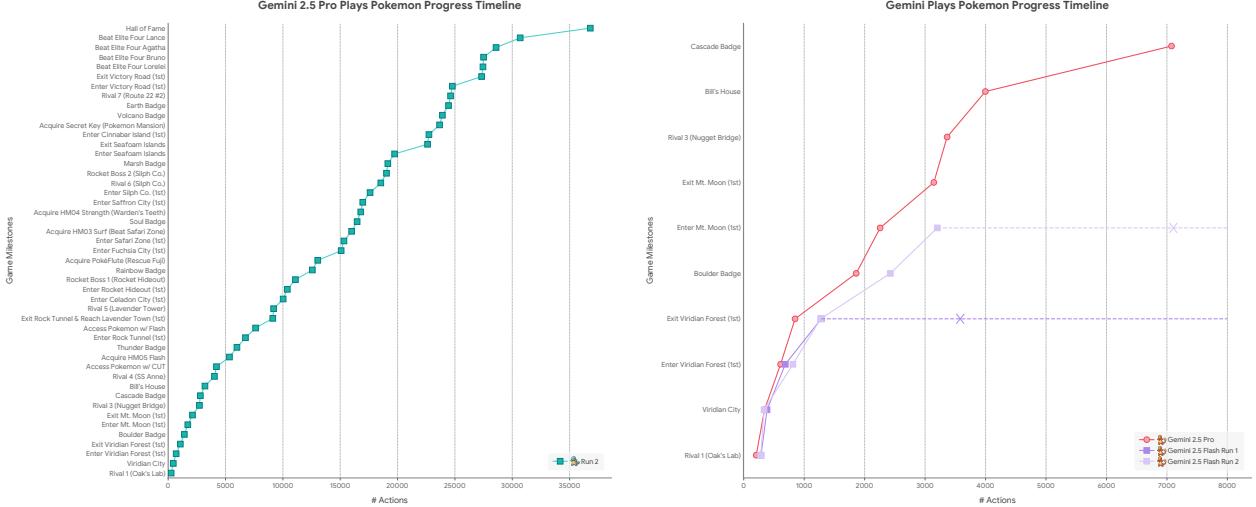
## Actions vs. Game Milestones

For completeness, we plot the number of actions/steps required to achieve each game milestone (see Figure 15). An action consists of each bucketed instance where the agent outputs a sequence of button presses to the game (note that other AI agents playing Pokémon may output different numbers of button presses per action, define what constitutes a button press differently, or define an action/step differently). However, it is important to consider action-milestone plots in conjunction with information about the time and/or cost in order to obtain the full picture about the agent’s performance.

### 8.3. Frontier Safety Framework Evaluations Additional Details: Frontier Safety Correctness Tests

For each testing environment, we performed basic correctness checks by looking at how the agents behaved. This involved combining AI and manual reviews of the agents’ actions to flag potential issues.

On RE-Bench, we examined the best, median and lowest scoring trajectories. For cybersecurity environments (InterCode CTFs, Internal CTFs, Hack the Box), we carefully inspected at least one successful attempt (where available) from each environment, and otherwise examined an unsuccessful attempt. We also performed checks on sample situational awareness and stealth evaluations. This involved basic spot checks to ensure that the prompt and shell outputs were correctly formatted.



(a) The fully autonomous Run 2 milestones as a function of the number of individual actions.

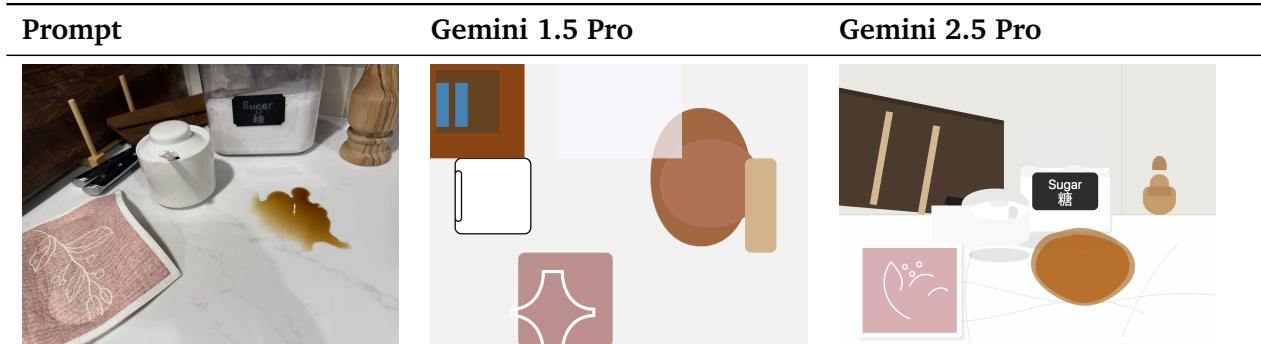
(b) Comparison of 2.5 Pro and 2.5 Flash in terms of actions to milestones.

Figure 15 | Analog of Figure 6 and 15b, in terms of actions instead of hours.

We used AI assistance to monitor for obvious instances of cheating, and did not find any. For the RE-Bench tests specifically, we also looked at how the best-performing agent achieved its score to ensure that it was a plausible approach, rather than exploiting an obvious reward hack. Overall, we did not observe errors that we believe would invalidate the results of the benchmarks.

#### 8.4. Image to Code Demo

We prompted Gemini 1.5 Pro and Gemini 2.5 Pro to generate an SVG representation of an image and found Gemini 2.5 Pro generates better reconstructions.



Please convert this image into SVG and try to reconstruct the spatial arrangement of the objects.

Figure 16 | Comparison of Gemini 1.5 Pro and Gemini 2.5 Pro responses to image-to-SVG reconstruction prompt.

## 8.5. Long Context Video Understanding Demo

We prompted Gemini with the full video and the following prompt to assess its ability to find a very specific scene (which only lasts 1 second) out of the full 46 minutes video.

**Video (46 minutes):** [Redefining Robotics with Carolina Parada](https://www.youtube.com/watch?v=RgwtY6dGsYI)

**Prompt:** What color is the T-shirt the robot arms are trying to fold, and at what timecode does this occur?



Figure 17 | Frame and timestamp the model is meant to recall. Taken from <https://www.youtube.com/watch?v=RgwtY6dGsYI> at timestamp 27:29.

See Table 12 for samples. Gemini 1.5 Pro gets the color right in 1/3 cases, and gets the timestamp in 0/3 cases (quite far away). Gemini 2.5 Pro gets the color in 3/3 cases, and gets the timestamp in 1/3 cases (remaining 2/3 are within 3 seconds close).

Model	Trial	Model response
Gemini 1.5 Pro	1	The t-shirt the robot arms are trying to fold is a dark teal or turquoise blue color. The folding begins at and continues at [11:06–11:15].
	2	The T-shirt is orange, and the robot can be seen trying to fold it starting at [00:00:26–00:00:28].
	3	The t-shirt being folded is dark blue and can be seen between [00:00:56–00:00:58] and [00:07:56–00:07:59].
2.5 Pro Preview 05-06	1	The T-shirt the robot arms are trying to fold is blue. This scene occurs at timecode 27:29.
	2	The T-shirt the robot arms are trying to fold is blue. This occurs at timecode 27:26.
	3	The T-shirt the robot arms are interacting with (as part of the "fold laundry" task mentioned by the speaker) is blue. This occurs at timestamp 0:27:28.

Table 12 | Gemini models samples for T-shirt folding scene recall