Benchmark Performance of Gemini3-Pro-Preview Across Reasoning, Mathematics, Coding, and Language Tasks

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 4 peer-reviewed papers addressing the following research question: What are the benchmark performance scores of Gemini3-Pro-Preview on reasoning mathematics coding and language understanding tasks. 10 claims were extracted from source literature; 10 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 8.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: CutVerse: A Compositional GUI Agents Benchmark for Media Post-Production Editing. Research question: What are the benchmark performance scores of Gemini3-Pro-Preview on reasoning mathematics coding and language understanding tasks.

2 Methodology

Systematic literature search across multiple databases yielded 4 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 8.2/10.

3 Results

4 papers retrieved. 10 claims extracted; 10 independently verified. Quality review score: 8.2/10.

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
GUI agents have made significant progress in web navigation and basic operating system tasks.	✓	0.27
The capabilities of GUI agents in professional creative workflows remain largely underexplored.	✓	0.25
Cutverse is a benchmark designed to systematically evaluate autonomous GUI agents in realistic media post-production env	✓	0.41
Cutverse curates expert demonstrations across 7 professional applications (e.g., Premiere Pro, Photoshop).	✓	0.20
Cutverse covers 186 complex, long-horizon tasks grounded in authentic editing workflows.	✓	0.30
Cutverse involves dense multimodal interfaces and tightly coupled interaction sequences.	✓	0.20
Cutverse supports scalable evaluation through a lightweight parser that transforms raw screen recordings and low-level i	✓	0.36
Existing agents achieve only 36.0% task success on realistic media editing tasks in Cutverse.	✓	0.30
Current models demonstrate promising spatial grounding, multimodal alignment, and coordinated action execution.	✓	0.28
Current models remain limited in long-horizon reliability and domain-specific planning.	✓	0.27

References

- https://openalex.org/W7154539175
- https://openalex.org/W7125352984

- https://openalex.org/W7162044470