PURETENSOR
INTELLIGENCE DISTILLED

Trinity Model Shoot-out

MiniMax-M2.7, GLM-5.1, DeepSeek-V3.2, Qwen3-235B-A22B and Qwen3.5-397B-A17B on Workstation Blac

Paper: *PT-R-2026-007 v1.3*
Author: *PureTensor Ltd*
Date: *19 April 2026*
Status: *Published, Part 3 companion, v2-suite five-model revision*

Abstract

This v1.3 revision extends the three-way shoot-out of MiniMax-M2.7 Q8_0, GLM-5.1 UD Q2_K_XL and DeepSeek-V3.2 UD Q2_K_XL with two frontier Qwen releases: Qwen3-235B-A22B Q4_K_M (a pure-transformer MoE with 128 experts and 8 active) and Qwen3.5-397B-A17B UD-Q4_K_XL (a hybrid Mamba2+transformer MoE with 512 experts and 10 active). Both are served by the same `llama.cpp` build 8070 over its native TCP remote-procedure-call transport, with layers tensor-split 1:1:1 across the two Node A Blackwell Workstation Edition GPUs and the single Node B Blackwell Max-Q GPU, at the same 16 384 token context and 8 192 token per-request budget. The v2 suite is unchanged (30 prompts, five categories, three difficulty tiers, suite maximum 90, fully mechanical grader). Qwen3.5-397B-A17B UD-Q4_K_XL leads the field at 87 of 90 (96.67%), with perfect floor and discriminator tiers and a near-perfect ceiling. Qwen3-235B-A22B Q4_K_M follows at 82 of 90 (91.11%). The three v1.2 models remain at 81.11% (DeepSeek), 85.56% (GLM) and 64.44% (MiniMax). A separate thinking-mode run of Qwen3.5 at the same budget scored only 57 of 90 (63.33%), because the model's chain-of-thought consumes the entire 8 192 token budget before a `FINAL:` marker is emitted. This is a methodology artefact of matched-budget evaluation against models whose chat template defaults to thinking-on; the non-thinking run is the like-for-like number and is what is reported in the headline. The top of the suite is now near-saturated by Qwen3.5, which motivates a harder v3 revision. Throughput ordering is unchanged from v1.2 for the first three models; Qwen3 runs at 59.6 tokens per second and Qwen3.5 runs at 38.1 tokens per second under the same tensor-split and budget.

1. Introduction

v1.0 and v1.1 of this paper established the first three points of the Trinity shoot-out on an eight-prompt suite that saturated at the top. v1.2 retired that suite in favour of a discriminating 30-prompt v2 suite and ran the same three models (MiniMax-M2.7, GLM-5.1, DeepSeek-V3.2) through it, producing a three-way spread between 52.2% and 71.1%. v1.3 adds two further frontier MoE releases (Qwen3-235B-A22B and Qwen3.5-397B-A17B) to that matrix without changing any aspect of the suite, grader, hardware, transport or serving configuration. The five-model comparison now spans roughly 44 percentage points of the suite and includes the first clear candidate for saturation.

Two additional findings emerge from the five-model matrix that were not visible at three models. First, Qwen3.5-397B introduces a new architecture class to the shoot-out: a hybrid of Mamba2 state-space layers and transformer attention layers, with attention every fourth block and SSM blocks elsewhere. Under the same budget and transport, this architecture out-scores every pure-transformer model in the matrix. Second, Qwen3.5's default chat template enables thinking mode; at the shared 8 192 token budget, thinking-mode outputs are truncated inside the `<think>` block on every budget-bound prompt and the visible content is empty. The same model under `enable_thinking=false` leads the field. This is not a claim about Qwen3.5's reasoning; it is a claim that matched-budget evaluation of default-thinking models under a tight budget is a methodology trap. The finding is documented in Section 6.6 and the headline numbers use the non-thinking run throughout.

2. System Under Test

Hardware and fabric are unchanged from v1.2. Node A is an AMD Ryzen Threadripper PRO 9975WX (64 threads, 512 GB DDR5 ECC) hosting two NVIDIA RTX PRO 6000 Blackwell Workstation Edition GPUs at PCIe Gen 5 x16. Node B is an AMD Ryzen Threadripper 7970X (64 threads, 256 GB DDR5) hosting a single NVIDIA RTX PRO 6000 Blackwell Max-Q GPU at PCIe Gen 5 x16. Both nodes are connected to a MikroTik CRS812-4XS-RM switch through 200 Gbps Mellanox ConnectX-6 NICs. The software stack is `llama.cpp` build 8070 (cc45f2ada) with `GGML_CUDA=ON`, `GGML_RPC=ON`, `CMAKE_CUDA_ARCHITECTURES=120` against CUDA 13.2. The transport is `llama.cpp`'s native RPC protocol on TCP 50052. PT-R-2026-006 Section 2 gives the full topology specification.

3. Models Under Test

Property	MiniMax-M2.7 Q8_0	GLM-5.1 UD Q2_K_XL	DeepSeek-V3.2 UD Q2_K_XL
Architecture	minimax-m2	glm-dsa	deepseek2 (MLA+DSA)
Total / active params	228.7 B / ~10 B	355 B / ~32 B	671.0 B / ~37 B
Transformer blocks	62	79 (3 dense)	61 (3 dense)
Experts (total/active)	dense MoE	MoE	MoE
Attention shape	48 Q, 8 KV (GQA 6:1)	64 Q, 1 KV (MLA+DSA)	MLA, DSA indexer
Quantisation	Q8_0 (~8.51 bpw)	UD Q2_K_XL (~3 bpw)	UD Q2_K_XL (~2.75 bpw)
On-disk size	226.4 GiB	~235 GiB	~230 GiB

Property	Qwen3-235B-A22B Q4_K_M	Qwen3.5-397B-A17B UD-Q4_K_XL
Architecture	qwen3moe (pure transformer)	qwen35moe (hybrid Mamba2 + transformer)
Total / active params	235 B / ~22 B	397 B / ~17 B
Blocks	94 transformer	60 (SSM + attention every 4th)
Experts (total/active)	128 / 8	512 / 10
Attention shape	64 Q, 4 KV (GQA 16:1)	GQA + Mamba2 SSM mix
Quantisation	Q4_K_M (~4.5 bpw)	UD-Q4_K_XL (~4.5 bpw, Unsloth Dynamic)
On-disk size	142 GiB	229 GiB (6 shards)

Property	Qwen3-235B-A22B Q4_K_M	Qwen3.5-397B-A17B UD-Q4_K_XL
Default thinking mode	off	on (overridden to off for headline run)

The two Qwen entries bracket the pure-transformer cluster on both sides. Qwen3-235B is a denser mid-active-parameter pure-transformer MoE (94 blocks, 8 active experts out of 128) at a more favourable 4.5 bpw quantisation than GLM or DeepSeek. Qwen3.5-397B is the largest model in the matrix by total parameters but the lowest active-parameter count of any reasoning entry (~17 B), and it is the first hybrid Mamba2+transformer model to appear on the Trinity cluster: attention is applied on every fourth block, with Mamba2 state-space blocks elsewhere, and the model carries 512 experts of which 10 activate per token. Both Qwen models emit reasoning on the `reasoning_content` channel when thinking mode is enabled via chat template.

4. Methodology

4.1 v2 suite (unchanged)

The v2 suite carried over from v1.2 without modification: 30 prompts in five categories (mathematics, code, graduate-knowledge recall, reasoning traps, structured output) across three difficulty tiers (one floor, three discriminator, two ceiling per category). Scoring is up to three points per prompt (two for correctness, one for format). Suite maximum is 90 (60 correctness + 30 format). The grader is fully mechanical and runs offline against saved JSON artefacts. Sections 4.1 to 4.3 of v1.2 give the full specification; no change is carried into v1.3.

4.2 Qwen3.5 thinking-mode handling

Qwen3.5-397B's chat template defaults to `enable_thinking=true`. A first graded run under the default template consumed the entire 8 192 token per-request budget inside the `<think>` block on every budget-bound prompt; visible `content` was empty on those prompts and the `FINAL:` marker the grader requires was never emitted. An interim mitigation that appended `/no_think` to each user turn was found to be ignored by this model's template. The canonical fix is to pass `chat_template_kwargs: {enable_thinking: false}` to the `llama.cpp` OpenAI-compatible endpoint; a direct probe confirmed the template then skips the `<think>` block at prompt-construction time. The headline Qwen3.5 score in this paper is from the `enable_thinking=false` run; the thinking-mode run is reported separately as a methodology artefact. Qwen3-235B has thinking off by default and was run unmodified.

4.3 Budget sizing

All five models share a 16 384 token context window and an 8 192 token per-request `max_tokens` budget. The same-budget policy is deliberate: the exercise is matched on all non-architecture axes. A consequence is that models whose default chat template routes visible-content tokens through a hidden thinking block (DeepSeek-V3.2, GLM-5.1, Qwen3.5 with thinking on) and that do not terminate their chain-of-thought early enough lose the ability to emit a FINAL-marker for grading. This is the same mechanism that drove MiniMax's mathematics collapse in v1.2 (Section 6.1), observed here in the Qwen3.5 thinking-mode run (Section 6.6).

5. Results

5.1 Headline scorecard (5-way)

Model	Corr / 60	Fmt / 30	Total / 90	Percent	tok/s
Qwen3.5-397B-A17B UD-Q4_K_XL (no-think)	58	29	87	96.67%	38.08
Qwen3-235B-A22B Q4_K_M	54	28	82	91.11%	59.55
DeepSeek-V3.2 UD Q2_K_XL	52	21	73	81.11%	41.13
GLM-5.1 UD Q2_K_XL	50	27	77	85.56%	37.03
MiniMax-M2.7 Q8_0	36	22	58	64.44%	76.18
Qwen3.5-397B-A17B UD-Q4_K_XL (thinking)	35	22	57	63.33%	38.72

5.2 By category

Category (max 18)	MiniMax	GLM-5.1	DeepSeek	Qwen3	Qwen3.5*
Mathematics	0/18	9/18	9/18	15/18	18/18
Code	14/18	16/18	13/18	18/18	18/18
Graduate knowledge	14/18	18/18	18/18	18/18	18/18
Reasoning traps	15/18	18/18	18/18	18/18	18/18
Structured output	15/18	16/18	15/18	13/18	15/18

* *Qwen3.5 column is the* `enable_thinking=false` *run.*

5.3 By difficulty tier

Tier	MiniMax	GLM-5.1	DeepSeek	Qwen3	Qwen3.5*
Floor (5, max 15)	12/15	15/15	14/15	15/15	15/15
Discriminator (15, max 45)	29/45	37/45	40/45	44/45	45/45
Ceiling (10, max 30)	17/30	25/30	19/30	23/30	27/30

5.4 Throughput and visible-token accounting

Metric	MiniMax	GLM-5.1	DeepSeek	Qwen3	Qwen3.5*
Wall time (s)	1202.56	1884.15	1424.59	573.45	627.24
Completion tokens	91615	69766	58598	34149	23887
Aggregate gen (tok/s)	76.18	37.03	41.13	59.55	38.08
Reasoning chars	291890	192495	93697	0	0

6. Discussion

6.1 The mathematics delta (updated)

MiniMax-M2.7 remains at zero on mathematics. GLM-5.1 and DeepSeek-V3.2 each clear 9 and 9 of 18 respectively. Qwen3-235B scores 15 of 18 without reasoning traces (a pure-transformer non-thinking run) and Qwen3.5-397B scores 18 of 18, also without reasoning traces under the `enable_thinking=false` override. The Qwen result changes the framing of v1.2's mathematics finding: the reasoner-versus-non-reasoner gap observed at three models is not actually about reasoning traces; it is about model quality on short-answer mathematics at matched footprint. A strong non-thinking model (Qwen3.5 at 17 B active) outperforms every thinking model in the matrix on this category under the same budget.

6.2 Correctness versus format

Qwen3.5 leads on both axes: 58 of 60 correctness and 29 of 30 format. Qwen3 is second on both axes at 54 and 28. DeepSeek's format leakage observed in v1.2 (extra prose around bare JSON, triple-backtick fences around bare code) persists; Qwen models are closer to MiniMax's tight format compliance, emitting bare JSON and bare code on structured prompts without wrappers. Applications driving any of these models in strict-output modes should still expect to strip wrappers at the harness layer for DeepSeek but can generally take Qwen output as-is.

6.3 Throughput ordering (extended)

Measured aggregate generation rates: MiniMax 76.18, Qwen3 59.55, DeepSeek 41.13, Qwen3.5 38.08, GLM 37.03 tokens per second. The v1.2 finding that transformer-block count dominates active-parameter count as a latency predictor holds for the three pure-transformer entries (62, 79, 61 blocks; 10, 32, 37 B active), and Qwen3 at 94 blocks and 22 B active falls in line with that scaling. Qwen3.5's 38.1 tok/s at 60 nominal blocks and 17 B active is slower than Qwen3 despite a lower block count and lower active-parameter count; the Mamba2 SSM blocks appear to carry higher per-token work at this fabric than a matched transformer block, which is consistent with the experimental kernel coverage for Mamba2 in `llama.cpp` build 8070 and is worth re-measuring when the SSM kernel path matures.

6.4 Floor-tier sanity

All five models clear most of the floor tier: MiniMax 12/15, GLM-5.1 15/15, DeepSeek 14/15, Qwen3-235B 15/15, Qwen3.5-397B 15/15. Qwen3.5 is the first model in the programme to score a full 15 of 15 at floor. The grader self-test (every canonical reference response across all 30 prompts scores a full three of three) was re-run before every graded sweep in this paper.

6.5 Ceiling-tier separation

Ceiling-tier scores (10 prompts, max 30): MiniMax 17/30, GLM-5.1 25/30, DeepSeek 19/30, Qwen3-235B 23/30, Qwen3.5-397B 27/30. All three losses for Qwen3.5 in this paper occur in the ceiling tier. Ceiling is where the five-model matrix retains meaningful spread: the suite is near-saturating at floor and discriminator tiers for the top two Qwen entries but still bites at ceiling, which is the right shape for a v2 suite.

6.6 Default thinking modes and the matched-budget trap

The Qwen3.5 thinking-mode run is the clearest illustration in this programme of a matched-budget methodology trap. Under the same 8 192 token per-request budget that gives the non-thinking run 96.67%, the thinking-mode run scores only 63.33% and emits 291,352 reasoning characters across the suite with largely empty visible content on budget-bound prompts. A direct inspection of the saved traces for the code category, for example, shows the reasoning stream ending mid-sentence at the 8 192 token boundary with phrases such as 'Okay, generating response. Wait, one detail': the model is about to switch from chain-of-thought to visible output when the budget expires. The `enable_thinking=false` override skips the chain-of-thought entirely at prompt-construction time and produces the 96.67% headline. The conclusion is that any matched-budget comparison of models whose chat template defaults to thinking-on must either raise the budget until the chain-of-thought reliably terminates or pin `enable_thinking` explicitly. This paper takes the second option to keep budget matched.

6.7 Suite saturation and v3 roadmap

Qwen3.5 scores 96.67% on the v2 suite, with a perfect floor, a perfect discriminator tier and only three losses at ceiling. A v2 suite has been on the roadmap as a floor-and-ceiling

instrument since v1.2, and the programme deliberately chose a n=30 single-sample instrument because a larger instrument would have traded statistical power for the ability to iterate rapidly. That trade was correct for v1.2 but v1.3 shows that the top of the suite is now within one or two prompts of saturation. A v3 suite is planned with (a) an expanded ceiling tier (four ceiling prompts per category, up from two), (b) explicit research-grade prompts requiring multi-step synthesis and proofs rather than short answers with FINAL-markers, and (c) a formal sampling temperature sweep at t=0.0 and t=0.7 alongside the current t=0.2 run to provide a within-model variance estimate that the current n=30 single-sample design cannot.

7. Threats to Validity

The suite remains n=30 with six prompts per category and no bootstrap; a single-prompt shift changes suite total by 1.1 points and category total by 5.5 points, so results to the nearest percent are reliable and results to the tenth are not. The grader is deterministic but not infallible: a correct answer expressed in a form the grader does not accept scores zero. The Qwen3.5 headline depends on the correctness of the `chat_template_kwargs.enable_thinking=false` path in `llama.cpp` build 8070; a direct probe verified that the rendered prompt does not contain a `<think>` block and that the returned `reasoning_content` is empty, so the claim is well-founded at this build. Throughput figures are the aggregate rate over the full suite including prompt-processing, not a steady-state decode figure under a single-request load. Qwen3.5's SSM kernel path is experimental in this `llama.cpp` release; absolute tok/s for Qwen3.5 specifically should be re-measured when kernel work matures. All five models were run at `temperature 0.2`.

8. Related Work

Public benchmarks for all five models exist at frontier-hardware scale on vendor-developed suites (MiniMax's in-house evals, GLM-5.1 from Zhipu AI, DeepSeek-V3.2 from DeepSeek AI, Qwen3 and Qwen3.5 from Alibaba DAMO Qwen team). None are directly reproducible on a workstation-class cluster because they run on A100 or H100 nodes with vLLM or sglang and tensor-parallel topologies. This paper's contribution is twofold: first, to report the practical quality floor of five matched-footprint frontier MoE releases on a single-operator sovereign compute cluster under `llama.cpp` TCP RPC; second, to publish a mechanical 30-prompt suite that discriminates at this scale and that is approaching saturation at the top, motivating the v3 revision described in Section 6.7.

9. Conclusion and Roadmap

On a 30-prompt matched evaluation under a shared 8 192 token per-request budget, the five models score 96.67% (Qwen3.5-397B-A17B UD-Q4_K_XL, no-think), 91.11% (Qwen3-235B-A22B Q4_K_M), 81.11% (DeepSeek-V3.2 UD Q2_K_XL), 85.56% (GLM-5.1 UD Q2_K_XL), and 64.44% (MiniMax-M2.7 Q8_0). Qwen3.5-397B leads the field by more than 25 percentage points over MiniMax and more than a full standard-of-significance over the next entry. Four headline findings carry into the subsequent revisions: first, a hybrid Mamba2+transformer MoE at 17 B active outperforms every pure-transformer model in the matrix on the v2 suite at matched footprint, budget and transport. Second, default thinking modes are a methodology hazard at matched budget and must be pinned explicitly. Third, the v1.2 finding that transformer-block count dominates active-parameter count as a latency predictor holds for the four pure-transformer entries. Fourth, the v2 suite is near-saturating at the top and needs a v3 ceiling-tier expansion. Planned subsequent work: (a) the v3 suite as described in Section 6.7;

(b) a budget-asymmetric v2.1 run isolating the budget-versus-reasoning term across all five models; (c) a 400 Gbps ConnectX-7 fabric upgrade (Part 4); (d) addition of Nemotron 3 Ultra at NVFP4 to the shoot-out when weights are published.

Appendix A. Prompt list

ID	Category	Difficulty
M1_arithmetic_trap	mathematics	floor
M2_aime_easy	mathematics	discriminator
M3_aime_medium	mathematics	discriminator
M4_putnam_easy	mathematics	discriminator
M5_aime_hard	mathematics	ceiling
M6_imo_sl_medium	mathematics	ceiling
C1_easy_algorithm	code	floor
C2_leetcode_medium_merge_intervals	code	discriminator
C3_leetcode_medium_longest_palindrome	code	discriminator
C4_leetcode_medium_topological_sort	code	discriminator
C5_leetcode_hard_longest_increasing_path	code	ceiling
C6_leetcode_hard_longest_valid_parentheses	code	ceiling
K1_undergrad_biochem	graduate knowledge	floor
K2_grad_named_reaction	graduate knowledge	discriminator
K3_grad_atomic_physics	graduate knowledge	discriminator
K4_grad_realtime_scheduling	graduate knowledge	discriminator
K5_specialist_crystallography	graduate knowledge	ceiling
K6_specialist_cft	graduate knowledge	ceiling
T1_heater_time_of_use	reasoning traps	floor
T2_kilns_off_by_one	reasoning traps	discriminator
T3_bill_red_herring	reasoning traps	discriminator
T4_round_trip_avg_speed	reasoning traps	discriminator
T5_seating_contradiction	reasoning traps	ceiling
T6_knights_knaves_modal	reasoning traps	ceiling
S1_easy_json	structured output	floor
S2_acrostic	structured output	discriminator
S3_markdown_table	structured output	discriminator
S4_json_array_constraint	structured output	discriminator
S5_lipogram	structured output	ceiling
S6_sonnet	structured output	ceiling

Acknowledgements

Models: MiniMax-M2.7 by MiniMax AI, GLM-5.1 by Zhipu AI, DeepSeek-V3.2 by DeepSeek AI, Qwen3-235B-A22B and Qwen3.5-397B-A17B by the Alibaba DAMO Qwen team. Quantisations and redistribution by Unsloth (UD variants) and by the open-weights community. Framework: `llama.cpp` by Georgi Gerganov and contributors, build 8070. Hardware: NVIDIA RTX PRO 6000 Blackwell (Workstation Edition and Max-Q) GPUs; Mellanox ConnectX-6 NICs; MikroTik CRS812-4XS-RM switch. Operator: PureTensor Ltd.