Teaching and Specialization: An Exploratory Pilot Archive

Abraham September 2026 studies; public archive prepared October 2026

Purpose and evidence boundary

This archive collects ten earlier pilots and diagnostics that shaped the apprenticeship program. It preserves null results, regressions and engineering failures rather than presenting only favorable experiments. Each section is a separate study with its own tasks, scoring and limitations. Results must not be pooled across studies or compared as a shared leaderboard. These are research summaries, not ten confirmatory replications or author-written personal essays.

1. Teaching packages across eight worlds - 8 September

An eight-condition pilot compared rules, examples, combined teaching and nominal dose increases. Across 96 paired tasks, 1x examples scored 27/96 (28.13%), 1x rules 22/96 (22.92%), 1x combined teaching 17/96 (17.71%), and no help 20/96 (20.83%). Larger nominal teaching budgets produced little or no additional accuracy.

The original 406 attempts stopped; a separately approved 362-attempt follow-on completed. Eight dependent worlds, phase/task-order confounding, and nine truncated responses retained as failures limit interpretation. This is descriptive pooled evidence, not significance or a universal best teaching method. Nominal package dose was not an equal realized-token experiment.

2. Examples, rules and their combination - 9 September

A five-group pilot compared no help, 1x rules, 1x examples, 1x both and 1.5x both. Examples scored 24/32 (75.00%), rules 8/32 (25.00%), combined teaching 17/32 (53.13%), larger combined teaching 15/32 (46.88%), and no help 1/32 (3.13%). In this development task set, combining modes did not beat examples alone.

All 160 calls completed across eight worlds. All 12 failures, including five capped outputs, remained in scoring. World means were equally weighted: 32 tasks are not 32 independent experimental units. This was fixed-weight in-context conditioning, not fine-tuning, and the development-world result does not establish a general prescription for teaching.

3. Answer supervision versus separate explanation supervision - 11 September

Three paired adapter seeds compared answer-only with answer-plus-explanation supervision under a common processed-token ceiling. On unseen arithmetic functions, answer-only adapters averaged 13.54% and combined adapters 14.06%, versus 17.97% for the unchanged model. Paired differences were 0.00, +4.69 and -3.13 points.

Every adapter fell below the unchanged model on both arithmetic slices. One training recipe and three seeds do not establish that explanation supervision is generally ineffective. The ceiling did not equalize updates, FLOPs or supervised tokens. Explanations were programmatically verified, not model-authored or adaptively taught. Recorded audit status was completed and audited.

4. Better likelihood without reliable fresh-task gains - 11 September

A post-hoc diagnostic evaluated the same frozen adapters without additional training. On sampled original training items, accuracy was 9.38% for base, 23.44% for answer-only, and 15.10% for combined supervision. Fresh training-style accuracy was 17.19%, 16.67%, and 14.06%; fresh test-style accuracy was 9.38%, 10.42%, and 9.90%, respectively.

Conditional likelihood improved consistently, but generated-answer accuracy did not show a reliable fresh-task advantage. The diagnostic used 64 original items and 64 fresh computations expressed in two paired formats. Teacher-forced likelihood is not inference competence or evidence that a generated reasoning trace is faithful. The report records successful audit and deterministic reproduction; it remains post-hoc exploratory evidence.

5. A verified lesson did not improve structural transfer - 12 September

In a fixed-weight, three-condition lesson pilot, instructions scored 25/48 (52.08%), a verified teacher lesson 20/48 (41.67%), and task-only prompts 16/48 (33.33%). The lesson and instructions tied at 13/24 on familiar-size tasks; on larger tasks they scored 7/24 and 12/24.

All 144 paired calls used one lesson package and one model, with no weight updates. The pooled rate reflects the chosen equal-size task mixture. Lesson correctness does not ensure the student uses it effectively. Replay and separately implemented scoring agreed; this does not establish a population teaching effect.

6. Compact lessons reduced context cost, not task error - 12 September

On fresh paired tasks, instructions scored 25/48 (52.08%), the full lesson 27/48 (56.25%), and the compact lesson 21/48 (43.75%). Compact teaching used 53.1% fewer student input-plus-output tokens and 48.4% less measured generation RPC time than full teaching, while answering six fewer problems correctly.

This was one package and model with fixed weights. Lower accuracy prevents an unconditional efficiency win. Timing is a systems measurement, not cognitive effort, and research preparation costs were excluded from inference comparisons. Recorded replay and subsequent independent scoring checks passed.

7. Feedback teaching at an arithmetic ceiling - 13 September

The unchanged base solved both 64-item arithmetic sets completely. Generic adapters scored 189/192 (98.44%) on each set. Adaptive adapters scored 190/192 (98.96%) on unseen functions and 188/192 (97.92%) on pool-domain arithmetic. The primary adaptive-minus-generic difference was +0.52 points, while all trained adapters lost retention accuracy.

The adaptive teacher received no incorrect practice examples, so the intended mistake-correction mechanism was not well exercised. A fixed-teacher/fixed-seed conditional interval of [0, +1.56] points is not robust superiority or equivalence. Repeated tasks across three seeds are dependent. Frozen replay and separate arithmetic reconstruction agreed on all 1,344 scores.

8. Route curriculum fine-tuning: repairs and regressions - 13 September

Final route-selection accuracy was 72.92% for base, 69.79% for generic teaching, and 70.49% for adaptive teaching. Adaptive-minus-generic was +0.69 points, with a conditional task-bootstrap interval from -5.90 to +7.29 points.

Both arms repaired some base errors while creating more new errors; net improvement was not established. Curriculum overlap was 23/32 problems. Three fixed seeds and one teacher pair limit inference; route selection

is not complete-answer correctness. Independent scoring corroborated route and retention metrics across 1,392 observations. This precedes and is distinct from the separately published teacher-versus-reference comparison.

9. Strong seen-task fit, limited evidence about transfer - 14 September

Five of six adapters selected the correct route on all 32 of their own training examples; the remaining adapter scored 31/32. Base scored 22/32 on generic curriculum items and 21/32 on adaptive curriculum items.

This is a training-only diagnostic, not fresh generalization. The curriculum panels differed, precluding a matched randomized comparison of arm totals. Load/reload parity on four fixed prompts found no discrepancy but does not independently validate the training system or establish equivalence to its training-time state. The recorded audit and independent selection check covered all 269 generations.

10. Adaptive curriculum: an audited interim comparison - 15 September

Round-one monitor complete-answer correctness was 145/288 (50.35%) for adaptive and 152/288 (52.78%) for static, a difference of -2.43 points. Parent monitor accuracy was 73/144 (50.69%).

This audited cross-attempt snapshot was interim evidence, not a completed final loop. Three of four adopted branches were adaptive and execution histories were asymmetric. There was no second-round, final or retention result in this snapshot. It cannot support causal, population or efficiency inference. The later completed adaptive apprenticeship study is published separately and does not silently rewrite this interim record.

What the sequence informed

These pilots motivated checks for task difficulty, action-format validity, teacher information, complete-answer grading, training-versus-transfer separation, retention, token accounting and saved-adapter recovery. They do not jointly prove that models teach models better than other curricula. Their value is in identifying which hypotheses the next controlled study can actually distinguish.

The cumulative-culture studies, interface failure, answer-only fitting curve and teacher-versus-reference route study already have separate reports in the research catalogue and are not duplicated here. Personal companion essays remain author supplied.