

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

Gemini Team, Google¹

In this report, we present the latest model of the Gemini family, Gemini 1.5 Pro, a highly compute-efficient multimodal mixture-of-experts model 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. Gemini 1.5 Pro achieves near-perfect recall on long-context retrieval tasks across modalities, improves the state-of-the-art in long-document QA, long-video QA and long-context ASR, and matches or surpasses Gemini 1.0 Ultra's state-of-the-art performance across a broad set of benchmarks. Studying the limits of Gemini 1.5 Pro'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 2.1 (200k) and GPT-4 Turbo (128k). Finally, we highlight 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 model from the Gemini line: Gemini 1.5 Pro. This is our first release from Gemini 1.5, a new family of highly-capable multimodal models which incorporates a novel mixture-of-experts architecture as well as major advances in training and serving infrastructure that allow it to push the boundary of efficiency, reasoning, and long-context performance. Gemini 1.5 Pro is built to handle extremely long contexts; it has 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. Gemini 1.5 Pro surpasses Gemini 1.0 Pro and performs at a similar level to 1.0 Ultra on a wide array of benchmarks while requiring significantly less compute to train.

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 \(1948\)](#), 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 4.2.1.1), near perfect recall (>99%) on synthetic retrieval tasks (Figure 1 and Section 4.2.1.2), and a host of surprising new capabilities like in-context learning from entire long documents (Section 4.2.2).

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

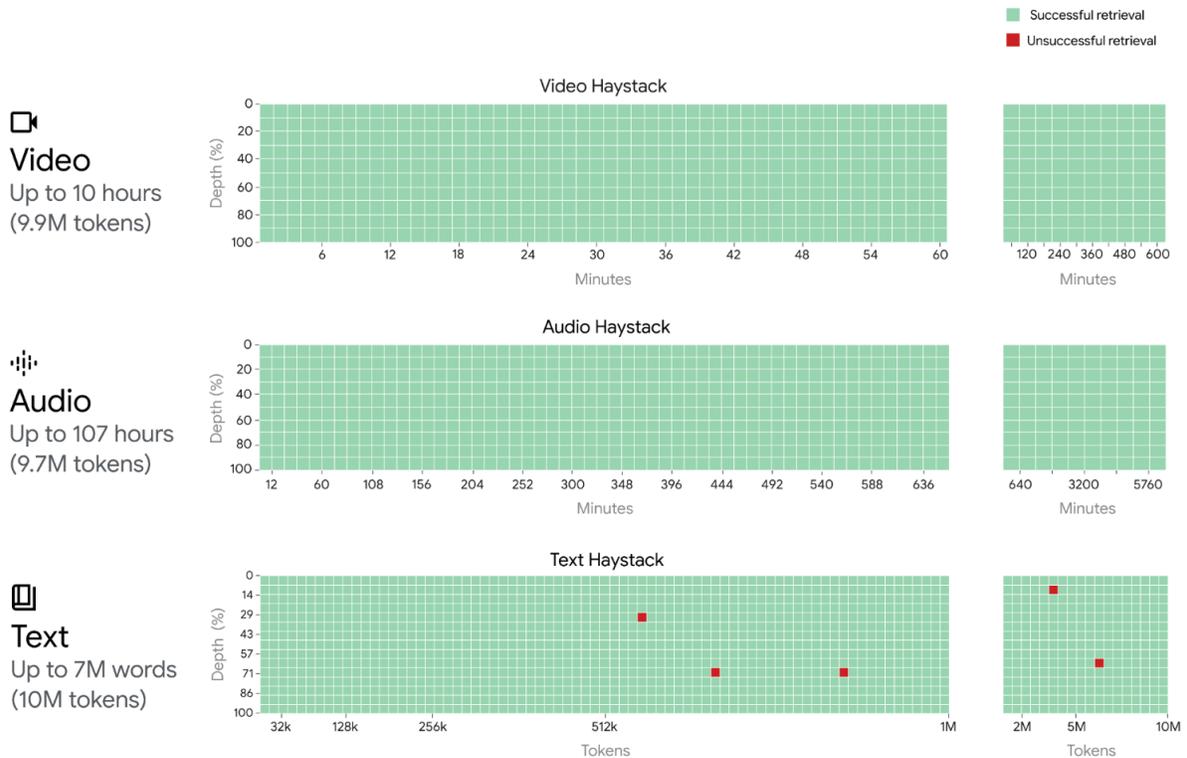


Figure 1 | Gemini 1.5 Pro achieves near-perfect “needle” recall (>99.7%) up to 1M tokens of “haystack” in all modalities, i.e., text, video and audio. It even maintains this recall performance when extending to 10M tokens in the text modality (approximately 7M words); 9.7M tokens in the audio modality (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.

To measure the effectiveness of our model’s 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 Gemini 1.5 Pro achieves near-perfect (>99%) “needle” recall up to multiple millions of tokens of “haystack” in all modalities, i.e., text, video and audio, and even maintaining this recall performance when extending to 10M tokens in the all three modalities. In more realistic multimodal long-context benchmarks which require retrieval *and* reasoning over multiple parts of the context (such as answering questions from long documents or long videos), we also see Gemini 1.5 Pro outperforming all competing models across all modalities even when these models are augmented with external retrieval methods. Finally, we qualitatively showcase the in-context learning abilities of Gemini 1.5 Pro enabled by very long context: for example, learning to translate a new language from a single set of linguistic documentation. With only instructional materials (a 500-page reference grammar, a dictionary, and ≈ 400 extra parallel sentences) all provided in context, Gemini 1.5 Pro is capable of learning to translate from English to Kalamang, a Papuan language with fewer than 200 speakers², and therefore almost no online presence. Moreover, we find that the quality of its translations is comparable to that of a person who learned from the same materials.

²Kalamang language: <https://endangeredlanguages.com/lang/1891>