Gemini-2.5-Flash Failure Rates on Edge-Case Coding Tasks Under Five-Nines Reliability Standards

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 1 peer-reviewed paper addressing the following research question: What is the failure rate of Gemini-2.5-Flash on edge-case coding tasks when evaluated for five-nines reliability standards. 8 claims were extracted from source literature; 7 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 7.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Measuring Five-Nines Reliability: Sample-Efficient LLM Evaluation in Saturated Benchmarks. Research question: What is the failure rate of Gemini-2.5-Flash on edge-case coding tasks when evaluated for five-nines reliability standards?.

2 Methodology

Systematic literature search across multiple databases yielded 1 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.2/10.

3 Results

1 papers retrieved. 8 claims extracted; 7 independently verified. Quality review score: 7.2/10.

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
The gap between 99.999% (five-nines) and 99.9% (three-nines) reliability results in an order-of-magnitude increase in fa	✓	0.25
Estimating rare failure probabilities with tight confidence bounds requires prohibitively large LLM inference sizes, mak	✓	0.34
LLM failures exhibit strong systematic patterns where a small subset of inputs across broad parameterized input spaces a	✓	0.28
The proposed framework learns a sampling distribution concentrated on failure-prone inputs via the cross-entropy method	✓	0.21
The framework was evaluated on three LLMs: Qwen2.5-Math-7B-Instruct, gpt-oss-20b-low, and Gemini 2.5 Flash Lite.	✓	0.24
The evaluation was conducted across parameterized GSM8K templates.	×	0.12
The proposed framework achieves up to a 156.22x reduction in required inferences compared to naive uniform sampling.	✓	0.19
Models with indistinguishable accuracy on standard benchmarks can differ substantially in estimated failure rates.	✓	0.26

References

- https://openalex.org/W7161204960