Comparative Performance of Gemini 1.5 Flash and Pro in Long-Context Needle Retrieval

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 9 peer-reviewed papers addressing the following research question: What is the comparative performance gap in needle-in-a-haystack retrieval tasks between Gemini 1.5 Flash and Gemini 1.5 Pro when context length scales from 100k to 1M tokens with embedded distractor. 11 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: LongBench Pro: A More Realistic and Comprehensive Bilingual Long-Context Evaluation Benchmark. Research question: What is the comparative performance gap in needle-in-a-haystack retrieval tasks between Gemini 1.5 Flash and Gemini 1.5 Pro when context length scales from 100k to 1M tokens with embedded distractor sequences?.

2 Methodology

Systematic literature search across multiple databases yielded 9 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.2/10.

3 Results

9 papers retrieved. 11 claims extracted; 0 independently verified. Quality review score: 3.2/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
LongBench Pro includes 11 primary task categories.	×	0.08
LongBench Pro covers all core capability dimensions evaluated by existing benchmarks.	×	0.04
LongBench Pro introduces the context requirement dimension with two categories: Full and Partial.	×	0.09
LongBench Pro crosses 11 primary categories with context requirements to design 25 secondary categories.	×	0.10
LongBench Pro documents are collected from diverse domains and formats, including news, medicine, science, literature, l	×	0.02
LongBench Pro documents are balanced across single-document and multi-document settings, as well as English and Chinese.	×	0.06
LongBench Pro documents are assigned to six target length buckets: 8k, 16k, 32k, 64k, 128k, and 256k tokens.	×	0.07
LongBench Pro documents undergo a compliance review by human annotators to exclude privacy-sensitive, copyrighted, or no	×	0.02
LongBench Pro uses two types of prompts: non-thinking and thinking prompts.	×	0.04
LongBench Pro collects predictions from five advanced models for each sample during the answer review process.	×	0.03
LongBench Pro samples are reviewed by two annotators independently, with an additional long-context expert evaluating sa	×	0.10

References

- http://arxiv.org/abs/2601.02872v1
- http://arxiv.org/abs/2403.05530v5
- http://arxiv.org/abs/2406.11230v2