Study 14 — Premise-Governed Context Scheduling Evaluation

Status: PREREGISTERED STUDY — COMPARATIVE RESULTS NOT YET AVAILABLE

Research question

Does premise-governed context scheduling improve correction behavior in multi-stage AI tasks when compared fairly with full context, status labels alone, and ordinary semantic retrieval? The preregistration focuses on three behavioral endpoints: unauthorized premise use, correction-propagation recall, and false completion, with accuracy, precision, overblocking, resource use, and human intervention as guardrails.

Contribution and current result

The current contribution is a detailed v1.1 preregistered evaluation design, not a performance result. Every held-out base case is planned to run under four conditions in independent stateless sessions: A, full context without status labels; B, the same context with status labels; C, a frozen competitive semantic retriever; and D, premise-governed scheduling with status, provenance, dependencies, contradiction state, unresolved obligations, and dependency-triggered reopening after a correction.

A development harness has been validated for packet construction and parity scaffolding: 40 of 40 parity checks passed. It executed **zero model runs**, consumed no pilot or held-out data, and did not freeze the final Condition C retriever. Prospective naturalistic director episodes were separately recorded with the confirmatory core unchanged and held-out outcomes unread. They are process records, not comparative estimates.

Methods

The design uses paired within-case contrasts over a frozen held-out task distribution, with mechanistic and applied tracks. Each case contains versioned evidence, correction events, completion obligations, an operational D state, and separately protected hidden authorization, dependency, and answer keys. Development, pilot, and held-out cases are cluster-separated. The two-stage episode gives every condition one initial answer and one correction/revision opportunity. Planned analysis uses case-level paired contrasts, stratified cluster bootstrap uncertainty, blinded grading, multiplicity control for nine confirmatory comparisons, and preregistered noninferiority or harm guardrails.

Zackary's role

Zackary is the study owner and human research director. He defined the research target and human-final boundaries, preserved the preregistration, and authorized prospective process recording without changing its confirmatory core or inspecting held-out outcomes. Human decisions are still required for the target task population, smallest meaningful effects, cost weights, acceptable error rates, and final execution authorization.

AI and tool role

AI systems assisted with protocol development, condition scaffolding, packet validation, instrumentation, and prospective process logging. Deterministic validators check condition parity and data boundaries. The evaluated models, retriever, graders, and runtime versions must be frozen before confirmatory execution; AI-generated state cannot use hidden answer or dependency keys.

Verification

The preregistration and development scaffold exist, and the harness reports DEVELOPMENT_SCAFFOLD_VALIDATED. Comparative execution has not started. No pilot outcomes, held-out outcomes, treatment-effect estimates, or replications are available. Condition C remains materially underspecified until its retriever, chunking, query, ranking, and context-allocation rules are frozen.

Exact nonclaims

No causal claim is made that ETI or Condition D outperforms any baseline. Naturalistic cases do not estimate randomized treatment effects. Harness parity is not efficacy evidence. A null result would not establish equivalence without frozen margins, and any gain that violates accuracy, precision, blocking, or cost guardrails would not count as unqualified improvement.

Public artifacts

None yet. The preregistration and harness remain in private/local research sources.

Employer relevance

Study 14 demonstrates rigorous evaluation design for AI systems: fair baselines, leakage control, paired experiments, blinded grading, reproducibility, correction-specific metrics, and explicit deflationary outcomes. It is relevant to AI evaluation, research engineering, model governance, human factors, and safety-focused experimentation.