COMPUTATIONAL UNDERSTANDING: CURRICULUM-DRIVEN CONSTRUCTION AND PROVENANCE-AUDITABLE EVALUATION OF DOMAIN REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Evaluations of LLM-based domain-knowledge systems are vulnerable to a quiet failure: the score can measure the substrate's parametric memory, or the evaluators' own leakage, rather than the constructed artifact. Auditing our own harness with a grounded rubric exposed two such channels — a hand-authored fact table and a construction cache — that supplied 13 of 50 development answers; we removed both, added specified controls, and rebuilt evaluation around a **curriculum** specification — a typed premise (syllabus), a corpus (textbook), and worked examples (problem sets) — scored by a deterministic certificate with the adaptivity of every later step disclosed under an explicit exposure taxonomy. On first evaluation of frozen, receipt-ruled, human-authored exams, the constructed stores lose to premise-conditioned retrieval-augmented prompting in both domains — 7/27 vs 17/27 (biochemistry) and 14/24 vs 15/24 (consumer-fraud law) — but every miss carries a provenance trail from which the builder diagnosed an addressable defect class. Across pre-registered repair iterations on those same fixed instruments (adaptive-reuse exposure disclosed throughout), the stores reach **20/27 and 23/24 with zero answer-time model calls and, under the stated confabulation definition, no invented text or sources identified in the audited campaign-final and fresh-exam outputs**. We then ran a pre-registered **zero-adaptivity generalization test**: one fresh exam per domain, single scored runs. The fraud system met its pre-registered floor (**10/13**); the metabolism system failed decisively (**2/19**), and post-hoc diagnosis identifies a representation limitation — the fresh exam demands relation classes the practice-shaped record ontology never carried, while the engine's failure mode stayed typed refusals and grounded near-misses, with all designed probes refused. The curriculum specification, the provenance-auditable answer artifact, the disclosed repair loop, and the generalization test that binds its own authors are the contributions; the domain systems are the worked examples.

1 INTRODUCTION

A 3-billion-parameter language model takes an exam on metabolic biochemistry and, scored on content alone, its marks do not move when its premise — the ∼200-word typed description of its domain — is stripped to a bare section header: 20/20 either way, because the model can answer from training data or the still-available corpus, and conclusion-checking cannot tell which. Score each answer instead by whether it is *grounded* — traceable to the declared premise and cited passages — and the full premise scores **20/20** while the header collapses to **0/20**. Same model, corpus, questions, cost: the first rubric cannot distinguish parametric memory from the constructed artifact; the second scores only the thing we built (§4.1).

What we built is a *curriculum-driven fact store*, and the education reading is exact: a **syllabus** (a ∼200-word premise along four axes — ontology, epistemology, axiology, praxeology — declaring a type lattice and predicate templates, no facts about any particular enzyme or statute); a **textbook** (a corpus read under that premise, extraction recipes populating the lattice by **mutual induction**, which is why a 200-word premise yields a store three orders of magnitude larger as a structural property); and **problem sets** (a small declared practice set of (question, answer, receipt) triples fixing the question shapes in scope). This third input is a correction this paper's experience forced: premise

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

and corpus under-determine the *shapes of the judgments an exam demands* — our store read every relevant sentence yet held regulation records asserting THAT an effector touches an enzyme but never HOW; and our development exam, we found, had been supplying question shapes illegitimately, by co-evolution. Worked examples are curriculum, firewalled from measurement by the frozen final.

Every record carries provenance and is type-checked against the premise; scope is a property of the storage container, never a model-assigned label. The model's only unrestricted role is extraction — in one domain even that is absent (§2) — and **downstream of construction the executed mechanisms are enumerable; in the primary configuration, no model runs at all** (construction-time interpretation assumptions remain the store's evidential premises; §3).

Grounded scoring is correspondingly mechanical: a certificate check (§3). Each question declares its premise dependencies; an answer scores only if content matches and every declared dependency exists in the premise actually given. Ablate an axis and the certificate fails exactly where it should: the grounded score decomposes axis by axis, reproduces within ± 1 per cell across substrates (Jiang et al., 2023), and predicted blocked sets match measured category-exactly (§4.1) — the dependence structure is a property of the premise, not the model.

Then we turned the rubric on ourselves: auditing our own harness exposed two implicit channels — a hand-authored fact table seeded at scoring time, and evaluation-era content persisting in a construction cache — that inflated the development score by a net 13 of 50 points (16 supplied, 3 poisoned; §4.3). We removed both, regenerated every number clean, demoted the original exam to a development set, and froze new held-out exams: human-authored, receipt-carrying, mechanically validated, sealed before the system saw them. The query class is the one contribution no inference over the corpus can supply — declared, not derived; a human author supplied it under scope-blindness and mechanical validation (§5.1), the single-author dependence disclosed in §6.

The first clean measurement was an honest loss, twice: retrieval-augmented prompting beat the construction 17/27 to 7/27 in metabolism, 15/24 to 14/24 in fraud law. But the misses partitioned into addressable defect classes read from typed gaps and provenance — a work order. Pre-registered iterations later, on the same frozen exams and byte-identical corpora, the constructions score **20/27** and **23/24**: above the baselines they started behind, zero identified confabulations, zero answer-time model calls in the primary configuration, typed refusals on every designed probe in every run. The trajectories — 7→6→8→11→11→11→19→19→20, with one backwards iteration and a stop-rule firing; then 14→16→17→20→21→23, with 9/17 flip and 15/15 hold predictions landing (one flip adjacent, disclosed) — are the paper's central exhibit, and they answer the question a careful reviewer should ask of any premise-driven system — how did the lemma shapes come to fit the domain so well? They didn't, at first; the loop that makes them fit is measured, honest about its failed steps, and it transfers.

Then the harder, pre-registered question: what do the finished systems do on exams they never influenced (§5.10)? Fraud met its pre-registered floor; metabolism failed decisively — and the falsification, reported as-is with its diagnosis, is as much a product of the methodology as the successes: the instruments bind their authors too.

Contributions. (1) The **curriculum specification** — syllabus (typed premise) + textbook (corpus) + problem sets (worked examples) — with the evidence demanding the third input. (2) **Provenance-auditable answer artifacts**: records with provenance, declared typed assumptions (the spectrum $U_0 \leq U_A \leq$ K), or typed gaps naming the missing step — Curry–Howard as design analogy and implemented target (§3). (3) An **instrument-hygiene methodology**: reflexive self-audit, frozen receipt-grounded human-authored exams, pre-registered iterations with falsifiable per-question predictions, adaptivity disclosures, stop-rules, sha-pinned artifacts. (4) The **measured loop, its transfer, and its generalization test**: 7→20 over a 17 retrieval baseline, replicated as 14→23 over a 15 baseline in a second domain, movement attributed by controls and per-layer pre-registrations under the contract of §2 — then the pre-registered zero-adaptivity test the fraud system passed (10/13) and the metabolism system failed (2/19), the failure released with per-item diagnosis under the same one-way door.

Organization. §2 construction and the kernel/adapter contract; §3 the certificate and artifact; §4 development-instrument results and the self-audit that retired them; §5 the frozen-exam protocol, campaigns, LLM arms, and generalization test; §6 discussion and limitations.

Related work (positioning). Attributed and citation-based QA score whether answers carry supporting sources, and the contamination literature motivates leakage-controlled testing (Bohnet et al., 2022; Rashkin et al., 2023; Sainz et al., 2023; Jacovi et al., 2023); our contribution there is the instrument discipline (frozen receipt-ruled exams, pre-registration, one-way doors) plus the grounded certificate that made our own harness's leakage measurable. The store and its folds relate to multi-hop QA, entailment trees, and KG reasoning (Yang et al., 2018; Dalvi et al., 2021; Lan et al., 2021), and competency-question ontology engineering prefigures question-driven design (Ren et al., 2014); what is distinctive is the contract (typed gaps first-class, machine-checked kernel/adapter separation, byte-identity acceptance) and the measured pre-registered loop. Retrieval baselines follow standard RAG practice (Lewis et al., 2020; Edge et al., 2024); selective prediction motivates first-class refusals (Geifman & El-Yaniv, 2017); why-not provenance prefigures structured failure explanation (Lee et al., 2019).

2 Construction, and the Kernel/Adapter Contract

The premise declares, in ∼200–220 words per domain, a type lattice and predicate templates along the four OEAP axes; Phase 0 bootstraps it into machine-checkable declarations (86% average type recovery, development-era; 100% on metabolism). Phase 1 populates the store, and the two domains diverge instructively: the corpus's *register* decides how much language-model construction needs.

Two construction regimes under one contract (detail: Appendix G). Biochemistry prose has no ritual formulas, so metabolism extraction is recipe-driven LLM reading, every triple type-checked at write time, a **type authority** gating merges, **container scoping** making perspective a property of storage. FTC documents are ritual, so fraud construction is **mechanical, zero model calls**: a pre-scan surfaces recurring formulas as cues; records and claim-vs-truth divergence pairs are mined by cue and document structure, every stored text a **machine-checked verbatim corpus substring** — 95 documents, 1,116 records, 122 divergences (100 paired), no model.

The kernel/adapter contract (both domains, machine-checked). Each answer engine splits into a **mechanism kernel** (linking, routing, IDF ranking — the same formula in both kernels — rank-then-cap rendering, typed gaps, aggregate folds) and a **reading adapter**: a data file, logically a section of the premise, holding every piece of domain reading vocabulary as a cited entry. The contract has teeth in three places. Every adapter entry cites its source, and a validator attests every pattern against the frozen corpus (fraud: 141 entries, 0 unattested at campaign close). A **purity self-test** asserts no adapter vocabulary appears in kernel source. And both refactors passed **byte-identity acceptance** on their specified acceptance sets — agreement on those cases, not program equivalence — the metabolism engine, back-fitted after its campaign closed, reproducing all 27 answer heads and the gap probe identically (acceptance artifacts in the supplement). Numeric assembly parameters are adapter data too, disclosed rather than pattern-attested. The two data models remain honestly different — structured records vs verbatim receipts, LLM recipes vs mechanical cues — so the claim is precise: one *contract*, instantiated twice.

The worked-examples channel enters construction as demand: the practice set's question shapes determine which recipe types, record kinds, and fold shapes the library must contain. The demand→recipe loop is executed as a human-directed frontier-LLM loop and measured (§5, §6); record families added this way include directed regulation triples, verbatim-finding annotations (metabolism), and a `procedural_vote` kind with a membership fold (fraud), the latter notable for out-reading the exam's own author (§5.9).

3 THE CERTIFICATE, THE PROVENANCE-AUDITABLE ARTIFACT, AND THE LEDGER

3.1 AXIOM-GROUNDED SCORING: THE CERTIFICATE

Where the substrate has memorized the field's literature, conclusion-only scoring measures the substrate. The grounded rubric scores an answer only if two deterministic checks both pass. The **content check** is keyword-certificate matching against corpus-attested key forms (modes any / all / any-of-k; numeric equality) — mechanical logic, shipped. The **grounding check** requires every declared premise dependency to exist in the premise actually given: the scorer derives a **paragraph-aware AxiomSet** per condition — axioms attributed to their premise sections, so ablating a section removes exactly its axioms — preventing scoring-time reconstruction of removed content. A question whose requirement set references a removed axis becomes ungroundable regardless of store content, which is what makes part of the §4 gradient predictable before any run. **No judge exists in the primary frozen scorer** — no LLM-as-judge, no human rater; the certificate is a deterministic function of (answer text, exam declaration, premise text), both implementations released (the *secondary* audit of §5.8 does use an assistant as a disclosed evaluator).

3.2 SCOPE ASSIGNMENT

In multi-scope domains a record's scope (LAW, SELLER, BUYER, TRUTH in the fraud premise) is a property of the **container** it is stored in, assigned by the construction step that wrote it, never a model-assigned label; the certificate consumes scope structurally, and `TRUTH` is **source-relative** — sworn allegations, not adjudicated fact (Appendix B).

Answer assembly. Answers are assembled by a deterministic question-driven engine under the §2 contract: retrieval seeded by entity mentions in the *question text*, never by gold keys; rendering rank-then-cap, every rendered record carrying its provenance into the artifact (detail: Appendix B).

The provenance-auditable artifact and the assumption ledger. Each answer is an auditable object of three step kinds: RECORD (a store record with provenance), ASSUMPTION (a non-mechanical inference by a contracted slim-LLM role, closed vocabulary, logged, attestation-gated), or GAP (a typed refusal naming the missing step). Verdicts grade $U_0 \leq U_A \leq \mathbf{K}$: U_0 the deterministic no-added-assumption configuration (the strictest grade and our headline), U_A admitting declared logged assumptions, **K** content-only. U_0 certifies provenance and rubric satisfaction, *not entailment*. A Curry–Howard reading (premise as type signature, store as lemma library, artifact as term, ASSUMPTIONs as open hypotheses, GAPs as a missing term with the lemma named) is the design analogy and implemented target, not a property the current scorer checks; metabolism measured the ledger (U_A - U_0 = 0), the fraud replication reports U_0 only, and the typed-path verifier is the declared next strengthening. What each layer checks, and what does not follow from that check alone, is stated once in a five-row table (Appendix B): premise checks declared dependencies are present, not that the answer used them; record checks well-typed fields and source location, not that the source entails the relation; receipt checks verbatim existence, not relevance; composition checks scope and fold rules, not an independently verified derivation; scoring checks the declared match, not every rendered assertion.

Two artifacts, verbatim (Appendix B). An ANSWERED fraud artifact — a machine-checked verbatim receipt with its checks applied — and a designed-probe GAP refusal with its recorded trap scan are reproduced verbatim in Appendix B; both, with full provenance fields, are in the released run JSONs.

4 RESULTS I: THE ABLATION GRADIENT, AND THE AUDIT THAT RETIRED THE DEVELOPMENT INSTRUMENTS

4.1 THE GRADIENT, ITS PREDICTION, AND THE CROSS-SUBSTRATE CHECK

The six-condition gradient table (both substrates, with the blocked column) is in Appendix C; its endpoints are §1's 20/20 vs 0/20, keyword score unmoved.

Much of the gradient is the rubric working as declared — the eligibility rule mechanically blocks a computable set per condition, so it is first a consistency check of the certificate, with the measured residual as its empirical content. The keyword score never moves (one 7B dip disclosed in the supplement); the grounded score decomposes. The **predicted-blocked exhibit** (worked per condition in Appendix C): each single-axis removal blocks its predicted question set exactly, with two small deviations — both of which **replicate identically on both substrates** (the 7B substrate adds exactly one of its own, disclosed): the dependence structure, including its imperfections, is overwhelmingly a property of the premise and exam, not of the model.

4.2 WHAT MOVES THE SCORE: RECIPES, NOT CORPUS

Two development-era claims survive the audit in weakened but genuine form (Appendix C). **Recipe expansion is the lever**: seeding removed, the extraction-only development score rises 6/50 $\rightarrow$ 15/50 across the recipe sequence, scorer fixed. **Corpus accumulation is not**: accumulation curves are flat before recipe expansion and *decline* past their peaks after it.

4.3 THE SELF-AUDIT: TWO LEAKAGE CHANNELS, ONE RETRACTED HEADLINE

Grounded scoring exists to catch free rides; applied to our own harness it found two implicit channels — a hand-written fact table the development scorer consulted, and a construction-era cache carrying its descendants into fresh stores (Appendix C). A 2×2 rescore closes the accounting exactly: FK on/cache on **28/50** (the reported config), on/off 31/50, off/on 25/50, off/off **15/50** (extraction-only).

Hand-authored content is worth +16 points (13 supplied redundantly by table and cache, 3 by the table alone); four garbled cache entries *poison* 3 points wherever the cache is consulted. Both closures hold: 28 = 15 + 16 - 3, 25 = 15 + 13 - 3; the 13 cache-side flips map one-to-one onto the 13 hand-authored entries reachable through it. One further development-era number (32/50) is retracted, with its provenance stated in Appendix C. The consequences structure everything downstream: both channels removed everywhere; the 50-question exam demoted to a labeled development set; every headline number regenerated clean; evaluation moved to frozen instruments (§5) the system had never seen.

4.4 WHAT THE DEVELOPMENT INSTRUMENTS ARE STILL GOOD FOR

Demoted is not discarded (two labeled uses: Appendix C). The reporting rule, enforced throughout: frozen-exam U_0 is the headline with its exposure label (§5.1); frozen-exam K vs the retrieval arm's K is the comparison pair; dev-set numbers are context.

---

5 RESULTS II: FROZEN INSTRUMENTS, HONEST LOSSES, AND THE MEASURED LOOP — TWICE

5.1 THE FROZEN INSTRUMENTS, AND THE EXPOSURE TAXONOMY

Every score carries one of four **exposure labels**: *development* (the co-evolved instruments of §4); *first evaluation* (a frozen exam's initial scored run, unseen by construction); *fixed adaptive instrument* (the same frozen exam under later campaign runs — the questions cannot change, but the builder has seen per-item outcomes, so endpoint scores are engineering predictions on exposed tasks: freezing does not preserve holdout status under reuse, Dwork et al., 2015); *fresh, zero-adaptation* (exam-2,

§5.10: new questions, one run, no change after). The campaigns are evidence about a disclosed repair process on fixed adaptive instruments; §5.10 is the transfer evidence.

Two exams, one discipline. Metabolism: 27 human-authored questions, each with a verbatim source receipt, sha-frozen, with a designed out-of-ontology gap probe. Fraud law: 24 graded items plus 3 memory-trap gap probes (FTX, Fyre Festival, the Made-in-USA rule — real, famous, absent from the corpus by design), same author, mechanically validated (catching key-authoring defects before the freeze), sha-frozen. Both one-way doors held through every scored run below; all shas are pinned in the supplement. One authorial discipline more: **the exam author deliberately never asked what was in scope for the store**, so each baseline measures the data model's reach on an independently conceived question distribution.

5.2 The comparison the field would run, run clean — in both domains

Both arms use the same model (Qwen2.5-3B-Instruct; Qwen Team et al., 2024), corpus, frozen exam, and keyword scorer, no seeding channel; construction answers with the deterministic engine of §2–§3, retrieval embeds the same corpus (MiniLM, Reimers & Gurevych, 2019) and prompts with the premise in context — a **fixed premise-conditioned reference configuration**: fair at first evaluation, where both arms are equally unexposed; the campaign-close comparison is a repaired system against this fixed reference, labeled so rather than claimed as superiority over RAG.

Property	Constr. (metab)	RAG (metab)	Constr. (fraud)	RAG (fraud)
First-evaluation content score	7/27	**17/27**	14/24	**15/24**
. . . at campaign close (fixed adaptive instrument; RAG arm unchanged)	**20/27**	17/27	**23/24**	15/24
Confabulated answers (graded; probes separate)	0	$\geq$4 confident false	0	0 invented / 3 wrong
Typed refusals / named-cause misses	✓ / all	0 / —	✓ / all	hedges / partially
Model calls at answer time	**0**	27	**0**	27

Retrieval wins the opening content score in both domains, reported without qualification; the rest of the table is what the paper is about. In metabolism, the construction arm's twenty baseline misses **partition into four named defect classes**, each localized by its provenance trail (enumerated in Appendix D); the retrieval arm's transcript offers no typed decomposition of its ten misses (retrieved evidence and generated claims can in general be audited separately, Gao et al., 2023; this configuration emits neither typed gaps nor per-claim attributions), and at least four are confident confabulations. Retrieval refused nothing in 27 answers; construction refused exactly where its store had nothing to say. A diagnosable failure list is a work order, and §5.4 executes it under pre-registration.

5.3 The loop, stated

Because every miss carries typed gaps and provenance from which an analyst can diagnose its defect class, the score is a *gradient*, not just a grade. An **iteration**: (i) read the defect classes from the run's provenance; (ii) pre-register the change and its predicted per-question flips, exam untouched; (iii) run once; (iv) report against the prediction, failures included. Stop-rules bind the builder; mid-iteration adaptivity is disclosed. The one-way door makes (iv) meaningful: the instrument cannot adapt to the system, so every point of movement is attributable to the system.

5.4 The metabolism campaign: nine scored runs, 7 $\rightarrow$ 20

(The nine-run ladder table is in Appendix D; the trajectory is §1's, every gap probe refused in every run.)

Iteration 1 made the system *worse* (7 → 6), and the audit convicted the construction from the artifact alone: a surface-lexicon extractor produced records asserting the *opposite* of their own quoted evidence ("activat" inside "inactivates"); we rejected the store and kept the score — a store contradicting its own receipts is the construction-side analogue of retrieval confabulation. 1b repaired extraction soundness; 1c landed on its pre-registered prediction (11/27, all flips as named, zero regressions); iteration 2 measured the assumption ledger at U_A - U_0 = 0, as observed (§6); iteration 3 opened the worked-examples channel (11 → 19); 3b–3c seesawed on kind-routing — **frozen by its pre-declared stop-rule** — while annotation records carried the campaign to **20/27** (controls and full narrative: Appendix D).

5.5 The fraud replication: the campaign transfers

The fraud bring-up replayed the whole protocol in a structurally different domain — multi-scope premise, legal register, mechanical construction. Baseline **14/24** against RAG's 15/24 (the honest loss, replicated), a ten-miss defect taxonomy in five provenance-level classes, then five pre-registered iterations, one scored run each:

(Per-iteration table and narrative: Appendix D.)

5.6 The one residual, fully attributed

The single remaining fraud miss is a paraphrase-shaped doctrine question whose key phrase appears in exactly one sentence of 4,664; the campaign closed rather than widen a cut against a known question (Appendix D). Token-overlap retrieval has a measurable paraphrase ceiling — **the item was missed by every tested configuration, LLM arms included (§5.7)** — of one item in 24 in this domain.

5.7 The fraud LLM arms: retrieval, premise, and parametric memory

Three additional single-run arms on the frozen fraud exam, identical scorer: **RAG premise-conditioned 15/24; RAG without premise 12/24; closed-book 7/24** — retrieval worth +8, premise-conditioning +3 more, every delta item-exact. Three assessments of the fraud RAG arm are reported together: **15/24** frozen (the number every comparison uses), **19/24** manual, **21/24 substantively defensible** under the secondary audit's rubric (§5.8) — six misses are grounded paraphrases failing key phrasing, a substring artifact favoring verbatim-receipt systems, reported rather than left for a reviewer to find; under that same rubric the engine is 23/24, a paired gap of two points versus eight frozen. The probe exhibits — closed-book's invented FTX complaint and Fyre Festival penalty (the engine typed-refused all three probes every run), RAG's one hedge fabricating a context attribution, and the corpus-wide vote item the engine's membership fold answered by reading all 95 documents — are verbatim in Appendix E.

5.8 Where the systems end, and what the scores mean

Final scorecards: metabolism U_0 **= 20/27** vs retrieval 17/27; fraud U_0 **= 23/24** vs 15/24 — each on an instrument the construction opened behind. A **confabulation** is rendered answer text asserting content absent from, or contradicting, its cited source — invented quotations, sources, or relations the receipt does not state; an authentic-but-irrelevant quotation is a *wrong answer*, not a confabulation, and counts against the score. Under that definition, at the strength the audit supports: **no confabulations were identified in the enumerated campaign-final and fresh-exam outputs** — the definition is what convicted iteration 1's rejected store (§5.4) — with zero answer-time model calls in the primary configuration and all 8 designed probes refused correctly. Historical mid-campaign configurations are outside this audited population (reporting a run in a campaign table does not bring it into the audit); the population is enumerated exactly below. The accounting is a property of the checked artifacts (machine-checked verbatim substrings; clause-level provenance audits), reviewed by the builder — not a blinded third-party audit. A **secondary output audit** (supplement: procedure, rubric, evaluator configuration, per-item ledger) applies one semantic-and-source rubric, at response level, to an enumerated population of 199 outputs — every output, probes included, of the campaign-final configurations on exam-1 (27 + 28), both exam-2 runs (15 + 21), and all four LLM arms (81 + 27): 91 engine, 108 LLM-arm — frozen scores untouched. For the one failed run the paper reports,

iteration 1, the audit's status is disclosed, not claimed (Appendix E): the released artifacts establish that defective-batch regulatory content reached the affected answers' rendered text, not whether the specifically documented polarity-inverted records did; the frozen scorer consumed the full strings either way. Audit findings (Appendix E): zero incidental passes among the engine's 55 (seven disclosed *diffuse*); zero invented content across all 91 engine outputs, including the 2/19 collapse run — extended to full-text strength by deterministic regeneration of every truncated engine cache, byte-consistent with every cached head and flag (the metabolism RAG arm's two cache-capped responses: excerpt-only, disclosed); fraud RAG 21/24 defensible vs 19/24 manual (both reported); closed-book inventions reconcile; one new LLM-side defect, a fabricated context attribution inside a RAG hedge (mechanically verified). The audit is assistant-applied, disclosed. Metabolism's seven remaining misses and fraud's one are all named and classified; every construction point exposes an inspectable provenance chain, while a retrieval output exposes prose whose support must be reconstructed (Appendix E). Inspectability, therefore diagnosis, therefore score.

One more sentence about what the endpoint numbers are and are not. The iterations ran against **fixed adaptive instruments** (exposure taxonomy, §5.1), under the declared protocol — single scored run per iteration, post-hoc provenance diagnosis, pre-registered class-level repairs, adaptivity disclosure — so the endpoint scores certify receipt-auditability on those instruments, not generalization beyond them. The generalization claim rests on the protocol's structural holds — the 15/15 hold predictions, the attestation and purity gates (which bar item-ID dispatch and unattested vocabulary — §6 states what they do not bar), the stop-rules, and the cross-domain replication — and it is tested directly, at zero adaptivity, in §5.10.

5.9 Instrument findings the campaigns returned

Three findings about the instruments — the sha discipline catching two stale-configuration runs on first contact, the membership fold out-reading the exam's own author (a third split-vote case, absorbed without touching the frozen exam), and cross-machine byte-identity of the final runs across ARM64 and x86 — are reported in Appendix E.

5.10 The generalization test: fresh exams, zero adaptivity

How much of 20/27 and 23/24 is the *domain*, and how much is convergence to the particular frozen instruments? We pre-registered the answer procedure before a single new question existed: one fresh exam per domain, same author and disciplines plus a per-item **shape tag** (`same-shape` vs `novel-shape`), mechanically validated, sha-frozen; **one scored run per exam against the campaign-final frozen systems, no iteration to follow, results reported whatever they say**, floors pre-set (fraud ≥ 0.70, metabolism ≥ 0.60), falsification condition stated in advance.

The results split. **Fraud: 10/13 (0.77) — floor met**, far above the first exam's 14/24 baseline rate, all three misses named. **Metabolism: 2/19 (0.11) — floor decisively missed**, below even the first exam's first-evaluation 7/27. All designed probes refused correctly (4/4), including one built around a fact present in the corpus but absent from the store; the failure mode under collapse stayed honest — grounded near-misses or typed refusals, zero fabrications.

The falsification has a post-hoc diagnosis — a representation limitation — that the released per-item table supports, short of an intervention-grade causal demonstration. Both metabolism hits are reaction-record questions, the store's native shape; sixteen of seventeen misses demand relation classes outside the record ontology entirely (organ localization, hormone source/effect, carriers, cell types, process-from-description). The first exam sampled the corpus's reaction/regulation shelf; the second its encyclopedic shelf. Under the receipt rule every miss's adjudicating sentence is in the frozen corpus, and a simple IDF ranking places that sentence at **retrieval rank 1 for 15 of 17 misses** — the evidence was one lookup away; the ontology was the boundary. Zero misses diagnose as the cue-vocabulary classes the campaigns bought, and the fraud same-shape stratum exercised repaired classes at 8/10 — evidence, at small denominators, that class-level repairs reused; what did not carry is the *ontology's reach* beyond the practice-shaped distribution.

Three conclusions, at the strength the data supports. First, the outcomes **motivate a hypothesis of register-dependent coverage**, to be tested, not confirmed here: fraud's stereotyped register carried a fresh exam past its floor (10/13; one fewer hit would have missed it); encyclopedic biochemistry's

inventory covered a small fraction of its register's relation classes. Both fresh exams are authored instruments, not random samples. Second, the authored shape tags tracked surface form, but predictive "shape" is mechanism-relative — which record type answers the question — and fourteen of seventeen surface-`same-shape` metabolism items were mechanism-novel; the corrected instrument declares shape mechanically, against the frozen record ontology. Third, and the reason the abstract reports the collapse: the one-way doors did exactly what they are for. A methodology whose instruments can only confirm is not measuring; ours returned a decisive negative against its authors' own pre-registered prediction, cause attributed — and the released miss set prices the next construction round by record family (five families cover sixteen of seventeen misses): a corpus relation census and pre-flight ontology check. Appendix H rebuilds the collapse record-by-record, from the schemas round 1 bought to the pre-flight check that would have caught it.

---

6 DISCUSSION

What the campaigns measured. Both trajectories (7→20, 14→23) occurred with corpora byte-identical — only the demanded lemma shapes and assembly discipline changed — while corpus growth under fixed demands measured flat-then-declining (§4.2): a definition of learning, **feasible-set growth under curriculum at fixed corpus**. The gap taxonomy gained a fifth measured class in fraud (ranking-weight defects).

Sample-exam convergence as hypothesis, and where it broke. Under the hypothesis — not established here — that a register has a fixed, finite class inventory, bring-up would be a coupon-collector over defect classes; §5.10 is its first out-of-sample test, consistent with fraud's narrow register and failed in encyclopedic biochemistry. Assumptions and the census argument: Appendix F.

The slim-LLM service inventory. What does the system need a model *for*? Measured service by service (Appendix F): extraction over open prose (LLM-driven in metabolism, **none** in fraud), answer-time linguistic assumptions (contribution zero in the tested configuration), and paraphrase bridging (the one measured answer-time residual, §5.6). *Builder-side services*: the demand-to-repair loop is executed by a **human investigator directing a frontier-LLM assistant**, disclosed rather than called "manual." The measurement survives structurally: the frontier model answered no exam question; the builder's output is auditable by type; the hygiene binds *the builder as an untrusted party*, human or model — a bounded builder (9/17 flip predictions, one regression, two stop-rule firings). Model-use table: Appendix F.

Reproducing the loop (the artifact is a download) requires comparable builder capability — a stated scope condition; at answer time the executed programs make no model calls, and the one item a ledgered role could plausibly have helped is named (§5.6).

The retrieval boundary; the evaluator's contribution. Every un-anchored ranking heuristic re-imports a little RAG, and the frozen instruments caught each import: the stop-rule is the methodology protecting itself from its builders. The evaluator contributes exactly two declared objects — the query class and the licensed inference vocabulary — fixed while the builder moves (Appendix F).

Limitations. The protocol is executed end-to-end in two domains under one machine-checked contract; a third-domain resettability test is designed, not run. The parametric-memory concern is answered structurally and by measurement: zero answer-time model calls in the primary configuration; every stored fraud text a machine-checked verbatim substring; closed-book at 7/24 on the actual frozen instrument; 2025 corpus documents postdating the extraction model's training data. The adaptivity ledger is disclosed per iteration; the structural protections (U_0 untouched by roles, one-way doors, stop-rules) keep disclosure from becoming license. Scoring is keyword-certificate, not yet path-verified, its verbatim-receipt bias measured (§5.7); the substrate is a 3B model throughout; the fraud replication reports U_0 only. The instruments are small (13–27 graded items), so every claim is about the named instrument and artifact, not a population estimate: the discipline buys auditability, not sample size. All four exams are single-author, mitigated by scope-blindness, mechanical validation, and a release built for third-party examination (second-author exams: future work). §5.10 bounds the headlines: the endpoints certify receipt-auditability and class-level repair within their registers, no more. We claim the human-directed loop, demonstrated twice, and its measured boundary.

AI USE STATEMENT

A frontier LLM assistant was used, under human direction, in four disclosed roles that are themselves objects of study in this paper: as the construction-time extraction substrate in the metabolism domain (the small local models named in the text); as the builder that proposed repairs during the pre-registered campaigns, bound as an untrusted party by the attestation, purity, and byte-identity gates of §2 and the model-use table of §6; as the mechanical validator of the human-authored exams before freezing; and as the disclosed, non-blinded evaluator of the secondary output audit of §5.8, with the frozen primary scores untouched. No generative model was called at answer time in the primary reported configuration, and no exam question was authored by a model. AI-assisted prose and code were reviewed by the authors, all reported numbers are regenerated from the released sha-pinned artifacts, and the authors take responsibility for the final content of this work.

ETHICS STATEMENT

The corpora are public documents (encyclopedia articles on metabolic biochemistry; public regulatory filings and press releases in the consumer-fraud domain). The paper notes explicitly that stored TRUTH values in the fraud domain are source-relative — an agency's sworn allegations, not adjudicated fact — and the released artifacts preserve that framing. No human subjects were involved; the single human exam author is one of the authors, with that dependence disclosed as a limitation.

REPRODUCIBILITY STATEMENT

The supplement releases the frozen exams and harnesses, both corpora manifests, the kernel, adapters, and stores (sha-pinned), every scored run JSON, all pre-registrations and freeze records, the regenerated full answer texts with their byte-consistency verification scripts, and the secondary audit's per-response ledger; §5.9 describes the one-command reproduction path, and Appendix A maps every number in the paper to its artifact and sha256.

REFERENCES

Bernd Bohnet, Vinh Q. Tran, Pat Verga, Roee Aharoni, Daniel Andor, Livio Baldini Soares, Massimiliano Ciaramita, Jacob Eisenstein, Kuzman Ganchev, Jonathan Herzig, Kai Hui, Tom Kwiatkowski, Ji Ma, Jianmo Ni, Lierni Sestorain Saralegui, Tal Schuster, William W. Cohen, Michael Collins, Dipanjan Das, Donald Metzler, Slav Petrov, and Kellie Webster. Attributed question answering: Evaluation and modeling for attributed large language models. *arXiv preprint arXiv:2212.08037*, 2022.

Bhavana Dalvi, Peter Jansen, Oyvind Tafjord, Zhengnan Xie, Hannah Smith, Leighanna Pipatanangkura, and Peter Clark. Explaining answers with entailment trees. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. The reusable holdout: Preserving validity in adaptive data analysis. *Science*, 349(6248):636–638, 2015.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. From local to global: A graph RAG approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*, 2024.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems 30 (NeurIPS)*, 2017.

Alon Jacovi, Avi Caciularu, Omer Goldman, and Yoav Goldberg. Stop uploading test data in plain text: Practical strategies for mitigating data contamination by evaluation benchmarks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023.

Yunshi Lan, Gaole He, Jinhao Jiang, Jing Jiang, Wayne Xin Zhao, and Ji-Rong Wen. A survey on complex knowledge base question answering: Methods, challenges and solutions. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI)*, 2021.

Seokki Lee, Bertram Ludäscher, and Boris Glavic. PUG: A framework and practical implementation for why and why-not provenance. *The VLDB Journal*, 28(1):47–71, 2019.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2020.

Qwen Team, An Yang, Baosong Yang, Beichen Zhang, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. Measuring attribution in natural language generation models. *Computational Linguistics*, 49(4), 2023.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

Yuan Ren, Artemis Parvizi, Chris Mellish, Jeff Z. Pan, Kees van Deemter, and Robert Stevens. Towards competency question-driven ontology authoring. In *The Semantic Web: Trends and Challenges (ESWC 2014)*, 2014.

Oscar Sainz, Jon Ander Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.

A THE RUN/ARTIFACT LEDGER (REALIZED)

The ledger exists as a released artifact: `APPENDIX_A_LEDGER.md` in the supplement maps every number reported in this paper to its artifact file, sha256, and date, organized by section; `SHA256SUMS` covers every file in the supplement tree except itself (the release archive's own sha256 is recorded below); `run_index.md` indexes every scored execution including exclusions; `instrument_inventory.md` lists every evaluation instrument with its exposure label; `secondary_output_audit.md` + `secondary_output_audit_per_item.json` carry the §5.8 audit's procedure and per-response ledger; the `full_answers_*.json` files retain the complete regenerated engine answer texts with their byte-consistency verification, alongside the regeneration and verification scripts.

Release archive, as uploaded: `cu_supplement-20260923.zip`, sha256 `bf873af3e9198771590f1fcdbc6d989120b863e0ef122882375d5389e65594db` (a zip repackaging of the canonical tree, whose tarball sha256 is `0e9dd9e8ca128f59...`). Container-level checksums depend on packaging; per-file integrity is container-independent: `SHA256SUMS` inside the archive covers every payload file individually.

B CERTIFICATE ARTIFACTS, VERBATIM

Two artifacts, verbatim. An ANSWERED artifact (fraud F01, campaign-final run): the engine links `teami` from the question text, routes to the claim/outcome genres, and renders "[claim] A marketer of teas and skincare products has agreed to settle Federal Trade Commission charges that it promoted its products using deceptive health claims and endorsements by well-known social media influencers who did not adequately disclose that they were being paid to promote its products. (ftc_cases/teami__pr.txt ¶10.0)" — the checks applied: stored text a machine-checked verbatim corpus substring with its document and paragraph cited; content check matches key `decept`; no grounding dependency ablated. A GAP artifact (metabolism designed probe H02, "glu296 is a residue produced by which enzyme"): the engine emits "TYPED_GAP: missing entity — no declared-type entity in the question resolves against the store or type authority" — the checks applied: refusal typed, the harness scans the answer for the probe's trap forms ("Acetyl-CoA carboxylase", the incidental subject of the receipt sentence) and records `trap_answer: false`. Both artifacts, with full provenance fields, are in the released run JSONs.

Answer assembly (updated). In the frozen-exam configuration, answers are assembled by a deterministic question-driven engine under the §2 contract: retrieval is seeded by entity mentions in the *question text* — linked against store names, corpus-mined abbreviations or alias banks, and (metabolism) the type authority — never by gold keys; rendering is rank-then-cap with every rendered record carrying its provenance into the artifact. In the fraud domain the exam harness carries plain-text questions only — no category metadata reaches the engine; in metabolism the declared category vocabulary of the exam structure parameterizes rendering order (disclosed, adapter-declared). The development-era gold-keyword-seeded configuration is disclosed in §4 and used only for dev-labeled context.

In multi-scope domains, a record's scope — which party's account it represents (LAW, SELLER, BUYER, TRUTH in the fraud premise) — is a property of the **container** it is stored in, assigned by the construction step that wrote it, never a label the model assigns its own output. Scope-specific extraction writes into scope-named collections, and cross-scope records (the claim-vs-truth divergences) are constructed only by dedicated boundary steps that record their pairing evidence and distance. The certificate consumes scope structurally: divergence questions require paired records from the declared scopes, and no record can drift across scopes after construction. One semantic note the labels require: `TRUTH` is **source-relative** — it holds the enforcement record's own factual assertions ("In truth and in fact…"), i.e., the agency's sworn allegations and findings, not independently adjudicated fact; container membership fixes whose account a record is, never whether that account is true.

What each layer checks (from §3.2).

Layer	Check performed	Not established by it
Premise	declared question dependencies present in the given premise	that the answer used those dependencies
Record	fields well-typed; source location valid (fraud: stored text a machine-checked verbatim corpus substring)	that the source entails the extracted relation
Receipt	exact substring exists in the frozen corpus	that the quotation is relevant to, or supports, the answer
Composition	container/scope compatibility; declared fold and pairing rules with recorded evidence and distance	that a complete derivation was independently verified
Scoring	keyword / numeric match under the declared mode	that every assertion in the rendered answer is correct

C DEVELOPMENT-INSTRUMENT FINDINGS

Two development-era claims survive the audit in weakened but genuine form, and both concern what the lever is. **Recipe expansion is the lever**: with all seeding channels removed (§4.3), the

extraction-only score on the 50-question development set rises from 6/50 to 15/50 across the recipe-expansion sequence — the type-authority rung and the numeric-recipe rung each contribute, and the contribution is store content, not scoring artifacts, because the scorer and its toggles are held fixed across the comparison. **Corpus accumulation is not**: the measured accumulation curves are flat at the extraction ceiling before recipe expansion and band-shifted after it — and both are non-monotonic, *declining* past their peaks as added articles dilute extraction specificity. The clean 139-article store scores within one point of the 60-article development measurement on the same instrument (16/50 vs 15/50, dev label). More corpus without more recipe specificity does not help and eventually hurts.

Demoted is not discarded. The development set retains two labeled uses: the ablation gradient of §4.1, read as a rubric consistency check with a measured residual (a shared instrument does not cancel a confound that interacts with the ablation, so per-condition differences are diagnostic, not conclusive); and regression context during iteration (§5.4, §5.8), where dev-set movement is reported alongside frozen-exam movement but never as evidence. The reporting rule, stated once and enforced throughout: frozen-exam U_0 is the headline, carrying its exposure label (§5.1); frozen-exam K and the retrieval arm's K are the comparison pair; dev-set numbers are context.

Condition	K (3B)	grounded (3B)	blocked	grounded (7B)
FULL	20/20	20/20	0	20/20
-PRAX	20/20	18/20	2	18/20
-AXIO	20/20	17/20	3	16/20
-EPIS	20/20	14/20	6	14/20
-ONTO	20/20	3/20	17	3/20
HEADER	20/20	0/20	20	0/20

(Cells verified from `metab_ablation_patched.json` and its Mistral counterpart; the predicted-blocked exhibit is computed by `predicted_blocked_exhibit.py`, both released.)

The **predicted-blocked exhibit**: removing praxeology is predicted to block exactly the intervention questions (2), and does, exactly; removing axiology, exactly the clinical questions (3), and does; removing epistemology, the regulation questions (5) — measured: those 5 plus one enzyme question whose answer leaned on an epistemology-declared relation. Ontology removal predicts total collapse and measures near-total (3/20 survive). The two deviations from exact prediction — the +1 enzyme spillover under -EPIS, one clinical survivor under -ONTO — **replicate identically on both substrates**, which adds exactly one deviation of its own (-AXIO, the same condition as its keyword dip; supplement): the dependence structure, including its imperfections, is overwhelmingly a property of the premise and exam, not of the model.

One further number is retracted: the development-era summary previously circulated for this system (32/50) is, by the run logs, the maximum of a noisy accumulation curve at a single corpus-size checkpoint under the seeded configuration; at matched corpus size the final recipe's contribution is +1, not +4. We do not report it as a result anywhere in this paper and flag it because it appears in earlier drafts and a related filing.

Grounded scoring exists to catch free rides. Before reporting any of the numbers above, we applied the same discipline to our own harness, and the audit found two implicit channels by which hand-authored content reached scored answers. **Channel 1:** the development scorer consulted a hand-written FACTUAL_KNOWLEDGE fact table when assembling answers for certain question types. **Channel 2:** a construction-era pipeline cache carried evaluation-era factual entries — descendants of the same table — back into freshly built stores. A 2×2 rescore with each channel toggled closes the accounting exactly: FK on/cache on **28/50** (the reported config), on/off 31/50, off/on 25/50, off/off **15/50** (extraction-only); hand-authored content +16 (13 via both channels, 3 table-only), cache poison -3 wherever the cache is consulted.

D CAMPAIGN NARRATIVES AND RESIDUALS

In metabolism, the construction arm's twenty baseline misses **partition into four named defect classes**, each localized by the answer's provenance trail: (1) *regulation direction never extracted* — five questions lost because regulatory records carried effector and enzyme but not activate/inhibit; (2) *cofactor sides missing* from the reaction template — two; (3) *recipe-coverage gaps* — six questions

about fact kinds no recipe targets, every one answered with a typed refusal rather than a guess; (4) *extraction noise and linking misses* — the remainder, each inspectable down to the garbage triple or unlinked surface form. The retrieval arm's transcript offers no typed decomposition of its ten misses (retrieved evidence and generated claims can in general be audited separately (Gao et al., 2023); this configuration emits neither typed gaps nor per-claim attributions), and at least four are confident confabulations: a wrong bond for fumarase, a wrong enzyme for β-oxidation's commitment step, "insulin" offered as an ACC *inhibitor*, "oxaloacetate" for a question it cannot answer. The retrieval arm refused nothing in 27 answers; the construction arm refused exactly where its store had nothing to say, including the designed probe. A diagnosable failure list is not an epitaph — it is a work order, and §5.4 executes it under pre-registration.

(Engine-only controls: baseline store + v4 = 7/27, + v5 = 8/27 — released alongside the nine scored runs.)

Iteration 1 made the system worse, and the audit says exactly why. A surface-lexicon extractor filled regulation directions and swept new regulatory records; its pre-registration predicted 2–5 flips, zero occurred, one prior hit regressed — and the audit convicts the construction: the substring "activat" fired inside "inactivates," producing records asserting the *opposite* of their own quoted evidence; contrastive sentences carrying both polarities were skipped; effectors matched by longest string anywhere in the sentence. We rejected the store and kept the score. The negative result is load-bearing: a store that contradicts its own receipts is the construction-side analogue of retrieval confabulation, and the audit caught it from the artifact alone, with the exam untouched. **Iteration 1b repaired extraction soundness and moved the failure one layer up**: a clause-scoped extractor — validated against the exact sentences that broke iteration 1 before any rescore — produced records faithful to their evidence, and the score moved only 7 $\rightarrow$ 8, with the audit localizing the residual to answer assembly at single-position precision (the three needed records ranked 12th, 14th, 15th under a rendering cap of twelve). **Iteration 1c ranked assembly by the question's own verb and landed on its pre-registered prediction**: 10–11 predicted with three named flips and one named non-flip; measured 11/27, all as named, zero regressions. **Iteration 2 measured the assumption ledger and got a zero**: with the contracted slim-LLM roles active, U_A - U_0 = 0 — the roles added no scored answers in this configuration, their most important recorded act a logged refusal to license an inference; the zero is reported as observed, with candidate explanations weighed in §6. **Iteration 3 opened the worked-examples channel at full strength**: new recipe and record families demanded by the practice set converted five of six typed refusals, 11 $\rightarrow$ 19. **Iterations 3b–3c** seesawed on kind-routing — which failed its prediction a third time and was **frozen by its pre-declared stop-rule** — while annotation records (findings that are verbatim corpus sentences) carried the campaign to **20/27**. The engine-only controls isolate the engine-side contributions they cover (two of the transitions); remaining attributions rest on per-layer pre-registrations and the audits, and one baseline hit changed epistemic status without changing score — its ✓ had ridden an incidental substring and now rides a genuine record with clause-level provenance, a movement invisible to any score-only evaluation and exactly what per-answer provenance is for.

Iter	Layer (changes)	U_0 (fixed adaptive)	Predicted	Flips landed / predicted
1	extraction cues + genre vocabulary (4)	16/24	18	2/4, both by stated mechanism
2	assembly: ranked pairs, render budget, caps (3)	17/24	18	1/2 (mechanism adjacent, disclosed)
3	shapes: scoped reference mining, membership fold, routing, strong-link (4)	20/24	21	3/5
4	ranking: linking-token exclusion, stop-token soundness, windows (3)	21/24	22	1/3

Iter	Layer (changes)	U_0 (fixed adaptive)	Predicted	Flips landed / predicted
5	named residuals: means genre, paragraph coverage, designed-to cue (3)	**23/24**	**23**	2/3

Campaign totals: **17 pre-registered changes across five layers, 9/17 flip predictions landed — eight by their pre-stated mechanisms, one by an adjacent mechanism, disclosed (iteration 2) — and 15/15 hold predictions correct; zero exam regressions across all six scored runs; probes 3/3 in every run; per-iteration score predictions 18→16, 18→17, 21→20, 22→21, 23→23.** Two dev-set incidents were caught by the practice set before scoring and fixed with named causes; mid-iteration adaptivity is disclosed per prereg. Items that resisted one layer were re-classified with measured causes and converted on the re-classified layer: the Zoom-AES item took three layers (routing, then eligibility, then ranking discriminativity), the dark-patterns comparison took linking, then ranking, then routing — the defect taxonomy working as designed.

The single remaining fraud miss is a paraphrase-shaped doctrine question whose key phrase appears in exactly one sentence of a 4,664-sentence reference corpus, sharing only common tokens with the question. Five mechanisms reached toward it across three iterations; the last put its paragraph at rank 6 of 995 under a top-3 coverage cut, and the campaign closed rather than widen a cut against a known question. Token-overlap retrieval has a measurable ceiling against paraphrase — **and the item was missed by every tested configuration, LLM arms included (§5.7)**. That ceiling is one item of 24 in this domain, and we can say so with a number.

Run	Change layer	U_0 (fixed adaptive instrument)	Gap probe
Baseline (clean)	—	7/27	✓
Iter 1	store: direction fill, surface-lexicon extractor	**6/27**	✓
Iter 1b	store: clause-scoped extractor; engine v4	8/27	✓
Iter 1c	engine v5: direction-aware ranking, pathway join	11/27	✓
Iter 2a/2b	assumption-ledger roles (U_A configuration)	11/27 ($U_A = U_0$)	✓
Iter 3	worked-examples channel: new recipe/record families	19/27	✓
Iter 3b	kind-routing	19/27	✓
Iter 3c	annotation records; kind-routing frozen by stop-rule	**20/27**	✓

E SECONDARY-AUDIT DETAIL AND INSTRUMENT FINDINGS

For the one failed run the paper reports, iteration 1, the audit's status is disclosed rather than claimed: the released run retains truncated answer heads, and the rejected store and its regulation-direction audit artifacts are not in the release, so the defective records' answer-level manifestation cannot be fully reconstructed. The released provenance for the affected questions (H22, H23) lists eleven reaction records followed by one regulatory record from the defective extraction batch — under the released renderer the provenance list is the rendered selection, so defective-batch regulatory content was part of those answers' rendered text beyond the retained head; whether the specifically documented glucagon polarity-inverted records rendered anywhere is not established. The frozen scorer consumed the full rendered strings (both questions scored wrong), and the construction defect itself is reported and attributed in §5.4.

Three findings about the *instruments*, worth reporting because the methodology predicts them. The sha discipline caught two stale-configuration runs on first contact (a macOS run and a Spark run whose kernel/adapter predated the freeze record) — the hygiene is not decorative. The membership fold built for the computational item out-read the exam's own author: it surfaced a third split-vote case (Facebook's 3-2 referral vote) the authoring pass had missed, and the membership-style key design absorbed the discrepancy without touching the frozen exam — mechanical folds read corpora more completely than authors, and instruments should be designed for that. And cross-machine byte-identity: the deterministic engines reproduce their final runs answer-head-identically across ARM64 and x86 — reproduction is a download, not an ordeal.

The exhibits: closed-book fabricated at least five confident specifics among its misses and one *inside a pass* (it scores on the token 'ROSCA' while expanding it to a nonexistent "Right of Consumers to Save Money Act" — a substring scorer cannot see a confabulated statute name wrapped around a correct token); it invented an FTC complaint against FTX and a Fyre Festival penalty on the designed memory-trap probes, while the engine typed-refused all three in every run and RAG hedged — though the secondary audit finds one of RAG's hedges itself fabricates a context attribution, claiming its retrieved passage mentions FTX when no corpus document does. And the corpus-wide computational item is the architecture in one line: RAG guessed "2" split votes from its five-passage window; the engine's membership fold read all 95 documents and enumerated the three split-vote cases correctly.

Metabolism's seven remaining misses and fraud's one are all named and classified. The margins understate the difference in what the scores *mean*: every construction point exposes a provenance chain a reader can inspect — records, receipts, folds — while a retrieval output exposes prose whose support must be reconstructed (and whose misses the secondary audit shows are often near-misses). Inspectability, therefore diagnosis, therefore score.

Audit findings: across the engine's 55 passes on four instruments the audit identified zero incidental passes (three flagged candidates inspected and cleared) and seven disclosed *diffuse* passes (the answer present within a multi-record rendering rather than isolated); across all 91 engine outputs zero invented content, including the 2/19 collapse run — a finding the supplement extends to full-text strength: the three engine populations whose caches truncate were regenerated in full from the frozen artifacts (byte-consistent with every cached head and scored flag), and every regenerated remainder verifies mechanically against the store or the cited corpus documents; the metabolism RAG arm's cache truncates two of its 27 responses, and for those two the audit covers only the retained excerpt (disclosed); the fraud RAG arm shows six near-misses that are substring-scorer artifacts (21/24 substantively defensible under the audit rubric, vs the 19/24 of the earlier manual review — both reported); the closed-book inventions reconcile exactly (five, plus one inside a pass); and one new defect surfaced on the LLM side — the RAG arm's FTX-probe hedge fabricates a context attribution, asserting its retrieved passage mentions FTX when no corpus document does (mechanically verified). The audit is assistant-applied and disclosed as such.

F Discussion extensions and the model-use table

Sample-exam convergence as hypothesis, and where it broke. The gates constrain repair form — no item-ID dispatch is expressible, every reading pattern must attest in the frozen corpus, no domain vocabulary may enter kernel code (syntactic checks, supplemented by review; a corpus-attested pattern chosen after an exposed failure can still target that item, which the exposure taxonomy discloses rather than denies) — so — under the hypothesis, not established here, that a register has a fixed, finite class inventory — each sampled exam would purchase coverage of the defect *classes* its misses expose, and the bring-up process would be a coupon-collector over defect classes, common classes bought early (fraud's iteration 1 was four adapter lines), the tail thinning toward rare shapes, the fixed point set by the mechanism vocabulary's expressiveness rather than at 100%. §5.10 is this account's first out-of-sample test: the collector picture was consistent with fraud's narrow register (second exam over its declared floor at zero adaptivity) and failed in encyclopedic biochemistry, where the first exam had sampled only one shelf of a much larger class inventory. The coupon-collector account is therefore a **hypothesis** — it assumes a fixed class inventory, exchangeable-ish sampling, durable class repairs, and no interference between repairs, none established by two campaigns (and the measured regression cautions against simple monotone growth); the register-dependent form arose after the failure and is stated for a later test, not confirmed by the observations that prompted

it — and the way to bound that inventory in advance is a relation census of the corpus itself: under the receipt rule, any admissible exam can only demand relations instantiated in the corpus, so the corpus's relation inventory bounds every exam's, though any practical census is a candidate inventory with declared granularity, not a completeness guarantee. That census, and the pre-flight check of a proposed record ontology against it, are designed future work with the §5.10 miss set as their first priced dataset.

(a) *Extraction over open prose*: the metabolism implementation used LLM extraction; the fraud implementation used **none** — legal complaints are ritual documents whose regularities are tic-shaped and mechanically minable. Necessity in metabolism was not tested; what is measured is that the demand varied with the corpus's register. (b) *Answer-time linguistic assumptions* (the U_A roles): measured contribution zero in the tested configuration — U_A - U_0 = 0 says the contracted roles added no scored answers there; role capability, prompt design, and task fit remain possible explanations alongside a matured mechanical layer — and the roles' enduring function is as a gauge for residual slack; their best recorded act was a logged refusal. (c) *Paraphrase bridging*: the one measured answer-time residual, named to the item (§5.6) — the service a future ledgered role would have to license, quote-attested.

Stage	Model used	Constraint
Premise & exam authoring	none (human drafts all questions, keys, receipts)	receipt rule; scope-blindness
Exam validation	frontier assistant runs fixed mechanical checks and flags defects with repair *directions*; the author decides all wording	pre-freeze only; the assistant's exposure to candidate items is part of the disclosed builder-exposure timeline
Phase 0 type bootstrap	slim LLM (metab); none (fraud)	machine-checked declarations
Extraction	slim LLM via recipes (metab); none (fraud)	type-check at write; provenance; authority-gated merges
Repair design (builder loop)	human-directed frontier LLM	attestation, purity, prereg, stop-rules, diffs shipped
Answer execution (U_0)	**none**	deterministic engines, sha-pinned
Answer execution (U_A)	slim-LLM roles	closed vocabulary, quote-attested, logged
Scoring	none	deterministic certificate
Comparison arms	slim LLM (by design)	self-manifesting artifacts
Post-hoc diagnosis & analyses	frontier assistant	disclosed; never touches frozen instruments
Secondary output audit (§5.8)	frontier assistant as disclosed evaluator	rubric + per-item ledger released; non-blinded; original scores visible; mechanical sub-checks (quote attestation) are code

The evaluator's declared contribution. Under this methodology the evaluator contributes exactly two declared objects — the query class (the frozen exam, extensionally; the worked examples, as curriculum) and the licensed inference vocabulary — and the measurement is meaningful because both are fixed while the builder moves. This division of labor is, we believe, the operational content of "understanding" as this paper uses the word: the query class is declared, not derived.

G CONSTRUCTION DETAIL PER DOMAIN

Metabolism (LLM-extracted construction). Biochemistry prose has no ritual formulas, so extraction is recipe-driven LLM reading: each recipe is a question chain whose earlier answers parameterize later questions, every extracted triple is type-checked at write time and carries provenance, and two disciplines keep the store honest at scale — a **type authority** resolving surface-form equivalence through a two-tier policy (trivial variants auto-confirm; substantive merges gate through an external authority, never the language model), and **container scoping** making perspective a property of storage.

Fraud law (mechanical construction; zero model calls in executing the construction and answering programs). FTC complaints and press releases are ritual documents: "In truth and in fact,"

"the representations set forth in Paragraph N . . . are false or misleading," "The Commission voted 3-2." A mechanical pre-scan (`tic_scan.py`) surfaces these recurring formulas as candidate cues; the construction then mines records and claim-vs-truth divergence pairs by cue and by document structure (numbered-paragraph cross-references), with every stored text a **verbatim corpus substring, machine-checked**, and every divergence pairing a declared structural inference with its distance recorded. The whole fraud store — 95 primary-source documents, 1,116 records, 122 divergences (100 paired) — is built without a model.

H WHERE THE SCHEMA STOPPED — A WORKED WALKTHROUGH OF THE METABOLISM COLLAPSE

This appendix rebuilds the §5.10 falsification from the artifacts, at the level of individual records and questions, for the reader who wants to *see* what the store's schema could and could not say — and why the exam author, applying the same discipline that produced exam-1, walked into the gap without noticing. Every record, answer, and sentence below is verbatim from the released artifacts (frozen store `e2_clean_constructed_regdir2_ann3.pkl`, scored run `cdbee264...`, regenerated full answers, frozen harness).

H.1 WHAT THE ROUND-1 CURRICULUM BOUGHT

The frozen metabolism store contains exactly six record families, each purchased by a demand the round-1 curriculum actually made:

family	count	fields (verbatim schema)	what demanded it
reactions	275	enzyme, substrates, products, pathway, reversible, step	practice/exam-1's dominant shape: "X converts A into what?"
regulatory	219	effector, enzyme, effect/type, mechanism, condition	"what activates/inhibits enzyme E?" (79 carry direction fields added by the campaign's regulation-direction repair)
factual	135	name→value pairs (`Pentose phosphate pathway_start` → `6-phosphogluconate`)	pathway endpoint questions
energetics	61	name→number (`oxphos_atp_per_nadh` → 3)	numeric yield questions
annotations	266	entity, kind, finding (verbatim sentence), article	added by campaign iteration 3c for verbatim-finding questions
diseases	0	(declared, never populated)	—

Two of these families did not exist at baseline: the directed regulation records and the annotations were **purchased mid-campaign by the worked-examples channel** (§5.4) when practice questions demanded them. The channel demonstrably buys families — the question §5.10 tests is whether the families one instrument demanded cover the next instrument's demands.

H.2 THE SHAPE THAT TRANSFERS

Exam-1's H05 — *"Acetyl-CoA carboxylase (ACC) converts acetyl-CoA into what molecule?"* — is the native shape. The linker resolves `Acetyl-CoA carboxylase` from the question text, `_records_for` retrieves reaction record M2:

```
{"id": "M2", "enzyme": "Acetyl-CoA carboxylase",
```

```
 "substrates": ["Acetyl-CoA"], "products": ["Malonyl-CoA"],
 "pathway": "Fatty acid biosynthesis", "reversible": false}
```

and the rendering opens `Acetyl-CoA --Acetyl-CoA carboxylase--> Malonyl-CoA [Fatty acid biosynthesis]`: key matched, grounded, scored. Exam-2's two hits are this same mechanism on fresh instances — H2-16 (*"glucose is converted into what product by glucokinase or hexokinase?"*) scores because the key form sits inside the rendered record set (the audit's disclosed diffuse-pass exemplar). When the question's demanded relation is substrate—enzyme→product, transfer worked.

H.3 Four walkthroughs of the trap

Each walkthrough gives the question (with the author's shape tag), what the engine actually rendered, the corpus sentence sitting at IDF retrieval rank 1 for the question's tokens, and the record that would have had to exist. The pattern to watch: the engine's answers are grounded, on-topic, and *structurally unable to say the answer*.

H2-05 — organ localization (tagged `same-shape`). *"What organ stores glycogen for use throughout the body?"* The engine linked `Glycogen` and rendered flawless glycogen chemistry: `Glycogen --Glycogen phosphorylase--> Glucose-1-phosphate [Glycogenolysis] ; Glucose, Glucose --Glycogen synthase--> Glycogen [Glycogenesis] ; ...` — every record grounded, no record anywhere in the six families with a slot for *where*. The adjudicating sentence was at retrieval rank 1: *"Liver glycogen stores serve as a store of glucose for use throughout the body..."* The missing row is a localization record — roughly `(entity: glycogen store, site: liver, role: whole-body glucose reserve)` — a family no practice question ever demanded, because exam-1 never asked *where anything happens*.

H2-04 — carriers (tagged `same-shape`). *"What is the major carrier of free fatty acids in blood?"* Rendered: fatty-acid reactions and two regulatory records (`Fatty acids inhibits Acetyl-CoA carboxylase (allosteric)`) — chemistry again. Rank-1 sentence: *"Serum albumin is the major carrier of free fatty acids in the blood."* Missing family: transport — `(species: free fatty acids, carrier: serum albumin, medium: blood)`. The store has no notion of anything *carrying* anything.

H2-10 — hormone action on a process (tagged `same-shape`). *"What hormone stimulates lipogenesis by activating two enzymatic pathways?"* This is the subtlest case, because **the store contains the fact**. Two directed regulation records, verbatim:

```
{"effector": "insulin", "enzyme": "Acetyl-CoA carboxylase",
 "type": "activate", "evidence_article": "Coenzyme_A"}
{"effector": "insulin", "enzyme": "Glycogen synthase",
 "type": "activate", "evidence_article": "Glycogen"}
```

But the regulatory schema is keyed effector→**enzyme**, and the question approaches from the other side: its only store-resolvable entity is *lipogenesis*, a pathway name, so the engine rendered the Lipogenesis reaction chain — in which insulin never appears, because no record points from a *process* to the hormone that stimulates it. Rank-1 sentence: *"Insulin stimulates lipogenesis primarily by activating two enzymatic pathways."* The gap is not fact coverage; it is **access shape** — the relation was in the store with the wrong key.

H2-08 — naming a process from its description (tagged `same-shape`). *"What is the name of the metabolic process that takes place during periods of low glucose, glycogen exhaustion... ?"* The question's resolvable entities are glucose and glycogen, so the engine rendered glycogenolysis records. The answer, *ketogenesis*, exists in the store only as a **pathway tag** on seven reaction records (`[Ketogenesis]`) — a label on other facts, never a subject with its own defining conditions. Rank-1 sentence: *"Ketogenesis takes place in the setting of low glucose levels in the blood, after exhaustion of other cellular carbohydrate stores..."* Missing family: process definition — `(process: ketogenesis, defining conditions: low glucose, glycogen exhausted)`. Five of the seventeen misses are this one inversion: exam-1 always gave the process and asked for its parts; exam-2 gave the parts and asked for the process.

H.4 THE FIVE MISSING FAMILIES, PRICED ON THE FROZEN WORKLOAD

missing record family	exam-2 items it unlocks	ΔV
organ/tissue/cell localization	H2-05, 07, 14, 17, 19	5/19
process/state naming (description → name)	H2-01, 03, 08, 09, 15	5/19
hormone source and effect	H2-06, 10, 11	3/19
entity–disease/gene association	H2-12, 13	2/19
transport/carrier	H2-04	1/19

Five families cover sixteen of the seventeen misses. The seventeenth, H2-18 (*"fatty acids combined with glycerol to form what product?"*), is the exception that sharpens the rule: it *is* the native reaction shape, and the store even holds triglyceride-producing reaction records (`Diacylglycerol -> Triacylglycerol [Lipogenesis]`) — but the question's entities link to other records and the right one never rendered: an ordinary linking/assembly miss of the kind the campaigns repaired, not a schema gap. Retrieval retrodiction places the adjudicating sentence at rank 1 for 15 of the 17 misses (H2-15 at rank 5, H2-17 at rank 9): the evidence was one lookup away in every case; the record ontology was the boundary.

H.5 WHY THE AUTHOR WALKED IN, AND THE REMEDY

The exam author tagged seventeen of nineteen items `same-shape`, and by surface form the tags are defensible: *"what organ stores glycogen"* and *"what enzyme converts acetyl-CoA"* are the same interrogative template. But shape, in the sense that predicts scoring, is **which record family answers the question**, and fourteen of the seventeen surface-same items were mechanism-novel. The author could not see this because the schema was invisible from where exam authoring stood — the scope-blindness discipline (§5.1) *forbids* the author from asking what the store can express, precisely so that exams measure reach rather than fit. The trap is therefore structural, not careless: the discipline that protects the measurement also guarantees the author cannot steer around ontology gaps.

The remedy cannot be foresight — no schema designer enumerates localization, carriers, and definitional inversions in advance of a domain's questions, and the next domain will demand different families. The remedy is a **computable denominator**: under the receipt rule, every admissible exam question's relation is instantiated in some corpus sentence, so a relation census of the corpus is a candidate inventory (at declared granularity) of every class any receipt-ruled exam can demand. A pre-flight check then compares *curriculum coverage* (what the practice set demanded) against *census coverage* (what the corpus can ask): the metabolism campaign would have shown high curriculum coverage against low census coverage — the collapse signature, visible before exam-2 was authored — while fraud's ritual register would have shown the two coverages nearly coinciding, which is why its second exam opened at 0.77. That check, with the five families above as its first costed work order, is the designed next experiment (§6).