Thinking Mode in Qwen3 Alters Latency-Accuracy Trade-offs on MATH Benchmark

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 12 peer-reviewed papers addressing the following research question: How does the introduction of thinking mode in Qwen3 affect latency and accuracy trade-offs on the MATH benchmark compared to non-thinking modes. 11 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Qwen3-Omni Technical Report. Research question: How does the introduction of thinking mode in Qwen3 affect latency and accuracy trade-offs on the MATH benchmark compared to non-thinking modes?.

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 3.5/10.

3 Results

12 papers retrieved. 11 claims extracted; 0 independently verified. Quality review score: 3.5/10.

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.

5 Extracted Claims

Claim	Verified	Confidence
Qwen3-Omni-30B-A3B-Instruct, Qwen3-Omni-30B-A3B-Thinking, Qwen3-Omni-Flash-Instruct, and Qwen3-Omni-Flash-Thinking were	×	0.05
Qwen3-Omni's evaluation results are divided into two main categories: understanding (X$\rightarrow$Text) and speech generation (X$\rightarrow$Sp	×	0.10
Qwen3-Omni was evaluated on text$\rightarrow$text tasks using MMLU-Redux, GPQA, AIME25, ZebraLogic, MultiPL-E, IFEval, Creative Writ	×	0.03
Qwen3-Omni was evaluated on audio$\rightarrow$text tasks using RUL-MuchoMusic, MMAU, MMSU, VoiceBench, GTZAN, MTG-Jamendo, and Magna	×	0.05
Qwen3-Omni was evaluated on vision$\rightarrow$text tasks using MMStar, HallusionBench, MM-MT-Bench, MathVista, MathVision, MMMU, an	×	0.04
Qwen3-Omni employs Thinker-Talker architecture.	×	0.09
Qwen3-Omni introduces Mixture-of-Experts (MoE) architectures for both the Thinker and Talker to support high concurrency	×	0.05
The Talker in Qwen3-Omni no longer consumes the Thinker's high-level text representations and conditions only on audio a	×	0.09
The Thinker and Talker in Qwen3-Omni can use distinct system prompts, independently controlling the Thinker's response s	×	0.06
The Talker in Qwen3-Omni adopts a multi-codebook autoregressive scheme.	×	0.12
The Code2Wav in Qwen3-Omni is implemented as a lightweight causal ConvNet.	×	0.09

References

- http://arxiv.org/abs/2601.21337v2

- http://arxiv.org/abs/2509.17765v1
- http://arxiv.org/abs/2103.03874v2