GLM-5 vs Qwen 3.5 vs Gemma 4

Coding Agent Performance Benchmark

Models Compared

- Qwen 3.5 (397B)
- Gemma 4 (31B)
- GLM-5

April 09, 2026

AI Model Benchmarking Analysis

Executive Summary

This comprehensive benchmark evaluates three state-of-the-art cloud-hosted language models across six critical software development tasks: code generation, bug fixing, code review, refactoring, test writing, and architecture design. Our analysis reveals significant performance differences in both speed and output quality.

Key Findings

Qwen 3.5 is the Clear Winner: 256.4 seconds total, providing the best balance of speed

Gemma 4 Offers Speed: 439.9 seconds total with strong code generation (19.7s), but st

GLM-5 is Thorough but Slow: 808.9 seconds (3.1x slower than Qwen) with verbose, con

Performance Summary

Metric	Qwen 3.5	Gemma 4	GLM-5	Winner
Total Time (seconds)	256.4	439.9	808.9	Qwen
Avg Test Time (seconds)	42.7	73.3	134.8	Qwen
Code Generation (s)	47.2	19.7	131.8	Gemma
Bug Fixing (s)	38.9	79.2	125.4	Qwen
Code Review (s)	42.1	68.4	138.2	Qwen
Refactoring (s)	43.5	72.1	142.3	Qwen
Test Writing (s)	41.8	74.5	129.8	Qwen
Architecture Design (s)	42.9	126.0	141.4	Qwen

Detailed Analysis by Task

Code Generation

Generating a complete feature implementation from requirements

Fastest: Gemma 4 (19.7s)

Best Quality: Qwen 3.5

Gemma 4 excels at rapid code generation, producing concise and functional code. Qwen offers better quality with slightly longer output. GLM-5 is overly verbose but thorough.

Bug Fixing

Identifying and resolving bugs in provided code

Fastest: Qwen 3.5 (38.9s)

Best Quality: Qwen 3.5

Qwen provides targeted, efficient bug fixes. Gemma struggles with complex debugging scenarios. GLM-5 offers exhaustive analysis but at 3.2x Qwen's cost.

Code Review

Thorough review and optimization recommendations for code

Fastest: Qwen 3.5 (42.1s)

Best Quality: Qwen 3.5

Qwen delivers balanced reviews with actionable insights. Gemma offers surface-level observations. GLM-5 is comprehensive but unnecessarily lengthy.

Refactoring

Improving code structure while maintaining functionality

Fastest: Qwen 3.5 (43.5s)

Best Quality: Qwen 3.5

Qwen produces clean, maintainable refactored code. Gemma's suggestions are less comprehensive. GLM-5 provides excessive detail.

Test Writing

Creating comprehensive unit test suites for code

Fastest: Qwen 3.5 (41.8s)

Best Quality: Qwen 3.5

Qwen generates well-structured, complete test suites. Gemma lacks edge case coverage. GLM-5 is overly exhaustive.

Architecture Design

Designing system architecture for complex requirements

Fastest: Qwen 3.5 (42.9s)

Best Quality: Qwen 3.5

Qwen provides balanced, implementable architectures. Gemma struggles with complex systems. GLM-5 produces lengthy, hard to follow designs.

Recommendations

Primary Model: Qwen 3.5

For production coding agent deployments, Qwen 3.5 is the recommended choice. It delivers the best balance of performance, quality, and cost-effectiveness across all development tasks.

Secondary Model: Gemma 4

Gemma 4 is suitable for scenarios where latency is critical, particularly for code generation tasks. Consider using it as a fallback model or for real-time features.

Analytics Model: GLM-5

GLM-5 should be reserved for offline comprehensive analysis tasks where speed is not a constraint and you benefit from its thorough, verbose approach to understanding code complexity.