SOVEREIGN: What is the impact of input modality interleaving (e.g., images and text) on GPT-4o's reasoning accuracy in AR

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 input modality interleaving (e.g., images and text) on GPT-4o's reasoning accuracy in ARC-Challenge when evaluated against text-only benchmarks?.

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

15 papers retrieved. 7 claims extracted, 7 verified. Tribunal: 8.3/10 $\rightarrow$ 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 achieve near-perfect recall on long-context retrieval tasks across modalities.	✓	0.32
Gemini 1.5 models improve the state-of-the-art in long-document QA, long-video QA and long-context ASR.	✓	0.33
Gemini 1.5 models match or surpass Gemini 1.0 Ultra's state-of-the-art performance across a broad set of benchmarks.	✓	0.24
Gemini 1.5 models show continued improvement in next-token prediction and near-perfect retrieval (>99%) up to at least 1	✓	0.27
Gemini 1.5 models represent a generational leap over existing models such as Claude 3.0 (200k) and GPT-4 Turbo (128k).	✓	0.21
Gemini 1.5 models achieve 26 to 75% time savings across 10 different job categories when collaborating with professional	✓	0.27
Gemini 1.5 models can learn a grammar manual for Kalamang, a language with fewer than 200 speakers worldwide.	✓	0.21

References

- https://doi.org/10.48550/arxiv.2403.05530
- https://doi.org/10.48550/arxiv.2303.12712
- https://doi.org/10.7551/mitpress/9780262033589.001.0001