When the Student Does Not Act

Abraham September 2026 · Feasibility failure report · Version 1.0

Abstract

A one-block apprenticeship feasibility experiment compared self-practice, static teaching, replayed teaching and responsive teaching in a cooperative route-choice task. The student, Qwen2.5-1.5B-Instruct, produced invalid actions in all 96 evaluation episodes. Every evaluation response consisted of prose and reached the 96-token output cap rather than producing the required action object. Recorded-run integrity passed audit, but the study did not establish behavioral readiness for a teaching comparison. This report preserves the failure, separates observed facts from causal hypotheses, and describes a prospective diagnostic without presenting that diagnostic as a completed result.

1. Task and intended comparison

The student sees four possible routes, their costs and the features each requires, such as a bridge or tunnel. A truthful partner knows which relevant features are blocked. The student may ask at most two status questions and take at most three total actions, including selecting a route.

The objective is safety first: choose only a route whose required features have been confirmed unblocked, or a feature-free fallback. Subject to that constraint, minimize expected route cost plus the cost of asking questions. Each relevant feature has an independent one-half prior probability of being blocked. Each question costs two units. Responses must contain exactly one query or selection action in JSON; prose is not an executable action.

The intended teaching comparison had four conditions: self-practice, static teaching, replay teaching and responsive teaching. Each recipient condition had eight practice episodes and 24 evaluation episodes. Prior learning records were supplied as retained context; no student weights were updated. A separate teaching model supported the teaching treatments, while the task partner answered status questions truthfully. There was no teacher intervention during evaluation.

2. Observed results

Condition	Evaluation episodes	Invalid actions	Policy successes
Self-practice	24	24	0
Static teaching	24	24	0
Replay teaching	24	24	0
Responsive teaching	24	24	0

All 96 evaluation responses consumed the full 96-token allowance. All 32 recipient practice episodes were also invalid. Across donor and recipient practice there were 43 student generations, of which five were schema-valid queries. A parseable query is not a successful episode; it may still be illegal in the current state.

The completed run contained 139 student generations and 24 teacher generations. Recorded teacher generation cost was $0.345196; auxiliary input-count billing was not independently verified and is not claimed to be free. Early invalid actions shortened episodes, so observed resource use does not estimate the cost of a functioning multi-action study.

3. Diagnosis: facts and hypotheses

Reapplying the strict parser to saved outputs reproduces every evaluation schema failure. The responses were not merely valid action objects surrounded by Markdown fences. They began with prose, and inspected outputs discussed prior routes or criticized learner mistakes instead of acting on the current observation.

The original prompt placed a serialized task, a potentially long learning history and task rules inside one user message. Sorting the serialized keys placed the current observation before the history. The history itself contained serialized records. Task/history confusion and instruction presentation are therefore plausible explanations, but their causal effects were not isolated by this run.

Accidental reusable generation state was considered. Source inspection found fresh generation caches and fresh detokenizers, rather than a demonstrated state-reuse defect. This inspection is not an independent numerical replication of the backend.

The output cap may contribute, but the failure begins before the cap: the model is already producing the wrong response type. Some valid practice queries required only ten tokens. More output space alone is not an established repair.

4. What the failure does not show

The result does not show that responsive teaching is ineffective, that all teaching modes are equivalent, or that the broader social-learning thesis is false. The shared action-interface failure prevents a useful efficacy comparison. Artifact integrity and behavioral validity are different requirements: faithfully recording an unusable response does not make it a successful learning trial.

One block, one student and one task family cannot support general significance or broad efficacy claims. This study measures context-mediated behavior, not parametric learning, inherited model weights or general intelligence.

5. Prospective diagnostic

A separately approved practice-only diagnostic compares the original prompt presentation with a clearer separation of rules, history and current task, at output allowances of 96 and 384 tokens. Four fresh practice tasks, each with empty and long histories, yield 32 episodes across the four combinations. The same student, strict parser and paired greedy decoding are retained.

The primary diagnostic question is whether complete legal action sequences become feasible. Safety and cost-optimal accuracy are reported separately. Clearer presentation is a bundled intervention, and the long-history test uses one fixed practice history; neither supports universal mechanism attribution.

This report contains no results from that follow-up. It is not pooled with the failed block. An engineering screen is not a statistically powered demonstration of improved teaching.

Evidence and reproducibility

The recorded-run audit passed checks for the frozen candidate, source and model identity, request/response receipts, full input token counts, accounting, closed file inventory and deterministic result replay. The original 116 apprenticeship tests passed. These checks verify consistency of recorded evidence, not independent regeneration or behavioral competence.

Student model: Qwen/Qwen2.5-1.5B-Instruct, revision 989aa7980e4cf806f80c7fef2b1adb7bc71aa306. Public replication materials and raw model outputs are not attached. The original failed block remains preserved separately from all proposed repairs.