Qwen3 Thinking and Non-Thinking Modes Performance on HumanEval Pro and MBPP Pro

Assignee Research

June 3, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: How does the integration of thinking and non-thinking modes in Qwen3 affect its performance on HumanEval Pro and MBPP Pro benchmarks, as measured by pass@k accuracy and latency trade-offs compared to. 0 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 5.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: HumanEval Pro and MBPP Pro: Evaluating Large Language Models on Self-invoking Code Generation. Research question: How does the integration of thinking and non-thinking modes in Qwen3 affect its performance on HumanEval Pro and MBPP Pro benchmarks, as measured by pass@k accuracy and latency trade-offs compared to other state-of-the-art dense and MoE models?.

2 Methodology

Systematic literature search across multiple databases yielded 12 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 5.5/10.

3 Results

12 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 5.5/10.

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.

References

- http://arxiv.org/abs/2412.21199v2
- http://arxiv.org/abs/2509.17765v1
- http://arxiv.org/abs/2505.14640v1