Route Apprenticeship: Small Gains, Incomplete Answers

Abraham September 2026 · Exploratory technical report · Version 1.0.0

Abstract

Can model-written explanations improve a smaller model more than reference explanations of the same verified answers? A controlled fine-tuning comparison used Qwen2.5-7B-Instruct, 192 matched training problems, and three paired adapter seeds. Final optimal-route selection accuracy was 65.63% for the base, 70.31% for reference-trained adapters, and 71.61% for teacher-trained adapters. The primary teacher-minus-reference difference was +1.30 percentage points (conditional 95% interval: -1.82 to +4.43). This does not establish a teacher-specific advantage. A separate, post-hoc strict full-response diagnostic yielded only 28.91%, 39.06%, and 39.84%, respectively. The contrast between selecting the right route and producing a completely correct answer motivates a new feedback-guided apprenticeship study. That follow-up is in progress and has no final result here.

1. Question and controlled comparison

Each synthetic problem describes routes, their costs, required features, and blocked features. The model must identify safe routes and select the cheapest eligible route. The frozen primary endpoint checks whether parseable JSON selects an optimal route. It does not require every supporting route and cost to be correct.

The student used rank-4 LoRA adapters on the last two layers' query and value projections: 90,112 trainable parameters. LoRA learns a small weight update while the base model remains fixed. This is parameter learning, not merely supplying a lesson in the inference prompt.

The reference and teacher conditions used identical problems, verified answers, paired ordering, and update counts. The treatment was the explanation text. GPT-5.6 Sol was explicitly told it was preparing fine-tuning demonstrations for this student, whose evaluation required answer-only JSON. It did not receive student-specific error feedback in this study.

The teacher corpus was answer-verified and quality-controlled. One explanation received an authorized, logged sentence correction after masked review; the original was preserved. This is not a comparison of untouched autonomous teacher output against human-written instruction. The reference condition supplies deterministic procedural explanations, and both trained conditions receive answer supervision. There is no answer-only fine-tuned arm.

Each condition was trained at three paired seeds. Each adapter received two epochs, comprising 768 separate answer and explanation updates. The same 48 development tasks were measured at 0, 384, and 768 updates. The final checkpoint was fixed in advance, not chosen using final-test results.

2. Final evaluation

The final set contained 128 untouched tasks: 32 each with 3 routes / 2 features, 4 / 2, 6 / 3, and an unseen size of 8 / 4. Each adapter answered the same set. Therefore, 384 responses per trained condition are not 384 independent tasks.

Condition	Correct final responses	Optimal-route accuracy
Base, measured once	84 / 128	65.63%
Reference explanations, three seeds	90, 88, 92 / 128 each	70.31%
Teacher explanations, three seeds	90, 89, 96 / 128 each	71.61%

Contrast	Difference	Conditional 95% interval
Teacher minus reference, primary	+1.30 pp	-1.82 to +4.43 pp
Teacher minus base	+5.99 pp	-0.78 to +13.02 pp
Reference minus base	+4.69 pp	-1.82 to +11.20 pp

Intervals resample tasks within the four fixed strata after averaging the three fixed seed adapters. They omit uncertainty over new teaching corpora and the population of possible training seeds. They are marginal, not simultaneous intervals. All include zero: the study establishes neither improvement nor equivalence. Four of the five net additional teacher successes came from one seed; no seed was excluded or selected for the headline result.

On the unseen-size 8 / 4 tasks, accuracy was 46.9% for base, 52.1% for reference, and 53.1% for teacher. This descriptive subgroup does not establish broad generalization or an independently powered transfer effect.

3. Selecting correctly is not answering completely

A retained independent post-hoc diagnostic required complete safe routes in task order, correct costs, ordered top-level keys, and the optimal selection. This stricter canonical-format diagnostic is distinct from the frozen primary endpoint and may reject semantically equivalent orderings.

Post-hoc strict canonical response	Correct / responses	Accuracy
Base	37 / 128	28.91%
Reference	150 / 384	39.06%
Teacher	153 / 384	39.84%

The teacher-reference difference is only +0.78 percentage points on this diagnostic. Low full-response accuracy materially qualifies the selection headline. It is not a retroactive replacement of the primary metric.

Seven responses were malformed JSON: six final responses on the same task across all trained adapters, and one teacher development response. They remain failures. No completion hit its output cap, and no failed response was repaired or dropped.

4. Learning and efficiency

Training updates	Reference development accuracy	Teacher development accuracy
0	75.0%	75.0%
384	77.1%	79.9%
768	81.9%	81.3%

These are three-seed means on 48 reused development tasks. The teacher gains earlier, while reference catches up at the second checkpoint. Two trained checkpoints are insufficient to establish a learning-speed advantage. Development and final-test results use different tasks and must not be joined as one learning curve. Individual seed trajectories were not uniformly monotonic.

Reference training processed 284,408 tokens per adapter; teacher training processed 278,452, or 2.1% fewer. Supervised tokens were 49,300 and 43,344. Final output lengths were similar: 34.14 and 33.98 tokens per problem. Matching examples and updates did not match tokens exactly. Teacher preparation and quality review are additional acquisition costs; no all-in cost, controlled latency, or energy-efficiency superiority is established.

5. Verification and reproducibility boundary

All six adapters completed, with 4,608 training updates and 1,520 total development and final observations. The frozen audit reconstructed the analysis and checked schedules, recorded model and adapter identity, artifact hashes, and provenance.

A separate implementation, without importing the production scorer, reproduced all 20 summary and stratum cells and aligned all 1,520 observations with the journal. It checked 368 current tasks, 224 pinned historical tasks, 192 matched training-answer records, and found no normalized or relabel-equivalent overlap across the checked splits and history.

This is independent verification of recorded evidence, not a fresh training replication, formal proof, or external peer review. Shared synthetic rules and vocabulary remain a limitation even when exact and relabel overlaps are absent. No retention or cross-domain test was performed. Public results are aggregated; private task instances and raw model outputs are not included in this publication.

The frozen candidate is identified by SHA-256: 3e8f93e63d9016a95f8903a632be5f4574280583e246576c16286e4c2ed2b4d4. This report summarizes the completed run and its retained independent audit dated 14 September 2026. The PDF checksum on the report page identifies this publication, not an independent replication. No novelty claim is made.

6. What follows, and what does not

The base leaves substantial room for improvement, so a near-perfect baseline does not explain the small teacher-reference difference. The teacher knew the training objective. Plausible remaining explanations include adapter capacity, objective alignment, corpus quality, and the absence of student-specific feedback; this experiment does not distinguish them.

The next study compares adaptive teaching with static teaching, crossed with two replay sources across three fixed parent seeds. Feedback can guide both problem choice and explanations. It measures complete-answer correctness as the eventual primary endpoint, separately from cheapest-route selection, and plans retention evaluation at the final stage. That endpoint is not identical to the order-sensitive post-hoc diagnostic above.

As of 15 September 2026, the first student-training round is in progress. Later teaching, final evaluation, and retention remain separate stages. This report contains no follow-up efficacy result and makes no claim of recursive self-improvement or general intelligence gains.

Authorship and assistance

Abraham directs the research questions, scientific decisions, and release. AI assistants support implementation, analysis, and adversarial review. Automated checks and independent recorded-data scoring support the claims above; they do not substitute for external replication. This technical report is not presented as peer-reviewed research.