Does Separate Teacher Reasoning Help Compact Students? A Controlled Distillation Technical Report (Kasui Tanku 2B)

Status: Completed technical report (v1.0).
Updated 24 November 2026. Controlled local training, three-seed development and held-out MATH-500 evaluations, AIME 2024 evaluation, seed-42 secondary equivalence adjudication, and Google Colab NVIDIA L4 and A100 40 GB paired EvalScope general-capability evaluations (including full IFEval, MMLU-Pro, MMLU-Redux, C-Eval validation, and SuperGPQA) are complete. Full MMMLU evaluation, controlled idle-GPU latency benchmarking, and independent human validation of secondary adjudication remain open. Authorship and institutional affiliations have not been assigned.

ABSTRACT

Teacher-generated reasoning datasets often return a separate reasoning field alongside a final explanation, introducing duplicated mathematical derivations and formatting instructions. We conduct a controlled empirical study testing whether retaining this separate reasoning field improves a compact student's accuracy enough to justify its added training and inference cost. Using `deepseek/deepseek-v4.1-flash-1017` through an explicitly pinned OpenRouter FP8 endpoint (`deepinfra/fp8/flex`) as the teacher and `Qwen/Qwen3.5-2B` as the local student, a deterministic gate accepted 79 of 100 selected MATH training responses; 71 fit a common 1,024-token training limit. From those 71 identical questions and reference answers, we constructed matched full-trace and final-explanation-only supervision sets and trained native BF16 LoRA adapters for 54 optimizer steps (216 example presentations) on an Intel Arc 140V across three random seeds (42, 43, and 44).

On 25 development questions at a 512-token budget with thinking disabled (seed 42), normalized boxed-answer agreement was 5/25 for the untuned student, 11/25 for full-trace supervision, and 10/25 for final-explanation-only supervision, with mean output lengths of 489.36, 398.12, and 265.28 tokens, respectively. On the 500 held-out MATH-500 questions (greedy decoding, thinking disabled, 512-token cap), the untuned baseline scored 119/500 (23.8%), whereas mean primary accuracy across seeds 42-44 reached 42.53% (213, 220, and 205/500) for full-trace supervision and 40.73% (205, 214, and 192/500) for final-explanation-only supervision; full-trace led by 8, 6, and 13 questions across the three seeds. The released seed-42 full-trace checkpoint (**Kasui Tanku 2B**) achieved 213/500 (42.6%, +18.8 percentage points over baseline) under primary grading and 235/500 (47.0%) versus 133/500 (26.6%) under blind AI-assisted secondary equivalence review. On all 30 AIME 2024 problems at the same 512-token cap, both the untuned baseline and Kasui Tanku 2B scored 0/30 because every response exhausted the token cap before emitting a boxed answer.

In paired cloud evaluations on Google Colab NVIDIA A100 40 GB (vLLM 0.34.0, EvalScope 1.15.0, BF16, non-thinking, temperature 0, seed 42), Kasui Tanku 2B improved over stock `Qwen3.5-2B` on full MMLU-Pro (+2.83 pp, 5,553 vs. 5,213 / 12,032; approx. 95% CI +1.97 to +3.69 pp), full MMLU-Redux (+1.95 pp, 3,953 vs. 3,842 / 5,700; approx. 95% CI +0.78 to +3.11 pp), and full SuperGPQA (+5.05 pp, 5,076 vs. 3,735 / 26,529), tied on full C-Eval validation (845/1,346, 62.78% each), and declined on full strict prompt-level IFEval (-2.96 pp on A100, 339 vs. 355 / 541, approx. 95% CI -6.58 to +0.66 pp; -2.22 pp on NVIDIA L4, 350 vs. 362 / 541). Paired error partitioning reveals that net gains on MMLU-Pro (+423 net items out-of-range vs. -83 in-range) and MMLU-Redux (+205 net items out-of-range vs. -94 in-range) are strongly associated with fewer out-of-range answer extractions under the frozen grader, so formatting, truncation, and underlying knowledge effects cannot be separated from these runs alone. Controlled latency benchmarking remains open.

1. Research Question and Motivation

Research Question: For the same answer-filtered training questions and optimizer-step budget, does retaining a teacher's separate reasoning field improve a compact student's accuracy enough to justify its additional training and inference costs?

Hypothesis and Epistemic Stance: Final-explanation-only supervision may preserve useful solution structure while discouraging duplicated reasoning. A null or negative result is equally reportable. Shorter training targets do not automatically imply shorter generated outputs or better mathematical reasoning, and this study does not claim a new distillation algorithm.

LiteCoT / difficulty-aware prompting motivates measuring reasoning length directly rather than assuming longer traces are always preferable. Mixed-Policy Distillation studies teacher compression of student trajectories under a distinct training framework; our deterministic removal of a separate teacher reasoning field is neither a reproduction of that method nor an implementation of difficulty-aware rewriting. The Qwen3.5-2B model card further motivates tracking thinking loops and generation-budget exhaustion separately from headline accuracy.

The intended contribution is a reproducible, hardware-constrained empirical study of a specific supervision choice, deriving its value from paired controls, auditable failures, multi-seed replication, cross-benchmark transfer audits, and transparent reporting of negative and mixed outcomes. Neither Intel GPU execution nor stripping a reasoning field constitutes methodological novelty on its own.

---

2. Data and Split Design

MATH training data are drawn from `EleutherAI/hendrycks_math`, pinned to revision `21a5633873b6a120296cce3e2df9d5550074f4a3`. Held-out competition math problems are drawn from `HuggingFaceH4/MATH-500`, pinned to revision `6e4ed1a2a79af7d8630a6b768ec859cb5af4d3be`.

From MATH's training split, we selected 100 training questions and 25 development questions using subject round-robin selection and a seeded shuffle. Selection restricts reference answers to scalar forms supported by the exact-answer grader and excludes Asymptote diagrams. Normalized exact-duplicate checks and a heuristic five-token-shingle overlap filter reduce surface leakage across training, development, and reserved test splits, though they do not guarantee semantic independence. Because of scalar filtering and sample size, the training pool is not representative of unrestricted MATH, and public pretraining contamination cannot be ruled out.

All 500 MATH-500 examples were held out from teacher generation, adapter training, and development tuning; they were evaluated only after training protocols were locked. Dataset revisions, raw source checksums, exclusions, and split hashes are recorded in the split manifest.

---

3. Teacher Generation and Filtering

Teacher responses were generated with `deepseek/deepseek-v4.1-flash-1017` via OpenRouter, explicitly pinning provider `deepinfra/fp8/flex`, requesting `quantizations: [fp8]`, and disabling provider fallback. All 100 training

requests returned provider `DeepInfra`. Advertised FP8 precision is an endpoint-reported property rather than an independently audited physical execution guarantee, and provider identity does not establish immutable weights or a deterministic serving stack.

Requests used temperature `0.6`, `top_p = 0.95`, high reasoning effort, and a 4,096-token completion ceiling. Saved records preserve effective request bodies, returned model and provider identifiers, response IDs, raw messages, finish reasons, token usage, and API-reported cost. API nondeterminism and backend sampling implementation details remain limitations.

A deterministic acceptance gate required:

1. A well-formed thinking block and normal completion termination.
2. A final `\boxed{ ... }` answer matching the reference under explicitly implemented normalization.
3. Passing repetition heuristics and an approximate text-token ceiling of 1,800 (distinct from the student's actual-tokenizer training limit).

A passing final boxed answer does not guarantee that every intermediate reasoning step is valid. All rejected records are retained alongside their rejection reasons:

- **Training pool (100 questions):** 79 accepted, 21 rejected; API-reported cost was `$0.0212813664`. These are acceptance-gate counts, not a benchmark of teacher accuracy.
- **Development pool (25 questions):** 18 accepted initially; a documented offline grader fix for unbraced single-digit TeX fractions raised offline acceptance to 19 without issuing new API calls. Both grader versions and raw responses are preserved; this reflects a grading fix rather than a model change.

---

4. Matched Supervision Conditions

From the 79 gate-accepted training responses, we constructed two matched supervision conditions:

- **Full-trace condition:** Preserves the complete teacher response (both the separate reasoning block and the final explanation with boxed answer).
- **Final-explanation-only condition:** Deterministically removes the separate think-body while preserving the entire final explanation and boxed answer verbatim, retaining an empty thinking block for native chat-template compatibility. No replacement answers or rewritten derivations were authored for either condition.

Both conditions use the exact same **71 question IDs**, defined as the intersection fitting the common **1,024-token** student sequence limit under the pinned Qwen3.5 tokenizer and native chat template. Eight gate-accepted examples exceeding 1,024 tokens were excluded from both arms so that no example was silently truncated. Prompt tokens through the user turn are loss-masked, while assistant header and completion tokens are supervised. The training loader verifies split membership against the frozen training manifest and rejects duplicate, unaccepted, or mismatched rows. Condition hashes and provenance are recorded in the ablation manifest.

---

5. Experimental Setup, Hardware, and Verification

5.1 Student Model and Local Training Configuration

- **Base Student:** `Qwen/Qwen3.5-2B`, revision `15852e8c16360a2fea060d615a32b45270f8a8fc`, loaded via the Transformers text-only causal LM class (`Qwen3_5ForCausalLM`). Full base model weights passed SHA-256 verification (download record).
- **Local Hardware and Software:** Intel Arc 140V iGPU on Windows with approximately 31.4 GiB system RAM; PyTorch `2.15.0+xpu`, Transformers `5.23.0`, PEFT `0.21.3` (resolved dependencies). Native reference kernels can be slow on this device.
- **LoRA Training Hyperparameters:** Native BF16 LoRA (not 4-bit QLoRA; teacher FP8 serving is unrelated to student training precision), initialized independently from the original base weights for each arm and seed (never from smoke-test checkpoints). Rank `16`, alpha `16`, dropout `0`, batch size `1`, gradient accumulation `4` (effective batch `4`), learning rate `2e-4`, AdamW (`betas = (0.9, 0.999)`, `epsilon = 1e-8`, `weight_decay = 0.01`), constant learning rate (no scheduler), gradient norm cap `1.0`, activation checkpointing enabled, **54 optimizer steps** without dropping the final partial epoch (`216` example presentations). Equal steps and example presentations do not constitute equal supervised-token or compute budgets across conditions; both conditions are reported with wall times.
- **Adapted Modules:** `q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, and `down_proj` (`10,911,744` trainable parameters; not every projection in the hybrid architecture is adapted). Loss is averaged over supervised tokens within each example, then averaged across the four accumulated examples.
- **Seeds and Released Checkpoint:** Replicated independently across random seeds `42`, `43`, and `44`. The released checkpoint (**Kasui Tanku 2B**) is the seed-42 full-trace adapter (weights SHA-256: `a989b5bb0f11876194b23fb012e6d433c4fa7de24bf14f35dc64133389168d0f`).

5.2 Pre-Pilot Engineering Checks and Artifact Audits

Before and during the controlled pilot, the pipeline passed the following verification checks:

- **Unit and compilation suite:** 26 initial regression tests (expanded to 30 with the pilot runner) covering answer normalization, provider routing, truncation rejection, split integrity, loss masking, ablation transforms, and interrupted-download recovery; all 20 Python source files compiled cleanly.
- **Tokenizer and XPU sanity checks:** Real-tokenizer formatting checks passed on all 71 matched rows per arm. A tiny random hybrid Qwen model completed two LoRA updates on Intel XPU with finite losses, unchanged frozen base weights, updated adapter parameters, and matching logits after adapter save/reload.
- **Convenience trace inspection:** Assistant inspection of seven accepted teacher examples confirmed valid final derivations with independent arithmetic/symbolic checks (a convenience sample, not an estimate of dataset-wide error rate).
- **Local smoke tests (Arc 140V):** A 128-token development inference smoke test completed in `36.38 s` at `3.54 GiB` peak XPU allocation (`3.52 tok/s` including cold-start overhead; hit budget without a boxed answer and made a geometric error). A one-step BF16 LoRA training smoke test completed in `38.57 s` with finite loss `1.1985`, `4.17 GiB` peak XPU allocation, and `10,911,744` trainable parameters. These smoke tests verify feasibility only and do not qualify as steady-state throughput benchmarks or total system-memory bounds.

- **Post-run artifact audits:** Both mid-run (audit_20261117T133638Z.json) and completed-run (audit_20261117T195630Z.json) audits verified that all 22 frozen input hashes matched, saved adapter tensor counts matched metadata, LoRA-B weights were nonzero, all tensors and training losses were finite, and grading replay across all 300 seed-42 development predictions yielded zero disagreements.

5.3 Evaluation Protocols (Local Intel Arc 140V and Cloud NVIDIA GPUs)

- **Local Math Evaluations (Intel Arc 140V):**
  - *Development set (25 questions):* Greedy decoding across four cells: `256` and `512` max new tokens crossed with thinking disabled and thinking enabled. Because seeds 42 and 43 produced byte-identical baseline predictions, extracted answers, grades, and token counts across all 25 questions in all four cells, the seed-44 protocol was amended to reuse the verified seed-42 baseline (seed-44 status); shared baseline records do not add independent baseline observations or seed-44 timing measurements, and an interrupted partial baseline run was excluded. Concurrent local GPU workloads (including Roblox and acceleration diagnostics) overlapped parts of the local pilot, so local wall times are not controlled speed benchmarks.
  - *Held-out MATH-500 (500 questions) and AIME 2024 (30 questions):* Greedy decoding, thinking disabled, `512`-token cap, identical prompts and frozen normalized boxed-answer grader. A missing valid final box is scored incorrect, making the primary metric jointly sensitive to reasoning accuracy and format compliance within the token budget.
- **Cloud General-Capability Evaluations (Google Colab NVIDIA L4 and A100 40 GB):**
  - Executed separately from local Intel runs to test cross-domain transfer and potential instruction-following regressions on stock `Qwen3.5-2B` versus the released seed-42 full-trace adapter (**Kasui Tanku 2B**).
  - *Hardware:* Verified NVIDIA L4 (`23,034 MiB` VRAM via `nvidia-smi`) for initial subset pilots and full IFEval, followed by NVIDIA A100 40 GB for fresh paired subset pilots and full-split evaluations.
  - *Inference and Grading Protocol:* Fresh BF16 inference per arm with `vLLM 0.34.0` and `EvalScope 1.15.0`, thinking disabled, greedy decoding (`temperature = 0`), `seed = 42`, matched unique prompts per pair. Generation caps were `2,048` tokens for IFEval and `1,024` tokens for multiple-choice benchmarks (MMLU-Pro, MMLU-Redux, C-Eval validation, SuperGPQA, and MMMLU). Subset pilots and full splits are strictly segregated in all tables.

---

6. Results

6.1 Local Development Pilot (Seed 42 and Three-Seed Replications)

In the seed-42 run, completing 54 optimizer steps (216 example presentations) required `732.85 s` for full-trace supervision and `475.26 s` for final-explanation-only supervision; different target sequence lengths and concurrent desktop GPU activity prevent interpreting this difference as a controlled throughput measurement.

Table 1 reports all twelve seed-42 evaluation cells on the 25 development questions (`300` total predictions across cells, not `300` distinct questions; values are `correct / 25`).

Generation Budget	Thinking Mode	Untuned Baseline	Full Trace (Seed 42)	Final Explanation Only (Seed 42)
256 tokens	Disabled	0/25	7/25	8/25
256 tokens	Enabled	0/25	5/25	2/25
512 tokens	Disabled	5/25	11/25	10/25
512 tokens	Enabled	0/25	9/25	7/25

At `512` tokens with thinking disabled, full-trace gained 6 baseline-incorrect items with 0 regressions, while final-explanation-only gained 5 with 0 regressions (a 1-question gap, or 4 percentage points, which is too small on 25 items to establish superiority or noninferiority). Mean generated output lengths were `489.36` tokens (baseline), `398.12` tokens (full trace), and `265.28` tokens (final explanation only), with budget-exhaustion rates of `19/25`, `13/25`, and `2/25`, respectively. Notably, final-explanation-only produced a valid final box on `23/25` questions, yet only `10/25` were correct—demonstrating that improved format compliance does not equate to improved mathematical reasoning. All baseline responses at 256 tokens, and at 512 tokens with thinking enabled, exhausted the generation budget before producing a boxed answer; zero scores in those cells reflect budget truncation rather than zero mathematical capability. Raw records are preserved in report.md, comparison.json, and failures_to_review.md.

Table 2 reports the verified three-seed development replications (`correct / 25` on the same 25 questions; seed 44 shares the verified baseline).

Evaluation Setting	Seed 42	Seed 43	Seed 44
Baseline, 256 tokens, thinking off	0	0	0
Full trace, 256 tokens, thinking off	7	9	8
Final explanation only, 256 tokens, thinking off	8	5	7
Baseline, 256 tokens, thinking on	0	0	0
Full trace, 256 tokens, thinking on	5	6	6
Final explanation only, 256 tokens, thinking on	2	4	3
Baseline, 512 tokens, thinking off	5	5	5
Full trace, 512 tokens, thinking off	11	13	10
Final explanation only, 512 tokens, thinking off	10	5	8
Baseline, 512 tokens, thinking on	0	0	0
Full trace, 512 tokens, thinking on	9	9	8
Final explanation only, 512 tokens, thinking on	7	7	7

At 512 tokens with thinking disabled, full-trace supervision (`11` , `13` , `10`) outperformed both the shared baseline (`5`) and final-explanation-only supervision (`10` , `5` , `8`) in all three training seeds, though these repeated 25-item counts capture training-seed variability rather than held-out generalization.

6.2 Held-Out MATH-500 Primary Results (Three Seeds)

All models were evaluated on the exact same 500 held-out MATH-500 questions using greedy decoding, thinking disabled, a 512-token cap, and the frozen normalized boxed-answer grader (paired primary record). The common

baseline was evaluated once.

Training Seed	Untuned Baseline	Full Trace	Final Explanation Only	Full-Trace Lead vs. Explanation-Only
Seed 42 (Released: Kasui Tanku 2B)	119/500 (23.80%)	213/500 (42.60%)	205/500 (41.00%)	+8 questions (+1.60 pp)
Seed 43	119/500 (23.80%)	220/500 (44.00%)	214/500 (42.80%)	+6 questions (+1.20 pp)
Seed 44	119/500 (23.80%)	205/500 (41.00%)	192/500 (38.40%)	+13 questions (+2.60 pp)
Three-Seed Mean	**119/500 (23.80%)**	**212.7/500 (42.53%)**	**203.7/500 (40.73%)**	**+9.0 questions (+1.80 pp)**

Key paired transition counts:

- **Seed 42 vs. untuned baseline:** Full-trace gained `107` items and regressed on `13` (net `+94`, `+18.80 pp`). Final-explanation-only gained `111` items and regressed on `25` (net `+86`, `+17.20 pp`). Across all three seeds, full-trace averaged `+18.73 pp` over baseline.
- **Full-trace vs. same-seed final-explanation-only:** Paired gain/loss transitions were `+64 / −56` (seed 42), `+60 / −54` (seed 43), and `+63 / −50` (seed 44). These differences are descriptive; evaluating three seeds on the same 500 questions does not create 1,500 independent test items, and primary scores remain format-dependent.

6.3 Seed-42 Secondary Boxed-Answer Equivalence Review

Because normalized exact string matching can penalize mathematically equivalent expressions, we conducted a blinded secondary equivalence adjudication on all 1,500 seed-42 MATH-500 responses (500 per arm, matched on question ID, text, and reference). Across 1,500 responses, `627` lacked a final box and received zero credit under both primary and secondary rules. The remaining `873` boxed responses collapsed to `583` distinct `(question, boxed_answer, reference)` triples: `260` resolved mechanically via exact string matching, and `323` were adjudicated in hash order with model identity hidden by Astra, assigning identical verdicts to identical triples across models (decision ledger). This review is AI-assisted adjudication and has not yet undergone independent human validation.

Model (Seed 42)	Frozen Primary	Secondary Equivalent	Secondary Incorrect with Box	Missing Box	Secondary Accuracy Bounds
Untuned Baseline	119/500 (23.8%)	133/500 (26.6%)	8	359	26.6% - 26.6%
Full Trace (Kasui Tanku 2B)	213/500 (42.6%)	235/500 (47.0%)	78	187	47.0% - 47.0%
Final Explanation Only	205/500 (41.0%)	228/500 (45.6%)	191	81	45.6% - 45.6%

The lower and upper bounds coincide because zero triples remained unresolved; they reflect adjudication coverage rather than statistical sampling uncertainty. Notably, full-trace answered 7 more questions correctly than final-explanation-only (`235` vs. `228`) despite leaving `106` more responses without a final box (`187` vs. `81`), whereas final-explanation-only produced `113` more wrong boxed answers (`191` vs. `78`).

6.4 AIME 2024 Evaluation (Null / Budget-Exhaustion Result)

On all `30` AIME 2024 problems, both untuned `Qwen3.5-2B` and the released seed-42 Kasui Tanku 2B full-trace adapter scored `0/30` (0 gains, 0 regressions) under greedy decoding, thinking disabled, a `512` -token output cap, and the frozen normalized boxed-answer grader on Intel Arc 140V. Every single one of the `60` generated responses reached the `512` -token ceiling without emitting an extractable final box. This result measures answer delivery within a strict 512-token budget; performance on AIME 2024 at longer generation budgets remains unmeasured.

6.5 Cloud General-Capability Evaluation: Paired Subset Pilots (NVIDIA L4 and A100 40 GB)

To assess whether math-only distillation transfers to broader reasoning or harms general instruction following, we ran paired subset pilots on Google Colab NVIDIA L4 (`23,034 MiB` VRAM) and subsequently fresh paired runs on NVIDIA A100 40 GB using `vLLM 0.34.0` and `EvalScope 1.15.0` (BF16, non-thinking, `temperature = 0`, `seed = 42`, `2,048`-token cap on IFEval, `1,024`-token cap on multiple-choice benchmarks).

Table 5a: Google Colab NVIDIA L4 Paired Subset Pilots

Pilot Benchmark Metric	Prompts	Stock Qwen3.5-2B	Kasui Tanku 2B	Difference (pp)
IFEval strict prompt accuracy	50	36/50 (72.00%)	34/50 (68.00%)	-4.00
MMLU-Pro accuracy	70	29/70 (41.43%)	37/70 (52.86%)	+11.43
MMLU-Redux accuracy	285	185/285 (64.91%)	193/285 (67.72%)	+2.81
C-Eval validation accuracy	260	161/260 (61.92%)	165/260 (63.46%)	+1.54
SuperGPQA accuracy	360	71/360 (19.72%)	83/360 (23.06%)	+3.33
MMMLU accuracy	70	28/70 (40.00%)	33/70 (47.14%)	+7.14

Table 5b: Google Colab NVIDIA A100 40 GB Paired Subset Pilots

Pilot Benchmark Metric	Prompts	Stock Qwen3.5-2B	Kasui Tanku 2B	Difference (pp)
IFEval strict prompt accuracy	50	34/50 (68.00%)	34/50 (68.00%)	0.00
MMLU-Pro accuracy	70	33/70 (47.14%)	37/70 (52.86%)	+5.71
MMLU-Redux accuracy	285	185/285 (64.91%)	194/285 (68.07%)	+3.16
C-Eval validation accuracy	260	166/260 (63.85%)	171/260 (65.77%)	+1.92
SuperGPQA accuracy	360	70/360 (19.44%)	93/360 (25.83%)	+6.39
MMMLU accuracy	70	23/70 (32.86%)	33/70 (47.14%)	+14.29

Small variations between L4 and A100 stock/adapter greedy subset scores reflect hardware-specific floating-point kernel nondeterminism in vLLM across GPU architectures; within each GPU, both arms were executed identically.

6.6 Cloud General-Capability Evaluation: Full-Split Benchmarks (L4 and A100 40 GB)

Table 6a: Google Colab NVIDIA L4 Full IFEval (541 Prompts / 834 Instructions)

Both arms answered the exact same 541 unique prompts containing 834 verifiable instructions (`2,048`-token cap). Kasui Tanku 2B satisfied every instruction under strict grading on `350` prompts versus `362` for stock `Qwen3.5-2B`. Percentage-point differences are computed from underlying counts before rounding.

IFEval Metric	Denominator	Stock Qwen3.5-2B	Kasui Tanku 2B	Difference (pp)
Strict prompt	541	66.91% (362/541)	64.70% (350/541)	-2.22
Strict instruction	834	76.02%	74.58%	-1.44
Loose prompt	541	71.35%	68.76%	-2.59
Loose instruction	834	79.74%	77.94%	-1.80

Table 6b: Google Colab NVIDIA A100 40 GB Full-Split Benchmark Results

All A100 full splits ran fresh BF16 inference with thinking disabled, `temperature = 0`, `seed = 42`, `vLLM 0.34.0`, and `EvalScope 1.15.0`. An independent CPU audit verified exact item-ID, prompt, and reference-answer alignment across both arms and reproduced totals from saved per-item grades on MMLU-Redux, C-Eval validation, and SuperGPQA (and paired transition audits on MMLU-Pro, MMLU-Redux, and IFEval).

Benchmark (Full Split)	Cap (tok)	Items	Stock Qwen3.5-2B	Kasui Tanku 2B	Net Items	Difference (pp)	Status
IFEval (strict prompt)	2,048	541	355/541 (65.62%)	339/541 (62.66%)	-16	-2.96	Complete
MMLU-Pro	1,024	12,032	5,213/12,032 (43.33%)	5,553/12,032 (46.15%)	+340	+2.83	Complete
MMLU-Redux	1,024	5,700	3,842/5,700 (67.40%)	3,953/5,700 (69.35%)	+111	+1.95	Complete
C-Eval (public validation)	1,024	1,346	845/1,346 (62.78%)	845/1,346 (62.78%)	0	0.00	Complete
SuperGPQA	1,024	26,529	3,735/26,529 (14.08%)	5,076/26,529 (19.13%)	+1,341	+5.05	Complete
MMMLU (14 languages)	1,024	-	-	-	-	-	Pending (70-prompt pilot only)

7. Error Analysis and Interpretation

7.1 Qualitative Local Failure and Trace-Audit Cases

Two documented local inspection cases illustrate why accuracy counts alone conceal critical failure mechanisms:

1. **Reasoning error prior to truncation (`train/algebra/1341`, untuned baseline):** The problem asks for the sum of positive integers on the six faces of a cube given that the sum of the eight adjacent-face vertex products is `1001`. For opposite-face pairs `(a, b)`, `(c, d)`, `(e, f)`, the vertex-product sum factors as `(a + b)(c + d)(e + f) = 1001 = 7 * 11 * 13`; since each pair sum is at least `2`, the face sum is `7 + 11 + 13 = 31`. The untuned student erroneously invented six additional two-face vertices (every cube vertex meets three faces) well before hitting the 128-token cap. Extending the generation budget would allow a final box but cannot be assumed to repair an early conceptual error. Because this occurred on the untuned student before distillation, no blame can be assigned to teacher FP8 quantization or fine-tuning.
2. **Correct answer with flawed reasoning (`train/prealgebra/494`, reasoning review):** An answer-correct response validly computed a mean of `10`, then appended a spurious alternative derivation claiming `70 / 5 = 10`

and asserting that the arithmetic mean is invariant under translation. We retained the primary answer-correct label and logged a separate reasoning-step error. This confirms that final-answer agreement does not verify trace soundness (and because review targeted a convenience sample of answer-correct items, it cannot estimate the population trace-error rate).

7.2 Paired Error Analysis on NVIDIA A100 Full Splits

Table 7 summarizes paired item-level transitions and extraction partitions across the audited A100 full-split runs (paired_change_analysis.json). Approximate 95% confidence intervals use the standard error of per-item paired differences under an independent-item approximation and omit training-seed and subject-clustering uncertainty.

A100 Full Benchmark	Items	Both Pass	Tanku Gains	Tanku Regressions	Both Fail	Paired Diff (pp)	Approx. 95% CI (pp)
MMLU-Pro	12,032	3,987	1,566	1,226	5,253	+2.83	+1.97 to +3.69
MMLU-Redux	5,700	3,323	630	519	1,228	+1.95	+0.78 to +3.11
IFEval (strict prompt)	541	297	42	58	144	-2.96	-6.58 to +0.66

Partitioning changed multiple-choice items by whether the incorrect arm's graded extraction fell inside or outside the valid single-letter option range exposes a crucial structural pattern:

Benchmark	Changed-Case Partition	Tanku Gains	Tanku Regressions	Net Item Contribution
MMLU-Pro	Incorrect arm extraction **out-of-range** (formatting / truncation / unparsed)	616	193	**+423**
MMLU-Pro	Incorrect arm extraction **in-range** (valid wrong option letter)	950	1,033	**-83**
MMLU-Redux	Incorrect arm extraction **out-of-range** (formatting / truncation / unparsed)	280	75	**+205**
MMLU-Redux	Incorrect arm extraction **in-range** (valid wrong option letter)	350	444	**-94**

Subject and constraint patterns in the A100 paired audit:

- **MMLU-Pro STEM subjects:** Chemistry (`183` gains vs. `90` regressions, `+93` net items), physics (`204` vs. `129` , `+75` net items), and mathematics (`208` vs. `154` , `+54` net items) contributed substantially to the overall `+340` net gain.
- **IFEval constraints:** Prompts requiring a specified number of highlighted sections (`detectable_format:number_highlighted_sections`) saw `4` gains and `11` regressions (`-7` net), and prompts prohibiting commas (`punctuation:no_comma`) saw `6` gains and `10` regressions (`-4` net). Because multiple constraints overlap within a single IFEval prompt, prompt-level constraint labels do not isolate which specific instruction failed.

7.3 Interpretation of Full-Split Findings

Taken together, the A100 full-split runs demonstrate a consistent stock-relative accuracy gain for Kasui Tanku 2B on MMLU-Pro (`+2.83 pp`), MMLU-Redux (`+1.95 pp`), and SuperGPQA (`+5.05 pp`), an exact tie on C-Eval validation

(`0.00 pp`), and an observed decline on strict prompt-level IFEval (`−2.96 pp`, with an approximate 95% interval crossing zero at `−6.58` to `+0.66 pp`; `−2.22 pp` on L4).

However, the out-of-range extraction decomposition demands restraint in causal interpretation: on both MMLU-Pro (`+423` out-of-range vs. `−83` in-range) and MMLU-Redux (`+205` out-of-range vs. `−94` in-range), the positive net delta is concentrated in cases where stock `Qwen3.5-2B` failed to emit a parseable in-range option letter within the `1,024`-token cap, whereas valid-letter transitions included both substantive corrections and newly introduced errors with a negative net balance. These descriptive partitions cannot cleanly separate output-formatting changes, generation-budget truncation, and underlying domain knowledge gains. Likewise, the hypothesis that math-only distillation weakened general instruction-following constraints on IFEval remains untested causally. All reported comparisons apply strictly to Kasui Tanku 2B versus its stock `Qwen3.5-2B` parent under the recorded protocol and establish neither a causal mechanism nor a competitive ranking against external compact models.

---

8. Limitations, Integrity, and Open Work

- **Dataset scale and selection bias:** Training used `71` scalar-answer MATH questions excluding Asymptote diagrams, limiting generalization across broader competition math and symbolic proofs.
- **Grading and adjudication bounds:** Exact normalized string matching undercounts mathematically equivalent answers; conversely, passing boxed answers can mask flawed intermediate derivations. Seed-42 secondary equivalence adjudication was AI-assisted (`Astra`) and awaits independent human validation on a documented sample.
- **Public contamination and endpoint opacity:** Public datasets (MATH, MATH-500, MMLU-Pro, SuperGPQA, IFEval) may overlap base-model pretraining. OpenRouter provider pinning is logged, but physical FP8 execution and serving-stack immutability at DeepInfra cannot be independently verified.
- **Hardware and timing controls:** Local training ran on a single Intel Arc 140V device and does not establish cross-device portability. Local development evaluations overlapped desktop GPU activity; steady-state latency and throughput under an explicit warm-up protocol on an idle GPU remain unmeasured.
- **Statistical uncertainty:** Reported A100 confidence intervals assume item independence and exclude training-seed variance and subject-clustering effects. Cloud general-capability runs evaluated only the released seed-42 full-trace adapter, not seeds 43/44 or the final-explanation-only comparator.
- **Transparency and attribution:** Astra assisted with study design, implementation, tests, and preliminary audits; DeepSeek generated the teacher training responses. Human review and responsibility for all final claims are required. The assistant's terse chat style and RTK terminal-output compression are not experimental treatments; all evidence derives from saved raw artifacts. No university endorsement, acceptance likelihood, or publication novelty is inferred from institutional names or cited literature.
- **Remaining open work:** (1) Full 14-language MMMLU evaluation; (2) independent human audit of a documented sample of secondary MATH-500 equivalence verdicts; (3) controlled idle-GPU steady-state latency benchmarking reporting both wall time and generated token counts; (4) AIME 2024 evaluation at longer token budgets (`≥2,048` tokens); and (5) public weight distribution and independent third-party reproduction.

Appendix A. Reproducibility, Checksums, and Adapter Loading

A.1 Reproducing Local Checks and Controlled Pilot Runs

From the repository root, after installing the recorded XPU environment (requirements-resolved.txt) and obtaining the pinned `Qwen/Qwen3.5-2B` revision, the following commands reproduce the unit tests, smoke checks, and seed-42 pilot. Each run writes a timestamped artifact directory; the one-step adapter is a disposable compatibility check excluded from controlled results.

```
.venv\Scripts\python.exe -m unittest discover -s tests -v
.venv\Scripts\python.exe scripts/05_local_baseline.py --revision 15852e8c16360a2fea060d615a32b45270f8a8fc --in
.venv\Scripts\python.exe scripts/07_train_local.py --input data/ablation_v1/full_trace.jsonl --tokenizer-path
.venv\Scripts\python.exe scripts/11_run_local_pilot.py
```

Primary evidentiary artifacts:

- Smoke inference: predictions.jsonl and hardware.json.
- Smoke training: metrics.json.
- Base download checksum: result.json.
- Controlled seed-42 live state: status.json.
- Held-out MATH-500 primary & blind secondary records: paired_primary.json and blind_decisions.json.
- General evaluation ledger and paired analysis: RESULTS.md and paired_change_analysis.json.

A.2 Loading the Released Kasui Tanku 2B Adapter

Released adapter weights (`models/Kasui-Tanku-2B/`, seed-42 full-trace BF16 LoRA) SHA-256:
`a989b5bb0f11876194b23fb012e6d433c4fa7de24bf14f35dc64133389168d0f`

```
from transformers import AutoTokenizer, Qwen3_5ForCausalLM
from peft import PeftModel
import torch

base = "Qwen/Qwen3.5-2B"
revision = "15852e8c16360a2fea060d615a32b45270f8a8fc"
device = "xpu" if torch.xpu.is_available() else "cuda" if torch.cuda.is_available() else "cpu"

tokenizer = AutoTokenizer.from_pretrained(base, revision=revision)
model = Qwen3_5ForCausalLM.from_pretrained(
    base, revision=revision, dtype=torch.bfloat16, attn_implementation="eager"
).to(device)
model = PeftModel.from_pretrained(model, "models/Kasui-Tanku-2B").eval()

prompt = tokenizer.apply_chat_template(
    [{"role": "user", "content": "What is the sum of the first ten squares?\nEnd with \\boxed{answer}."}],
    tokenize=False,
    add_generation_prompt=True,
    enable_thinking=False,
)
inputs = tokenizer(prompt, return_tensors="pt").to(device)
with torch.inference_mode():
    tokens = model.generate(**inputs, do_sample=False, max_new_tokens=512)
print(tokenizer.decode(tokens[0, inputs.input_ids.shape[1]:], skip_special_tokens=True))
```

Sources

- Wu et al. (2025), Concise Reasoning, Big Gains: Pruning Long Reasoning Trace with Difficulty-Aware Prompting.
- Yang et al. (2026), Reasoning Compression with Mixed-Policy Distillation.
- Qwen3.5-2B model card.
- Transformers Qwen3.5 documentation.
- MATH training dataset (`EleutherAI/hendrycks_math`) and MATH-500 (`HuggingFaceH4/MATH-500`).
- OpenRouter provider routing documentation.