SOVEREIGN: What is the impact of context length scaling from 1M to 2M tokens on the performance degradation of Gemini 1.5

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 29, 2026

Abstract

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 nea

1 Introduction

Analysis of: Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. Research goal: What is the impact of context length scaling from 1M to 2M tokens on the performance degradation of Gemini 1.5 models across different inference tasks?.

2 Methodology

Multi-query arXiv search (4 parallel queries, Relevance-sorted). TF-IDF cosine semantic verification (bigrams, threshold=0.15). NIM nv-embedqa-e5-v5 (dim=1024) for semantic indexing. Tribunal v2: 3-role parallel review (SKEPTIC/VALIDATOR/SYNTHESIZER) with revision round if score $<$ 6.5.

3 Results

11 papers retrieved. 9 claims extracted, 8 verified. Tribunal: 8.2/10 $\rightarrow$ APPROVE (revision_round=0). Policy: AUTO_APPROVE.

4 Uncertainties

NIM free tier latency varies. TF-IDF verification is a weak signal. arXiv Relevance ranking is query-dependent. Tribunal consensus is LLM-based and prompt-sensitive.

5 Extracted Claims

Claim	Verified	Confidence
Gemini 1.5 models can process millions of tokens of context	✓	0.19
The Gemini 1.5 family includes Gemini 1.5 Pro and Gemini 1.5 Flash models	✓	0.16
Gemini 1.5 Pro exceeds the February version on the great majority of capabilities and benchmarks	✓	0.23
Gemini 1.5 Flash is a more lightweight variant designed for efficiency with minimal regression in quality	✓	0.22
Gemini 1.5 models achieve near-perfect recall on long-context retrieval tasks	✓	0.30
Gemini 1.5 improves the state-of-the-art in long-document QA, long-video QA and long-context ASR	✓	0.30
Gemini 1.5 matches or surpasses Gemini 1.0 Ultra's performance across benchmarks	×	0.09
Gemini 1.5 achieves near-perfect retrieval (>99%) up to at least 10M tokens	✓	0.19
Gemini 1.5 achieves 26 to 75% time savings across 10 job categories when collaborating with professionals	✓	0.19

References

- https://doi.org/10.1017/dce.2020.16
- https://doi.org/10.48550/arxiv.2403.05530
- https://doi.org/10.48550/arxiv.2307.06435