Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context

Gemini Team, Google[1]

In this report, we introduce the Gemini 1.5 family of models, representing the next generation of highly compute-efficient multimodal models capable of recalling and reasoning over fine-grained information from millions of tokens of context, including multiple long documents and hours of video and audio. The family includes two new models: (1) an updated Gemini 1.5 Pro, which exceeds the February version on the great majority of capabilities and benchmarks; (2) Gemini 1.5 Flash, a more lightweight variant designed for efficiency with minimal regression in quality. Gemini 1.5 models achieve near-perfect recall on long-context retrieval tasks across modalities, improve the state-of-the-art in long-document QA, long-video QA and long-context ASR, and match or surpass Gemini 1.0 Ultra's state-of-the-art performance across a broad set of benchmarks. Studying the limits of Gemini 1.5's long-context ability, we find continued improvement in next-token prediction and near-perfect retrieval (>99%) up to at least 10M tokens, a generational leap over existing models such as Claude 3.0 (200k) and GPT-4 Turbo (128k). Finally, we highlight real-world use cases, such as Gemini 1.5 collaborating with professionals on completing their tasks achieving 26 to 75% time savings across 10 different job categories, as well as surprising new capabilities of large language models at the frontier; when given a grammar manual for Kalamang, a language with fewer than 200 speakers worldwide, the model learns to translate English to Kalamang at a similar level to a person who learned from the same content.

1. Introduction

We present our latest multimodal models from the Gemini line: Gemini 1.5 Pro and Gemini 1.5 Flash. They are members of Gemini 1.5, a new family of highly-capable multimodal models which incorporates our latest innovations in sparse and dense scaling as well as major advances in training, distillation and serving infrastructure that allow it to push the boundary of efficiency, reasoning, planning, multi-linguality, function calling and long-context performance. Gemini 1.5 models are built to handle extremely long contexts; they have the ability to recall and reason over fine-grained information from up to at least 10M tokens. This scale is unprecedented among contemporary large language models (LLMs), and enables the processing of long-form mixed-modality inputs including entire collections of documents, multiple hours of video, and almost five days long of audio.

The Gemini 1.5 Pro presented in this report is an update over the previous Gemini 1.5 Pro February version and it outperforms it predecessor on most capabilities and benchmarks. All in all, the Gemini 1.5 series represents a generational leap in model performance and training efficiency. Gemini 1.5 Pro surpasses Gemini 1.0 Pro and 1.0 Ultra on a wide array of benchmarks while requiring significantly less compute to train. Similarly, Gemini 1.5 Flash performs uniformly better compared to 1.0 Pro and even performs at a similar level to 1.0 Ultra on several benchmarks.

The ability to model data of increasingly longer contexts has tracked the development of more general and capable language models, from the now toy 2-gram language model proposed by Shannon

[1] Please send correspondence to gemini-1_5-report@google.com.

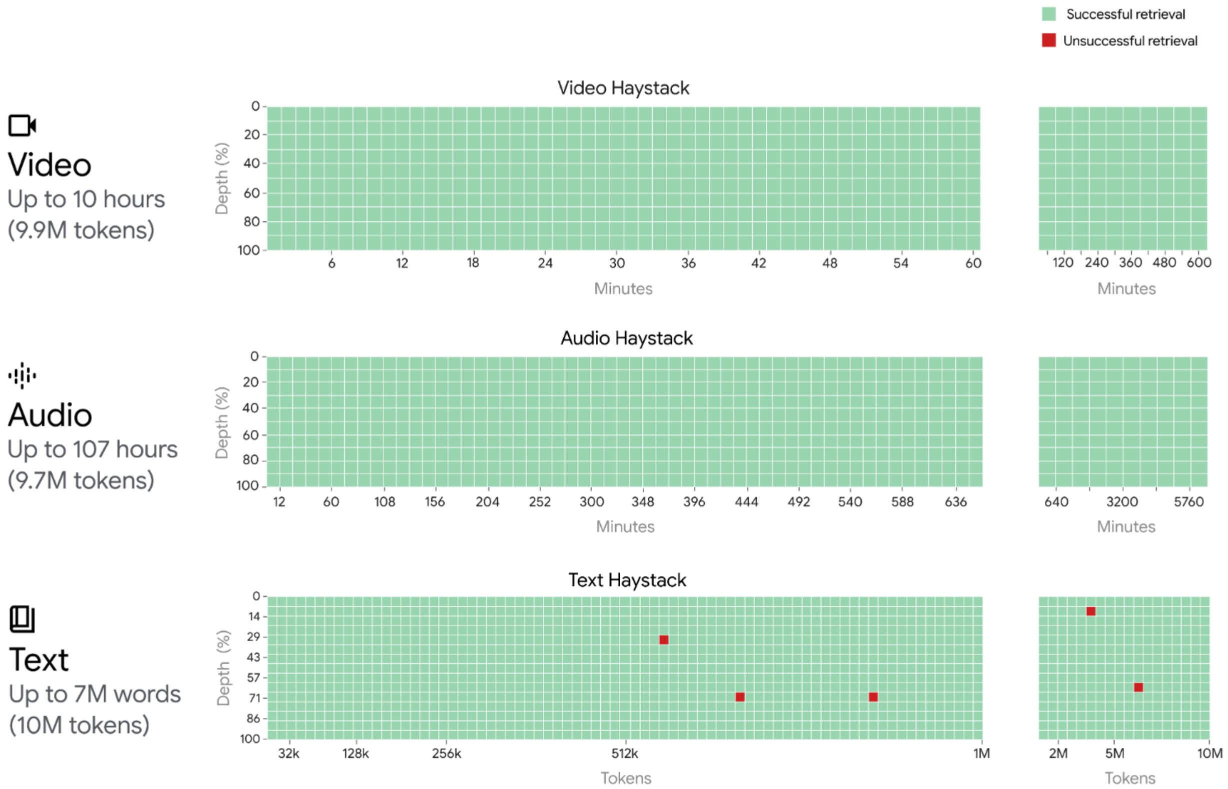

Figure 1 | Gemini 1.5 Pro achieves near-perfect "needle" recall (>99.7%) up to 1M tokens of "haystack" in all modalities, i.e., text, video and audio. It even maintains this recall performance when extending to 10M tokens in the text modality (approximately 7M words); 9.7M tokens in the audio modality (up to 107 hours); 9.9M tokens in the video modality (up to 10.5 hours). The x-axis represents the context window, and the y-axis the depth percentage of the needle placed for a given context length. The results are color-coded to indicate: green for successful retrievals and red for unsuccessful ones. Note that the performance for all modalities is obtained with the previously reported Gemini 1.5 Pro version from February.

(1948), to the modern n-gram models of the 1990s & 2000s typically constrained to 5 tokens of context (Brants et al., 2007; Chen and Goodman, 1999; Jelinek, 1998; Kneser and Ney, 1995), to recurrent neural networks language models from the 2010s which could effectively condition on hundreds of tokens (Jozefowicz et al., 2016; Mikolov et al., 2010), to the modern Transformer (Vaswani et al., 2017) which can condition on hundreds of thousands of tokens (Anthropic, 2023a). Gemini 1.5 Pro continues this trend by extending language model context lengths by over an order of magnitude. Scaling to millions of tokens, we find a continued improvement in predictive performance (Section 5.2.1.1), near perfect recall (>99%) on synthetic retrieval tasks (Figure 1 and Section 5.2.1.2), and a host of surprising new capabilities like in-context learning from entire long documents and multimodal content (Section 5.2.2).

To measure the effectiveness of our models' multimodal long-context capabilities, we conduct experiments on both synthetic and real-world tasks. In synthetic "needle-in-a-haystack" tasks inspired by Kamradt (2023) that probe how reliably the model can recall information amidst distractor context, we find that both Gemini 1.5 Pro and Gemini 1.5 Flash achieve near-perfect (>99%) "needle" recall up to multiple millions of tokens of "haystack" in all modalities, i.e., text, video and audio. As part of our experimental setup, we also assessed the performance of Gemini 1.5 Pro when extending