

Qwen2.5 Post-Training Strategy and Instruction-Following Benchmark Performance

Assignee Research

May 30, 2026

Abstract

This report synthesises findings from 11 peer-reviewed papers addressing the following research question: Does the improved post-training strategy in Qwen2.5 yield higher alignment scores on instruction-following benchmarks compared to models trained with equivalent data but earlier alignment techniques. Despite widespread deployment of Large Language Models, systematic evaluation of instruction-following capabilities remains challenging. While comprehensive benchmarks exist, focused assessments that quickly diagnose specific instruction adherence patterns are valuable. 20 claims were extracted from source literature; 2 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.8/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: When Models Can't Follow: Testing Instruction Adherence Across 256 LLMs. Research question: Does the improved post-training strategy in Qwen2.5 yield higher alignment scores on instruction-following benchmarks compared to models trained with equivalent data but earlier alignment techniques?.

2 Methodology

Systematic literature search across multiple databases yielded 11 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.8/10.

3 Results

11 papers retrieved. 20 claims extracted; 2 independently verified. Quality review score: 3.8/10.

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 evaluation framework was implemented using the OpenRouter API, which provides unified access to multiple language models	×	0.06
The evaluation system executes each test systematically across all target models, maintaining consistent parameters to ensure reproducibility	×	0.04
Temperature was set to 0.0 to minimize response variability and enable more deterministic evaluation of instruction-following capabilities	×	0.07
Response timeout limits were established at 10 seconds to balance comprehensive model coverage with practical execution	×	0.04
Primary metrics include binary pass/fail determination based on adherence to specified instructions, response time measurement, and token usage	×	0.03
Results are automatically compiled into a structured Excel workbook with multiple analytical views.	×	0.01
The overview sheet provides a high-level pass/fail matrix across all models and tests, enabling rapid identification of trends and issues	×	0.03
Individual test sheets contain detailed response data, allowing deeper investigation of specific failure modes.	×	0.03
A model summary sheet aggregates performance statistics, including success rates, average response times, and total tokens used	×	0.02
Each test prompt includes programmatically verifiable success criteria.	×	0.03
The verification approach accounts for common variations in model outputs while maintaining objective assessment standards	×	0.02
The verification process operates in two stages: primary verification applies strict matching criteria to determine exact matches, while secondary verification uses fuzzy matching for partial success	×	0.01
The comprehensive evaluation of instruction-following capabilities encompassed all 331 models available via OpenRouter API	✓	0.17
331 models were verified for basic functionality, of which 256 passed verification and were subsequently evaluated using more complex prompts	×	0.10
The two-stage verification protocol first assessed basic endpoint functionality and subsequently evaluated instruction-following capabilities	×	0.08
The evaluation of instruction-following capabilities was conducted on October 14, 2025.	✓	0.20
The evaluation involved querying the OpenRouter API to retrieve all available models, yielding a total of 331 models across various categories	×	0.06
The verification step involved querying each model with a standardized set of prompts to assess basic functionality	×	0.03

References

- <http://arxiv.org/abs/2405.19874v3>
- <http://arxiv.org/abs/2308.16149v2>
- <http://arxiv.org/abs/2510.18892v1>