

Fall Through on Doubt: Pre-Registered Two-Domain Evidence for a Deterministic Serve Layer in Front of a Local LLM

Abstract

1 Introduction

2 System Model: the substrate contract

3 Experimental design

4 Results

5 Limitations

6 Related work

7 Discussion and future work

8 Conclusion

Reproducibility statement

Ethics statement

Appendix A — the traffic-scenario weight matrices

Appendix B — the domain-2 rule base and its mechanism sources

References

Vincent Grandjean — AME Labs — vincent.grandjean@ame-labs.eu

All data is synthetic: no real person, no real household, no real prescription, no real clinician anywhere in the evidence.

Abstract

A language model asked about a user's private record has two failure modes: it invents an answer, or — if well calibrated — it declines almost everything. Retrieval-augmented generation (RAG) narrows the gap but keeps the defining defect: when it fails, it still answers. We present and evaluate a deterministic **serve layer** that, placed in front of a local LLM, answers bounded-domain record questions directly from a curated knowledge graph — and **falls through to the model on any doubt**, at every stage, carrying its near-miss facts as grounding. The layer is stated as a seven-point substrate contract that any system can satisfy; we describe one instantiation.

The evaluation is pre-registered throughout: verdict bands were frozen before the exam fixtures existed, workloads and knowledge graphs are hash-pinned in git, per-cell outcomes were predicted in writing from the shipped code before every run, and both baseline arms ran under criteria frozen before the arms were built. Across two synthetic

domains — a domestic household and a medication-and-care record chosen for grammatical distance — the layer served **376 answers with zero wrong** (a count, never a pooled rate: the domains are never pooled, and the count sums the three measured legs, domain 1 plus domain 2 zero-shot and post-growth, which put the same questions to two configurations of the layer), over-routed **0 controls on every leg** (0/40 in domain 1, Wilson upper bound 8.8%; 0/40 on each of domain 2’s two legs), and separated **84 of 84** pre-authored near-neighbor question pairs — where retrieval over the *same facts in the same words* asserted between 39 and 44 wrong answers in domain 1 and between 55 and 63 in domain 2, each a range spanning the frozen judge and a lenient bound (\$4.6), and conflated 14 pairs. Zero-shot on the second domain, the layer served 36 questions — all correct — and refused the rest, with the fall-throughs sitting exactly where its vocabulary tables end; after a human wrote 159 vocabulary rows it reached exactly its domain-1 score, cell for cell, at 100% served-precision. The adaptation-cost criterion contained a designed falsifier, and it fired: two non-table code changes were required, and we quote the pre-named FAIL sentence verbatim. Against dense retrieval the correctness comparison is a statistical tie in domain 1 ($p = 0.694$ under the frozen judge; $p = 1.000$ under the lenient bound, whose point estimate straddles the layer) and an established win in domain 2 ($p = 0.0052$) — a lead robust in direction but, for the dense retriever, sitting at the edge of significance under that bound ($p = 0.0817$), a sensitivity we quantify with one symmetric audit applied identically to every arm of both exams, which moves every baseline and leaves the layer unmoved. Supplying the declined facts as grounding flips fall-through answers from wrong to right in both domains (exact McNemar $p < 0.001$ and $p = 0.003$). A served answer returns in about a millisecond at roughly three orders of magnitude ($\sim 10^3\times$) less energy than the model call it displaces, metered with hardware counters under a pinned power profile.

1 Introduction

Ask a modern LLM about the world and it is remarkable. Ask it about *your* world — “*Is the thermostat battery-powered?*”, “*Where is the metoprolol kept?*” — and something has to give. A bare model does not hold these facts. We measured both failure modes on the same exam: a local 7B model (mistral:7b-instruct, Apache-2.0) guesses its way to 75 of 210 household questions, while a stronger 9B (qwen3.5:9b, 9.7B parameters, Q4_K_M, Apache-2.0, thinking disabled) answers 14 correctly, asserts 17 wrongly, and declines the other 179 — because it correctly knows what it cannot know [D1]. Honesty is the better failure mode. It is still a failure mode.

Retrieval-augmented generation is the standard cure: embed the user’s facts, fetch the top-k into the prompt, let the model compose. It works — until it does not, and when it does not, *it still answers*. On our two exams, retrieval over the same facts in the same words asserted between 39 and 44 wrong answers about a household [D1][J1] and

between 55 and 63 about a synthetic medication record [E4] — each a range spanning the frozen judge and a lenient bound (§4.6) — fluently, and in one measured household case with every correct chunk already in its prompt [D1][E4]. Both the bare model and the retrieval pipeline also pay a full model call — the better part of a second of GPU time — for every question, including the ones a database lookup answers in about a millisecond for about a thousandth of the energy [E4].

Both options answer. When one of them refuses, it refuses by temperament rather than by an auditable rule — the 179 declines above are a property of that model’s calibration, not of a stated policy anyone can inspect. This paper is about closing that gap deterministically, and about measuring the closure so that its failure modes are as visible as its wins.

The claim, stated precisely. A deterministic knowledge-and-reasoning layer, placed in front of a local LLM, answers a substantial share of in-domain information questions exactly, auditably, in ~1 ms and at $\sim 10^3\times$ less energy than the displaced model call — **with zero wrong served answers** across every measured leg — while the LLM keeps fluency everywhere else. The claim is bounded: two curated domains at hundreds-of-facts scale, exam traffic, local models. §5 states every boundary we know of.

The experimental conditions, and the broader aim. The inference these experiments exercise is the narrowest the architecture offers — deliberately. The measured decision core is direct lookup plus **one** derivation family: forward rule application over curated causal implication rules, bounded at depth-2 — and under the frozen serve bar every measured serve is in fact depth-1 (§7). The host system’s inference engine registers **49 reasoning primitives across 11 categories** (deductive, relational, abductive, temporal, spatial, and non-monotonic among them); almost none of that breadth is exercised here, and none of it is claimed. The experiment isolates the property the wider neuro-symbolic program depends on first: that a symbolic substrate can *license* what it asserts — serve exactly, refuse honestly, ground what it declines — demonstrated on the smallest inference form that makes the question measurable. The designed trajectory is to widen the exercised inference set under the same licensing discipline, one pre-committed measurement at a time (§7); the numbers in this paper say nothing about that wider engine, by construction.

Contributions.

1. **A substrate contract (§2).** Seven properties — a typed schema, a provenance-carrying symbolic memory, a confidence rail, a serve seam, a grounding-and-audit surface, *fall-through-on-doubt*, and a bounded deductive depth — such that the measured results follow from the contract rather than from our codebase, under the KB-fidelity assumption §2 states. Any system satisfying the contract can host the evaluation; we describe one instantiation, and the generalization is a prediction from the contract’s structure, not a theorem.

2. **A pre-registered, eval-first evaluation protocol (§3)** — bands frozen before fixtures, both outcomes of every band pre-named, hash-pinned workloads, per-cell predictions derived from shipped code before each run, and byte-identity regression contracts — reusable for measuring any serve-vs-generate architecture.
3. **Two-domain evidence with a measured transfer boundary (§4.1–4.4)**: zero wrong served answers on every leg; a zero-shot stage decomposition locating the domain boundary exactly at the vocabulary tables; post-growth cell-for-cell parity with the home domain; and an adaptation cost priced by a criterion whose designed falsifier fired and is quoted firing.
4. **Baseline arms under frozen criteria (§4.5–4.7)**: a same-knowledge, same-words RAG head-to-head whose domain-1 tie does not survive transfer; a stronger-model arm with a hand-audited refusal decomposition; and a symmetric judge-sensitivity audit that bounds how much of every arm’s number is morphology.
5. **Grounding-bite evidence (§4.8)**: supplying the declined facts as grounding flips fall-through answers from wrong to right in both domains, established twice from independently derived populations.

2 System Model: the substrate contract

The serve layer is an architectural seam, not a standalone artifact — its claim lives in the linkages between a knowledge substrate and a language model. Describing our full system would make this an unreviewable systems paper; describing the layer without its substrate would make the results unreproducible. Following the design-theory position that boundary conditions are a mandatory component of a design contribution (Gregor & Jones 2007), we state the substrate as a **minimal contract**, under one assumption we make explicit because it is load-bearing: **the knowledge base is faithful to the domain** — its stored claims and rules are true of the world the questions ask about. A contract-satisfying system over a knowledge base of false facts serves wrong answers with exactly the same discipline; nothing in C1–C7 audits the facts themselves. Under that assumption the results in §4 are **predicted for any system satisfying these seven properties** — a prediction from the contract’s structure, established here on one instantiation, not a theorem — and we describe that instantiation. The instantiation is one slice of **AME-1**, a larger personal-assistant framework built specification-first (21 modules); the serve path crosses four of them — semantic extraction and routing, the symbolic memory stores, the per-claim confidence rail, and the reply composer, whose served-answer rendering is the template verbalizer of C4. Everything runs locally: models via Ollama, the embedder in ONNX, the knowledge in PostgreSQL. A general architecture overview is maintained at ame-labs.eu/ame-1/architecture; for this paper, this paragraph and the contract below are the system description.

C1 — Typed symbolic schema. A shared ontology (T-Box) with, at minimum: **typing** (`instance_of`), **containment** (`located_in`), and an **event-reification pair** (`event_type` / a participant predicate) so that n-ary happenings decompose into plain triples. The substrate *schema* is what transfers across domains; the *vocabulary* (attributes, properties, event types, states) is domain-specific. In our instantiation the sharing resides in the read orchestration, not in any grammar: the binding planner carries only an event-type surface and is predicate-agnostic [D2-A]. Typing is what lets the layer refuse — “*How many planets...?*” extracts as a valid question form and dies at binding, because `planets` is not a type the household ontology holds [D1].

C2 — Symbolic memory with provenance. A private, per-assistant claims graph (A-Box), queryable against the schema. A claim is a triple of subject, predicate, and object whose edge carries **polarity, source, and confidence** as first-class properties; events are reified as nodes. Provenance is what makes a served answer *auditable*: it cites the exact claim ids it stands on. Isolation is structural (per-assistant graphs), not promised.

C3 — Per-claim confidence and source-trust rail. Every stored claim and causal rule carries a confidence — in our instantiation hand-set and documented by ratified design decision, never fitted to an exam (the conformal-calibration slot exists and is deliberately empty; its dormancy marker is what proves the bar predates the exams). This rail is load-bearing arithmetic, not decoration: the serve law of C7 reads it directly.

C4 — A serve seam. A deterministic point in the reply path where a query is answered *from the knowledge base instead of invoking the model*. The served reply is rendered by a template verbalizer filled from the stored facts: **no model writes a served answer**. In our instantiation the seam sits after grounding assembly and before prompt composition.

C5 — A grounding and audit surface. Two distinct terms, both required. (a) *Fall-through grounding*: a refusal is not an apology — the near-miss facts the attempt turned up ride into the model’s prompt as an explicit grounding block. (b) *No reply is delivered un-gated*: served replies still pass a read-side consistency gate; an override is a defect plus an alarm counter, never a shipped wrong answer. Auditability rests on both.

C6 — Fall-through on doubt. The layer’s fail posture, and the property the results depend on most: **doubt, at any stage — no extraction, no unique binding, confidence under the bar, any error or timeout — degrades to exactly the system’s prior behavior** (the bare model plus whatever grounding C5 provides). It is the read-side inversion of a write-side admission gate: the write gate rejects on doubt; the reader falls through on doubt. We state this as a contract term rather than an implementation detail because the headline results are consequences of it and not of C1–C5: a system satisfying C1–C5 with a *serve-on-doubt* posture reproduces none of the zero-wrong or zero-over-route results below. A contract that under-determines the results it is quoted for is a description, not a contract.

C7 — A bounded deductive depth. Symbolic inference chains curated rules to a fixed maximum depth (here: two hops), composing confidence at each application, and prices

the result against a serve bar frozen before any exam existed (here: $\text{serve iff rule_conf} \times 0.95 \geq 0.80$ — the rule’s own confidence, discounted once per application — with the bar untouched throughout). Together with C3 this term *derives* measurable results in advance: it fixed the 10/30 chain-cell serve ceiling in both domains at design time, and measurement confirmed it in both (§4.1, §4.3).

The layer that instantiates this contract processes a question in five stages — extract (a grammar whose surface forms are vocabulary tables mapping English onto ontology ids), bind (resolution against the typed graph; ambiguity is a failure, not a choice), infer (forward rule application to C7’s bounded depth, composing confidence at each step), clear the bar (C3 + C7’s serve law), render (C4) — with C6’s fall-through at every exit. Figure 1 shows the seam with the contract terms annotated on the components they name; Figure 2 shows the entailment structure the seven-point cut exists for — which term produces which measured result. Table 1 maps each term to its instantiation and the evidence that exercises it.

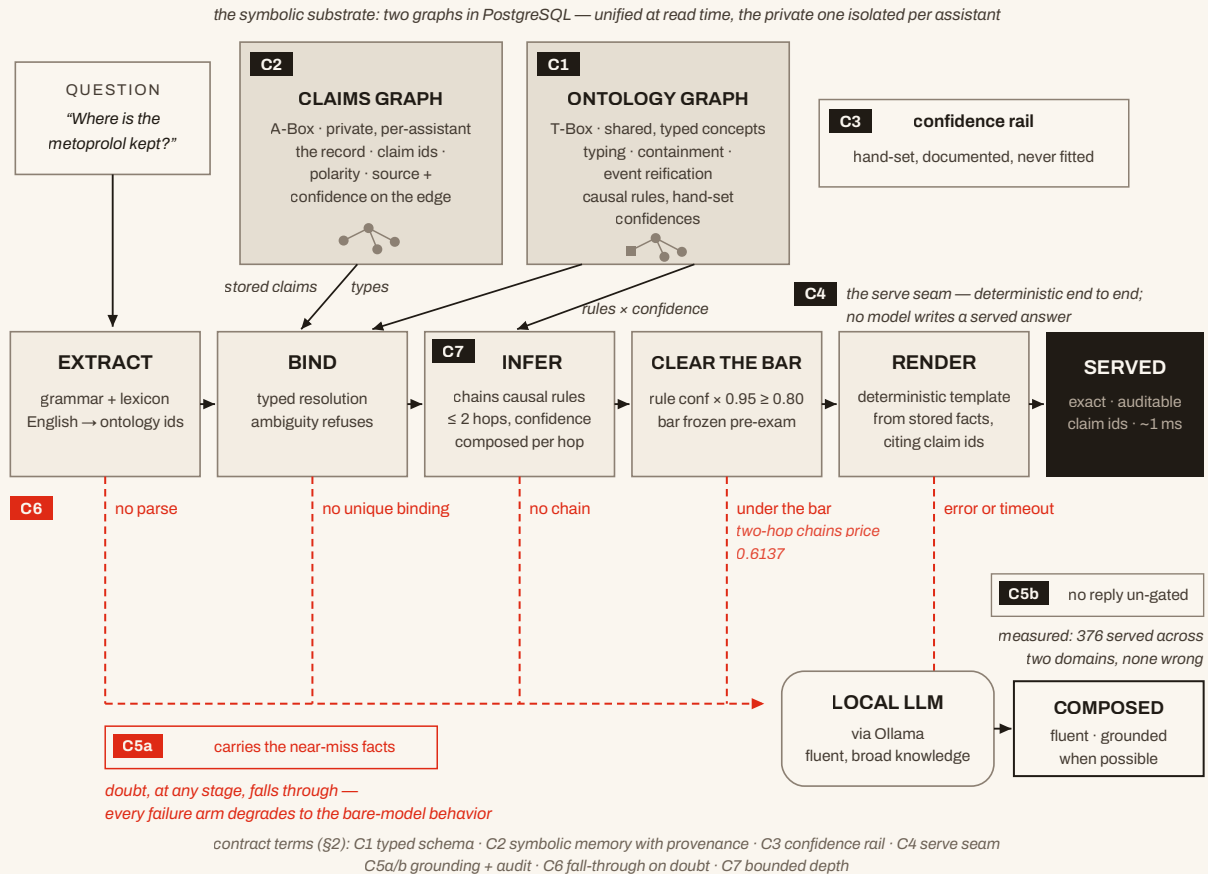
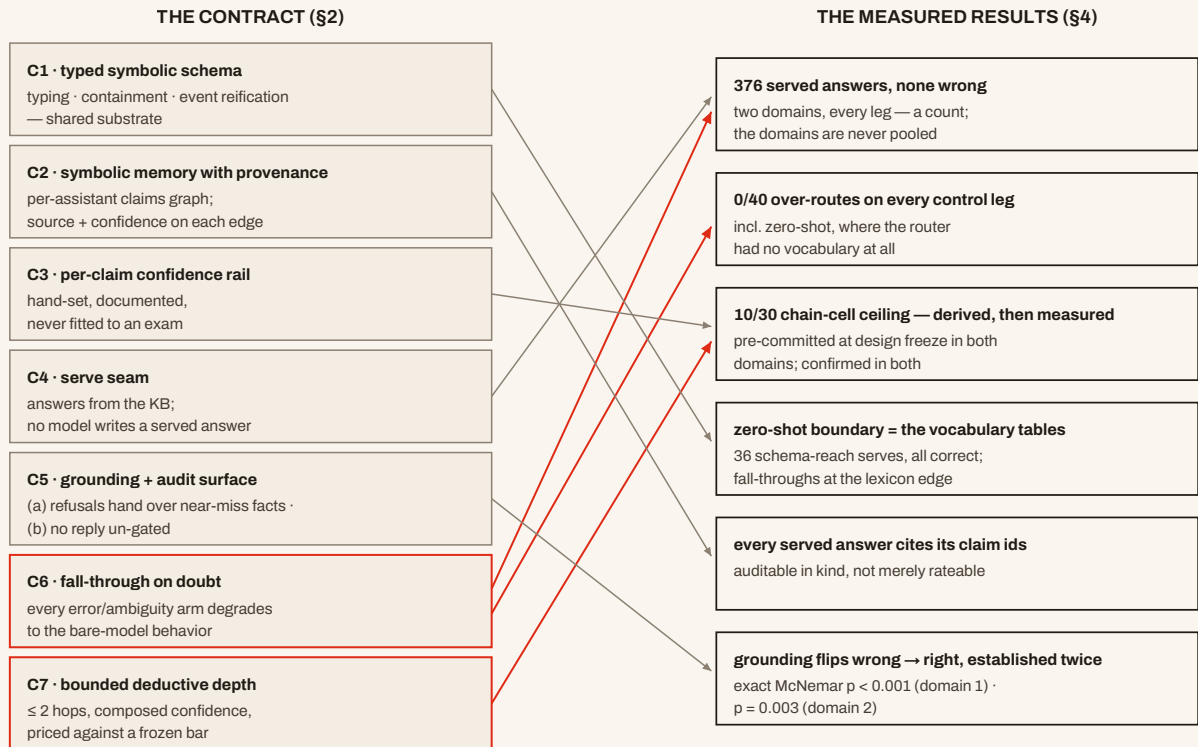


Figure 1 — the serve seam. Five deterministic stages, each with an honest exit; doubt at any stage falls through to the local LLM, carrying near-miss facts as grounding (C5a). No model writes a served answer (C4), and a served reply still passes the read-side consistency gate (C5b).



Each arrow runs from a contract term to a result it entails. The two highlighted terms are the ones a five-point cut omits, and the headline results (red arrows) are consequences of them; grey arrows carry the remaining terms to the results that exercise them. A system satisfying C1–C5 with a serve-on-doubt posture reproduces none of the top three boxes.

Figure 2 — the contract entails the results. The two highlighted terms, C6 and C7, are the ones a five-point cut omits — and the headline results are consequences of them, which is the reason they are contract terms and not implementation details.

Table 1 — the contract, its instantiation, and the evidence that exercises each term.

term	our instantiation	exercised by
C1 typed schema	shared <code>_ontology_global</code> T-Box: <code>instance_of</code> · <code>located_in</code> · event reification	zero-shot schema-reach serves; control refusals at binding [D2-A] [D1]
C2 symbolic memory	per-assistant claims graph; polarity/source/confidence on the edge	claim-id citation on every served answer [D1][D2-BC]
C3 confidence rail	hand-set rule/claim confidences; empty conformal slot as dormancy proof	the serve law; the 10/30 ceilings [D1][D2-BC]
C4 serve seam	reply-route provider after grounding assembly, before compose	376 served answers, template- rendered [D1][D2-A][D2-BC]
C5a grounding surface	near-miss facts $\rightarrow$ the prompt’s grounding block	the §4.8 probe, both domains [D1] [E4]
C5b audit gate	read-side consistency gate; override = defect + alarm	the write-side instrument set [W]
C6 fall-through on doubt	every error/ambiguity/timeout arm degrades to bare-model behavior	0 wrong served, 0/40 over-routes on every leg [D1][D2-A][D2-BC]
C7 bounded depth	$\leq 2 \text{ hops} \cdot \text{rule_conf} \times 0.95 \geq$ <code>0.80</code> , bar frozen pre-exam	all 20 two-hop refusals per cell, both domains [D1][D2-BC]

Two names, fixed here. We call the instantiated serve layer **the router** from §3 onward — it is the component that decides, per question, whether the knowledge base answers or the model does. And *knowledge base* and *knowledge graph* name one store throughout: C1’s typed schema together with C2’s claims graph.

3 Experimental design

Why an authored exam. No standard benchmark measures what this paper measures. A recent survey of some twenty RAG and KGQA benchmarks concludes that no end-to-end question-answering benchmark over a structured knowledge base with golden answers exists (Wedge et al. 2026 — a study funded by Neo4j, whose authors declare the funder had no role in its design, analysis, or interpretation). Two further constraints follow from the claim under test. The domains must be **private records**, because on public-knowledge material a bare model can answer from its weights rather than from any substrate — the same study reports exactly that inversion, its zero-shot arm beating both of its retrieval arms on correctness and relevancy over an open-domain encyclopedic corpus, and its authors note the result “would not be achieved for QA tasks on proprietary data”. And the questions must be authored against a **known** knowledge base, because whether a refusal was correct is only checkable when what the base holds is known exactly. Both exams are

therefore authored, pre-registered, and hash-frozen; §5 states what that costs and §7 states the held-out leg it owes.

3.1 The eval-first protocol

Every measurement in this paper ran under the same discipline:

1. **Criteria before instruments.** Each exam’s criterion — question cells, verdict bands, and pre-named sentences for *both* outcomes of every band — was committed before a single fixture, KG file, or code change for it existed ([C-D1] frozen 2026-08-12; [C-D2] frozen 2026-08-18; the arm criteria [C-RAG], [C-SM] frozen 2026-08-11 before any arm code; the grounding-bite criterion [C-F] frozen 2026-07-17). A band’s **pre-named sentence** is the exact wording each outcome will be reported in, written before the measurement that selects it exists; when we say a sentence **fired**, we mean the measurement selected that pre-written outcome and we print it verbatim rather than paraphrasing it. The headline bands are one frame throughout — under 10%, 10–15%, 15–30%, and 30% or above, the **pre-committed acceptance band** we call the GO band — carrying an overlay that voids every hit-rate unless served-precision reaches 95% and at least matches the LLM baseline on the same served turns. The frame was frozen before any measurement of this layer existed and has never been re-anchored.
2. **Frozen workloads.** Fixtures and knowledge graphs are git-hash-pinned at authoring time and verified before and after every session; a post-freeze edit voids and re-runs the full measurement. The pins are enumerated in the Reproducibility statement.
3. **Predictions before measurements.** Per-cell outcomes — including the *stage* at which each fall-through would occur — were derived from the shipped extraction and binding code (via a read-only sweep tool) and recorded in writing before each run.
4. **Byte-identity regression contracts.** Every additive instrument (both baseline arms, the grounding probe) is required to leave a plain router evaluation byte-comparable to its frozen artifact; growth sessions re-prove the home domain byte-identical after vocabulary changes.
5. **An adversarial instrument review preceded every measurement session**, raising 17–19 findings each time, and every action-required finding was fixed before any number was read.

Figure 3 shows the protocol as a sequence; Table 2 is the ledger a reviewer can check it against.

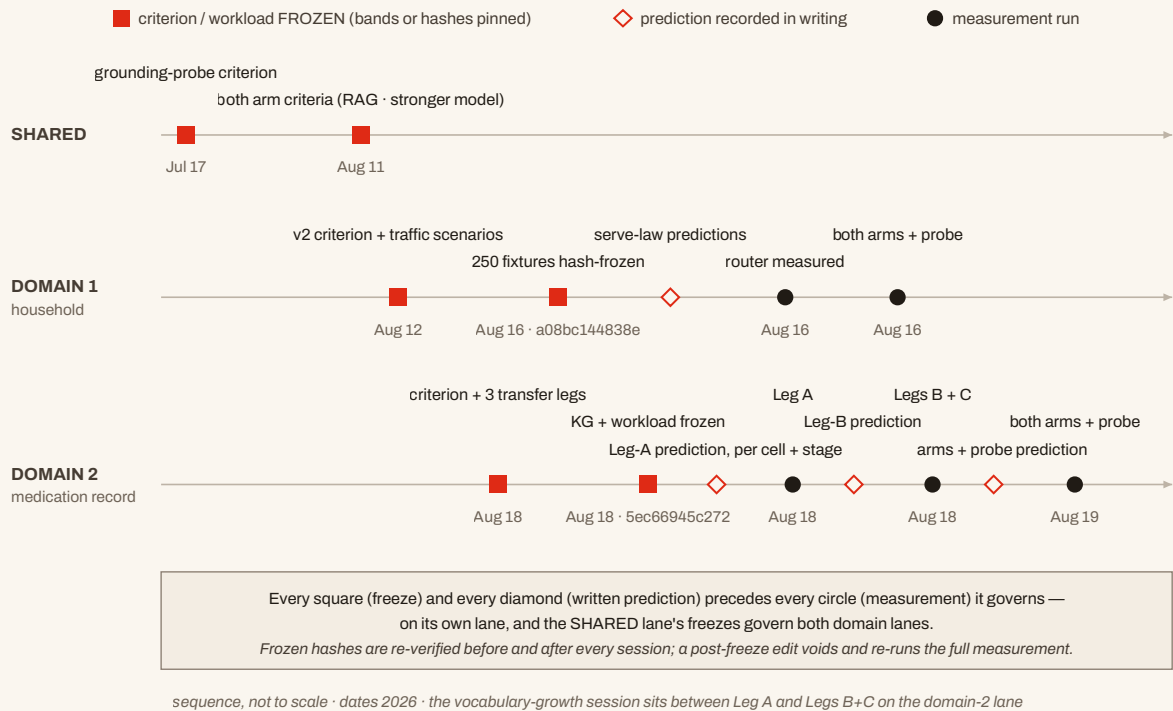


Figure 3 — the protocol as a sequence (not to scale). On every lane, every freeze (square) and written prediction (diamond) precedes every measurement (circle) it governs.

Table 2 — the freeze ledger: what was frozen, when, and what it preceded.

artifact	frozen	precedes
grounding-probe criterion [C-F] — arms, corruption rule, bands	2026-07-17	the first probe LLM call
RAG arm criterion [C-RAG] — corpus recipe, retrievers, k = 5, bands	2026-08-11	any arm code
stronger-model criterion [C-SM] — model pin, bands, energy leg	2026-08-11	any arm code
domain-1 v2 criterion [C-D1] — cells, bands, five scenarios, gold rules	2026-08-12	any v2 fixture, code, or measurement
domain-1 v2 workload — 250 fixtures + KG extension, tree a08bc144838e ; serve-law predictions recorded at the freeze	2026-08-16	the v2 answer-shape code being measured
domain-2 criterion [C-D2] — domain, scope law, three transfer legs, bands	2026-08-18	any domain-2 fixture, KG file, code, or measurement
domain-2 workload + KG — tree 5ec66945c272 (KG 63ab8f4834cc); Leg-A per-cell and per-stage predictions recorded at the freeze	2026-08-18	Leg A
Leg-B per-cell prediction — derived from the shipped, grown grammar	2026-08-18	the first Leg-B measurement call
E4 prediction — 7 directional claims + 10 point predictions	2026-08-19	the first arm measurement call
shadow-probe pre-registration (observational; §7)	2026-08-19	the probe instrument existing

3.2 Domain 1: the household exam

The first domain is the synthetic domotic household the system was developed against: 46 devices, 194 stored facts (including 24 reified past-tense event records and 8 explicit stored negatives), 16 curated causal rules. The exam (v2) is 250 questions: 30 in each of seven measured cells and 20 in each of two control cells, giving **210 banded questions** and a raw ceiling of 84.0%. The cells are anchored in a verified question taxonomy (Graesser & Person 1994, checked against the original scan; every fixture declares its category) — Table 3:

Table 3 — the workload cells (identical structure in both domains).

cell	n	content	Graesser category
private_polarity	30 (15 YES / 15 NO)	yes/no over stored facts	Verification
private_lookup	30	attribute values	Feature specification / Quantification
private_relational	30	where is X / relations	Concept completion
private_inventory	30	counting + which-container aggregation	Quantification / Concept completion
event_history	30 (15/15)	did event E ever happen over reified records	Verification
grounded_chain	30 (20 one-hop / 10 two-hop)	causal consequence, hybrid grounding	Causal consequence
generic_chain	30 (20/10)	causal consequence, no private grounding — the bare-LLM-parity control	Causal consequence
llm_required	20	general knowledge — over-route control	Instrumental/ procedural
out_of_domain	20	out-of-scope — over-route control	(topic-defined)

Design decisions that matter for reading §4: the `event_history` NO-half is served by a **closed-world rule bounded to recorded events** (“no record of...” — a claim about the records, not the world); the two chain cells are shape-identical apart from private grounding, emitted from one table, so `generic_chain` is a true knowledge-source control; all 10 two-hop questions per chain cell are **designed refusals** under the frozen bar (§2 C7 — best two-hop confidence product $0.6800 \times 0.95^2 = 0.6137 < 0.80$); and the exam carries **52 same-cell near-neighbor distractor pairs**.

v1 as design history. A first exam (N = 120, measured 58.3% [49–67%] hit-rate at 100% [95–100%] served-precision against a 72.5% answerable ceiling) was retired from results to methods when the v2 criterion froze, for three measured defects: an inverse cell structure, an unreported 67% polar guess floor, and golds authored in KB surface forms — the source of the judge asymmetry that v2 cures at the data layer (§3.5). v1 numbers are never compared band-to-band with v2 numbers, and are not evidence in this paper.

3.3 Domain 2: the medication record and the three transfer legs

Five of domain 1’s seven measured cells sit at ceiling (§4.1), so domain 1 cannot distinguish a router that generalizes to a new domain from one built around its own KG. Domain 2 exists to measure that distinction, and its criterion [C-D2] was frozen before any domain-2 fixture, KG file, code change, or measurement.

The domain is a **personal medication and care record for one synthetic individual**: 26 invented medications with real published mechanisms, 5 storage containers, 4 prescribers, 24 reified care events (refill, administration, prescription change), 66 T-Box

concepts (27 causal-state), 16 causal rules each carrying a source note (the mechanisms and their published sources are tabulated in Appendix B), and 341 stored fact triples (293 ordinary A-Box facts + 24 events at 2+ triples each). It was chosen for **grammatical distance** (the attribute/event lexicon is disjoint from domotics — that distance is the transfer evidence), for **near-neighbor supply** (metoprolol tartrate vs succinate; two strengths of levothyroxine; brand vs generic inhalers), and because the stakes argue the thesis: on a medication record, a wrong served answer is self-evidently worse than a refusal. **All data is synthetic. The exam asks about the record and about published mechanisms — never what a person should do;** advice and deontic forms are out-of-grammar and refuse by construction, and the advice temptation is cured by rewording (“*Can I take X with grapefruit?*” is out-of-grammar; “*What does taking X with grapefruit cause?*” is a causal lookup over a published mechanism).

The workload inherits the v2 structure verbatim (250 questions, same nine cell names, same polar balance, same 20/10 hop mix, 32 distractor pairs) with one deliberate arithmetic control: **the rule base’s 16-value confidence multiset mirrors domain 1’s element-for-element**, so the chain-cell serve ceiling is 10/30 in both cells *by construction* — any chain-cell difference between the domains is attributable to transfer, never to arithmetic. Table 4 puts the two domains side by side: matched shape, mirrored arithmetic, disjoint vocabulary.

Table 4 — the two domains: matched shape, mirrored confidences, disjoint vocabulary.

	domain 1 — household	domain 2 — medication record
nature	fully synthetic; one invented household (maison_jasper)	fully synthetic; one invented patient
instances	46 devices, rooms	26 medications · 5 containers · 1 patient · 4 prescribers
stored fact triples	194 (incl. 24 reified events, 8 stored negatives)	341 (293 ordinary + 24 reified events)
causal rules	16	16 — confidence multiset mirrored element-for-element
two-hop chains	6 depth-2 paths	5 depth-2 paths (the one structural non-mirror, disclosed)
distractor pairs	52	32
traffic scenarios	5	4 — assistant_anchored empty, reason stated (§3.4)
workload	250 questions · 210 banded · the nine cells of Table 3	identical structure, verbatim
persona sentence	household assistant	medication-record assistant — a parameter, same shape, pinned in the criterion

The three transfer legs, with their sequencing law:

- **Leg A — zero-shot.** The router measured on the frozen domain-2 workload with its vocabulary tables *exactly as domain 1 left them*, before any domain-2 vocabulary edit; the vocabulary files’ git blob hashes are pinned in the Leg-A report. The criterion **forbids quoting Leg A as a hit-rate**; it reports, per cell, the serves and the fall-through stage decomposition (extraction / resolution / low-confidence).
- **Leg B — post-growth.** The full exam was re-measured after vocabulary growth. Leg B carries domain 2’s verdict bands.
- **Leg C — adaptation cost.** The pre-committed tables-as-data claim itself — “*adapting the router to a new domain is a data edit, not a code change*” — measured as the git diff from the Leg-A-pinned state to the Leg-B state. **The falsifier is designed in:** any hunk outside a constant table counts against the claim and is named, and the criterion pre-names a verbatim FAIL sentence for that outcome.

3.4 Traffic scenarios and the weighted headline

A raw hit-rate over an authored cell mix measures the author’s mix. The criterion therefore pre-commits **sourced traffic scenarios** — each mapping a published distribution onto the nine cells under a declared mapping law (uniform-within-mapped-group as the null; out-of-domain weighted only by topic-axis sources; form-authoring pins) — and requires the conclusion to be reported under **all** scenarios. Domain 1 has five: `conversation_anchored` (Stivers 2010; polar dominance), `assistant_anchored` (Bentley et al. 2018, engagement-proportional reading), `reference_anchored` (the reference-desk studies of Arnold & Kaske 2005 and Radford & Connaway 2013, ready-reference share pinned at the harsh-to-thesis 33%), `uniform`, and `jasper_house` (the mechanical element-wise mean of the three sourced scenarios — no editorial profile exists). Domain 2 has four: the three transferable scenarios with **weight columns numerically identical to domain 1’s** (a pinned instrument property), plus `med_record` (the mechanical mean of its two sourced scenarios). **The `assistant_anchored` slot is empty in domain 2, with the reason stated:** Bentley is smart-speaker-specific and says nothing about a medication-record assistant; a weak anchor is worse than a named absence.

Two rules bind every weighted number in this paper: **a weighted headline never appears without its scenario’s weighted ceiling and weighted guess floor** (the ceiling is the weight mass on measured cells; the floor is what a constant-answer model earns on the balanced polar cells), and **the two domains are compared only under the three shared scenarios — never pooled**. The per-cell transfer table (§4.3) is the primary transfer evidence, weight-free. The complete weight matrices, with each scenario’s derivation, ceiling, and floor, are reproduced in Appendix A so every weighted headline in §4 is recomputable from this paper alone.

3.5 The judge

All arms are scored by one byte-frozen judge: `answer_matches` requires the gold answer as a contiguous token subsequence with exact morphology, set-semantics on enumerations. Morphology tolerance is provided **at the data layer, symmetrically**: each gold carries an accept-list (morphological number, hyphenation, spelled numerals), the runner ORs the judge over $\{\text{gold}\} \cup \text{accept}$, and router, RAG, and LLM arms are judged under the same lists. The direction is disclosed: the router’s renderer matches the gold by construction, so accept-lists only ever help the baselines. A 96-probe paraphrase audit gate ran before the workload froze (every failure cured by extending accept-lists, never the judge; each cure proven against both a render test and an inverted-predicate test). One residual gap is disclosed in §5: polar YES-golds cannot be made tolerant, and that gap runs *against* the baselines. §4.6 quantifies the judge’s remaining asymmetry with a symmetric lenient re-judge.

The judge errs in the other direction too, and that error is measured rather than assumed. Because a gold counts when it appears as a contiguous token subsequence, a reply that merely *mentions* the right value can score correct without asserting it. A blind hand-audit of thirty judge-correct retrieval replies — drawn by a stride rule over the sorted fixture ids, fixed in writing before any reply was read, fifteen per domain — found **one** such false positive: 3.3% (Wilson 95% [0.6%, 16.7%]), or 13.3% [5.3%, 29.7%] if modal phrasing in causal answers is also counted as hedging [FP]. The error runs in the **baseline’s** favor and cannot run in the router’s, whose served answers are rendered from the gold surface, so every router margin reported below is conservative to about that degree.

3.6 The baseline arms

The two arms vary different levels of the stack, in the sense of Chen et al. (2026): the retrieval arm varies the *embedded* level — the resources placed in front of the same composer — while the stronger-model arm varies the *trained* level, the model itself.

RAG (the head-to-head, built to be losable). The retrieval arm gets the *same knowledge in the same words*: every stored fact verbalized through the router’s own production render table, every rule in the same bullet form the router’s grounding uses; T-Box concepts excluded (they answer no question); **no pre-composed derivation chains** (in-context composition is exactly the burden the structural argument says retrieval carries). Corpus: 210 chunks in domain 1 (194 facts + 16 rules), 357 in domain 2 (341 + 16), each SHA-256 hashed and persisted. Two retrievers: dense BAAI/bge-base-en-v1.5 (Xiao et al. 2024; ONNX, CLS pooling — the same embedder family the live system trusts) and BM25 ($k_1 = 1.2$, $b = 0.75$); **the headline is the better of the two, disclosed** — the generous-to-RAG posture. Retrieved chunks are injected verbatim through the same prompt surface the router’s grounding uses. **$k = 5$ is the banded configuration, committed before any run**; $k \in \{1, 3, 10\}$ is an observational sweep (disclosure, never tuning — and in both domains the sweep shows $k = 10$ would have scored better, so the pinned k is

demonstrably not chosen to win). Composer: `mistral:7b-instruct` (Jiang et al. 2023), $t = 0$, single pass. RAG has no refusal path and pays embed + search + a full LLM call on every turn — zero displacement by construction.

Stronger model. *“Is the LLM-alone baseline an artifact of a 7B model?”* The arm re-runs the LLM passes only, against `qwen3.5:9b` (Qwen Team 2026; 9.7B, Q4_K_M, Apache-2.0, hybrid thinker with **thinking pinned off on the wire** — reasoning tokens break $t = 0$ comparability). Quantization, size, and license are quoted beside every number. Transport errors abort a run; an error is never scored as a wrong answer. One disclosure about the 7B comparator, which appears in both domains as a reference point: its per-case replies were never persisted — only its aggregate score survives in the router-evaluation reports — so that arm alone cannot be re-scored offline, and it is absent from the judge-sensitivity audit of §4.6 for that reason [J1].

3.7 The grounding-bite probe

Faithfulness $\neq$ retrieval: putting a fact in the prompt does not mean the model uses it. On every banded fall-through that carries a grounding block, three arms run with one variable (same model, same template, $t = 0$): **BARE** (block removed), **TRUE** (block verbatim as the router produced it — reusing the existing evaluation replies, never re-run), **CORRUPTED** (exactly one bullet’s effect swapped for a false-but-plausible state, by a deterministic rule that never injects a KB truth; every skip is surfaced explicitly, and any skip voids the bite verdict). Serve-path faithfulness is trivially 100% by construction (the KB substance *is* the served answer) and is not probed. Bands: helped/hurt with exact McNemar for the lift; a follow-rate threshold ladder for the bite, both outcomes pre-named.

3.8 Statistical apparatus

Wilson 95% intervals beside every rate; exact two-sided McNemar (binomial on discordant pairs, $p_0 = \frac{1}{2}$) for paired comparisons; the weighted headline is $\sum w_i \cdot \text{rate}_i$ with variance $\sum w_i^2 \cdot p_i(1-p_i)/n_i$. No number is quoted bare: weighted headlines carry ceiling and floor; model numbers carry quantization, size, license; energy carries its referent and is quoted as a magnitude only. Zero counts carry their Wilson upper bound at first mention, since a bare zero over a small denominator is the easiest number in the paper to over-read. **No family-wise correction is applied**, and the reason is structural rather than convenient: every comparison reported here was named in a criterion before the measurement existed, so there is no post-hoc selection over a family of tests to correct for; the p-values are individual tests of pre-registered comparisons and are read as such.

4 Results

4.1 Domain 1: it works — and five cells at ceiling is why it cannot settle generality

170/250 served · 170/170 correct · 0/40 over-routes (Wilson upper bound 8.8%) · 52/52 distractor pairs separated (Wilson lower bound 93.1%) [D1]. Per cell: the five lookup-family cells are 30/30 each; both chain cells are 10/30 — exactly the ceiling the serve law pre-committed, with all 20 two-hop questions refused and the 10/20 one-hop split landing on precisely the rules whose confidence clears $\text{rule_conf} \times 0.95 \geq 0.80$. Every pre-recorded prediction — counts and the composition arithmetic — was confirmed exactly.

The weighted headline clears the pre-committed GO band (§3.1) under **all five** scenarios [D1] — Table 5:

Table 5 — domain 1: the weighted headline under all five traffic scenarios.

scenario	weighted headline [95%]	ceiling	guess floor
conversation_anchored	89.1% [87.9–90.2%]	95.3%	35.9%
assistant_anchored	38.0% [34.2–41.8%]	56.4%	5.8%
reference_anchored	47.9% [42.6–53.2%]	77.7%	6.6%
uniform	63.0% [60.3–65.6%]	77.8%	11.1%
jasper_house	58.3% [55.0–61.6%]	76.5%	16.1%

The spread (38 → 89%) is driven almost entirely by each scenario’s *ceiling*, not by router performance — `assistant_anchored` is lowest because the honest reading of Bentley puts most assistant traffic outside the user’s own house (ceiling 56.4%; the router reaches 67% of what the scenario allows). The pre-named band sentence fired verbatim: “*the weight choice is not load-bearing for the conclusion.*” Note that `jasper_house` = 58.3% is a numerical coincidence with the retired v1 exam’s 58.3% headline — a different exam, a different quantity, a different weighting; the two are not continuous and must not be read as though they were.

The result that carries the paper’s central claim is the precision gate: **0 wrong served answers in 170** (100.0% [98–100%]), on an exam that includes a closed-world answer shape — asserting the *absence* of an event record — that is the shape most able to be confidently wrong. The 40 controls all fell through (35 at extraction, 5 at resolution — the five nearest approaches all being “how many X” forms whose type surface the household KB does not hold: the grammar generalizes to the form; the binding layer refuses).

The 40 *banded* fall-throughs are the two chain cells’ designed refusals, with one wrinkle §4.8 depends on: in each chain cell, 19 of the 20 fall at the confidence bar and carry a near-miss grounding block, while one falls earlier, at resolution, and carries none. That is why the grounding probe’s domain-1 population is 38 and not 40.

But **five of the seven measured cells sit at ceiling**. An exam this saturated cannot tell you *why* it was passed: every vocabulary row was authored with this household in view, so domain 1 cannot distinguish a router that generalizes to a new domain from one built around its own KG. That question is domain 2's, and it was designed as its own measurement with both verdicts pre-named.

4.2 Domain 2 zero-shot (Leg A): the boundary sits exactly where the tables end

The criterion forbids quoting Leg A as a hit-rate; the finding is the **stage decomposition** [D2-A]. Pointed at the frozen medication-record workload with its vocabulary tables byte-identical to the state domain 1 left them in (blob-hash-pinned), the router:

- **served 36 questions — 36 correct** (100.0% [90–100%]);
- **over-routed 0 of 40 controls** — the criterion calls a zero-shot over-route “the worst outcome this exam can produce”; there were none;
- **fell through on the remaining 174 banded questions at exactly the stages the shipped grammar predicted: 106 at extraction** (words the vocabulary tables do not hold — the router never reaches the record), **28 at resolution** (forms that extract on generic English and bind nothing), and **40 at the confidence bar** (the chain cells’ designed arithmetic, identical to domain 1 by construction).

Every cell’s count *and* its fall-through stage were predicted in writing from the shipped code before the run, and every one was confirmed exactly — Table 6, visualized in Figure 4. The pre-named Expected sentence fired verbatim: *“interrogative forms and substrate schema transfer; the lexicon does not — by design and now by measurement.”*

Table 6 — Leg A: predicted vs measured, per cell and per stage. Every row was recorded at the workload freeze, before the leg could run.

cell	n	predicted	measured	predicted stages	measured stages
private_polarity	30	0/30	0/30 ✓	extraction 30	extraction 30 ✓
private_lookup	30	0/30	0/30 ✓	extraction 30	extraction 30 ✓
private_relational	30	16/30	16/30 ✓	extraction 3 · resolution 11	extraction 3 · resolution 11 ✓
private_inventory	30	0/30	0/30 ✓	extraction 13 · resolution 17	extraction 13 · resolution 17 ✓
event_history	30	0/30	0/30 ✓	extraction 30	extraction 30 ✓
grounded_chain	30	10/30	10/30 ✓	low-confidence 20	low-confidence 20 ✓
generic_chain	30	10/30	10/30 ✓	low-confidence 20	low-confidence 20 ✓
llm_required	20	0/20	0/20 ✓	extraction 20	extraction 20 ✓
out_of_domain	20	0/20	0/20 ✓	extraction 18 · resolution 2	extraction 18 · resolution 2 ✓
total	250	36/210	36/210 ✓		

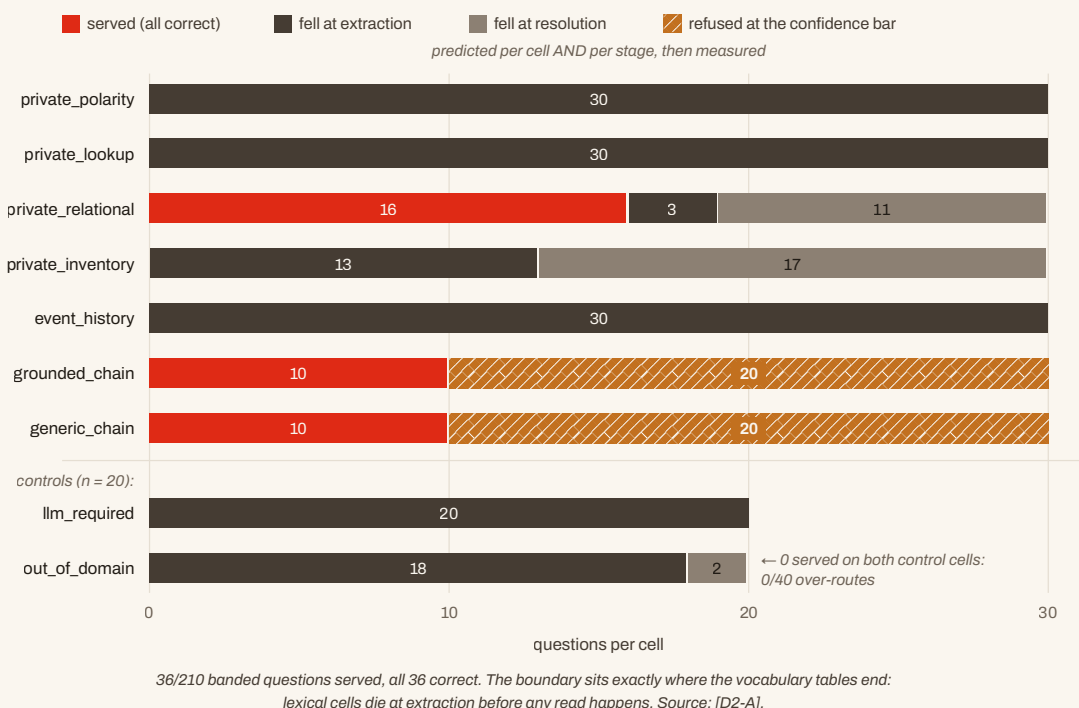


Figure 4 — the zero-shot decomposition as stacked stages. Lexical cells die entirely at extraction — before any read happens; inventory splits between extraction and resolution for two different reasons with two different cures; the chain cells serve exactly their arithmetic ceiling; both control cells fall through completely.

The 36 serves sit precisely where shared substrate reaches: 16 locative questions riding the shared `located_in` predicate plus generic English (“Where is X kept?”), and 10 causal serves per chain cell — the arithmetic ceiling, zero-shot. The 16 relational serves bound through instance ids alone (no synonym rows exist for domain 2), which is what makes them attributable to the architecture rather than to naming luck. The three distractor pairs that were evaluable zero-shot — a salt-form pair, a two-strengths pair, a brand pair — all separated.

What Leg A establishes: at maximum vocabulary mismatch, the router asserted nothing wrong and claimed no authority it did not have. A system that degrades by guessing would have produced its wrong answers here. **What it does not establish:** anything about usefulness in domain 2 — 36/210 is the zero-shot floor, and the criterion forbids reading it as a capability.

4.3 Domain 2 post-growth (Leg B): cell-for-cell parity

After vocabulary growth (§4.4), the full exam was re-measured [D2-BC]:

170/250 served · 170/170 correct · 0/40 over-routes · 32/32 distractor pairs separated (Wilson lower bound 89.3%) — the same 170/210 domain 1 reaches, cell for cell (Table 7, Figure 5).

Table 7 — the per-cell transfer table (the primary transfer evidence, weight-free; every cell predicted before it was measured, twice):

cell	D1 v2 measured	D2 Leg A predicted	D2 Leg A measured	D2 Leg B predicted	D2 Leg B measured
private_polarity	30/30	0/30	0/30	30/30	30/30
private_lookup	30/30	0/30	0/30	30/30	30/30
private_relational	30/30	16/30	16/30	30/30	30/30
private_inventory	30/30	0/30	0/30	30/30	30/30
event_history	30/30	0/30	0/30	30/30	30/30
grounded_chain	10/30	10/30	10/30	10/30 (pre-committed)	10/30
generic_chain	10/30	10/30	10/30	10/30 (pre-committed)	10/30



Figure 5 — Table 7 as a picture. In every cell the post-growth dot lands exactly inside the home-domain ring; zero-shot, only the shared-substrate relational serves (16) and the arithmetic chain ceiling (10 per cell) exist at all. The boundary moves by writing vocabulary rows.

The weighted headline clears the GO band under **all four** domain-2 scenarios [D2-BC] — Table 8:

Table 8 — domain 2 (Leg B): the weighted headline under all four traffic scenarios.

scenario	weighted headline [95%]	ceiling	guess floor
conversation_anchored	89.1% [87.9–90.2%]	95.3%	35.9%
reference_anchored	47.9% [42.6–53.2%]	77.7%	6.6%
uniform	63.0% [60.3–65.6%]	77.8%	11.1%
med_record	68.5% [65.3–71.7%]	86.5%	21.3%

Under the three shared scenarios — the only license for cross-domain comparison — the two domains’ weighted headlines are **numerically identical**, because the per-cell counts are identical and the shared weight columns are pinned identical. `med_record` (68.5%) is domain 2’s own ecological claim and has no domain-1 counterpart. The pre-named PASS sentence fired verbatim: “...*the weight choice is not load-bearing, and neither is the domain.*”

The precision sentence completes across both legs: **206 served turns in a domain the router was not built for, and not one wrong answer** — and the under-route decomposition collapses to a single arm: every banded fall-through at Leg B is one of the 40 pre-committed confidence-bar refusals. The control decomposition is byte-identical to Leg A’s: growth did not widen the door. A later session re-ran the full exam and reproduced it with **zero of 250 per-case rows differing** — including, to the digit, the stochastic LLM leg [E4].

What cell-for-cell parity is and is not evidence of. The same saturation §4.1 discloses applies here: five of the seven measured cells sit at ceiling in *both* domains, so once a cell clears at all, matching the home domain in that cell is close to guaranteed. The parity’s content is therefore the *clearing*, not the coincidence of the counts — that the grown vocabulary reaches every cell of a lexically disjoint domain, with the two chain cells landing on an arithmetic ceiling that was fixed by construction (§3.3) rather than reached by performance.

4.4 The adaptation cost (Leg C): the falsifier fired, and we quote it

Adapting the router to domain 2 cost **159 vocabulary-table entries across 9 tables in 2 source files** — Table 9 [D2-BC]. Two of the four pinned files did not move at all: the render table needed **zero** new rows (its humanized fallback had already passed the frozen render audit 210/210), and the synonym table for instances gained zero rows because every instance id is the underscore-join of the phrase a question uses, making the id itself the route.

Table 9 — the adaptation cost, itemized. The diff runs from the blob-pinned Leg-A grammar state to the state Leg B was measured on.

vocabulary table	rows added	carries
EVENT_PREDICATE_SURFACES	41	the 13 domain-new event types, with synonyms
TYPE_SYNONYMS	42	the record's whole T-Box type vocabulary
PROPERTY_SURFACES	21	the four complementary polar pairs, both members
ATTRIBUTE_SURFACES	20	strength · ingredient · form · frequency · prescriber · brand · interaction
CONTAINER_NOUN_SURFACES	12	new table — the place noun
RELATIONAL_VERBS	8	the care relations
VERB_PROPERTY_SURFACES	7	the polar pairs in the <i>does/can</i> voice
ROOM_SYNONYMS	6	the five storage containers
RELATIONAL_PARTICIPLES	2	new table — the passive relational form
total rows	159	in 2 source files
DEVICE_SYNONYMS	0	the instance-id law makes the id itself the route
the answer render table	0	the humanized fallback already passed the 210/210 render audit
non-table code hunks	2	readers of the two new tables; each deletes a hardcoded English literal — the falsifier's subject

And the pre-committed falsifier fired. The criterion pre-named a FAIL sentence for any change outside a constant table, and the data required it; it was fired verbatim rather than softened:

“Adaptation required 2 non-table changes, named: (1) the container-noun table's shared reader, replacing a hardcoded English noun in two function bodies; (2) the relational-participle table's reader in the copular locative branch, replacing a hardcoded predicate — the claim is weakened to that exact extent, and we say so.” [D2-BC]

Read in both directions: five of the seven measured cells — 138 questions' worth — were reached by table rows alone, and 156 of the eventual 170 were measured served before either code change existed. Both changes are one finding twice (an English literal hardcoded *beside* a table), each change *deletes* the literal, and the cost is therefore one-time — but they were code, and softening the sentence would have defeated the reason the falsifier exists. Whether a third domain costs zero code is a claim this measurement cannot make, and does not.

Two measured guards. The 159 rows plus 2 hunks moved domain 1 by **zero routes** — proven by reverting the vocabulary files to their pinned blobs, measuring domain 1 (both exams), restoring growth, and measuring again: byte-identical reports, three times over [D2-BC]. And one ratified finding keeps the leg honest, in the project’s own words: every one of the 159 entries was mechanically derivable from the seeded ontology, but they are hand-authored *beside* the KB, not generated *from* it — “**tables-as-data**” today means data a human writes, not data the system derives (§5).

4.5 The retrieval baseline: a tie at home, a loss under transfer

Every margin in this subsection carries §4.6’s caveat: the frozen judge’s morphology strictness favors the router, and §4.6 quantifies that sensitivity for **both** domains, under one mechanical lenient re-judge applied identically to every arm of both exams. Table 10 is the one-look summary of every arm in both domains — per-domain columns, deliberately never pooled — and Figure 6 plots it.

Table 10 — the arms across both domains (weight-free; the domains are never pooled). Wrong-answer counts carry their judge range from §4.6; every other number is under the frozen judge, the only banded one.

	domain 1 v2	domain 2
router	170/210 @ 170/170 served-correct	170/210 @ 170/170 served-correct
RAG (bge, k=5)	166/210 · 39–44 wrong · p = 0.694 (tie)	147/210 · 55–63 wrong · p = 0.0052 (win)
RAG on the router-served turns	139/170 vs 170/170	127/170 vs 170/170
near-neighbor pairs	52/52 router · 40/52 RAG (9 conflations)	32/32 router · 25/32 RAG (5 conflations)
stronger model (qwen3.5:9b)	14/210 · 10/170	29/210 · 28/170
grounding probe (§4.8)	N = 38, 0 skips, p < 0.001	N = 40, 0 skips, p = 0.003

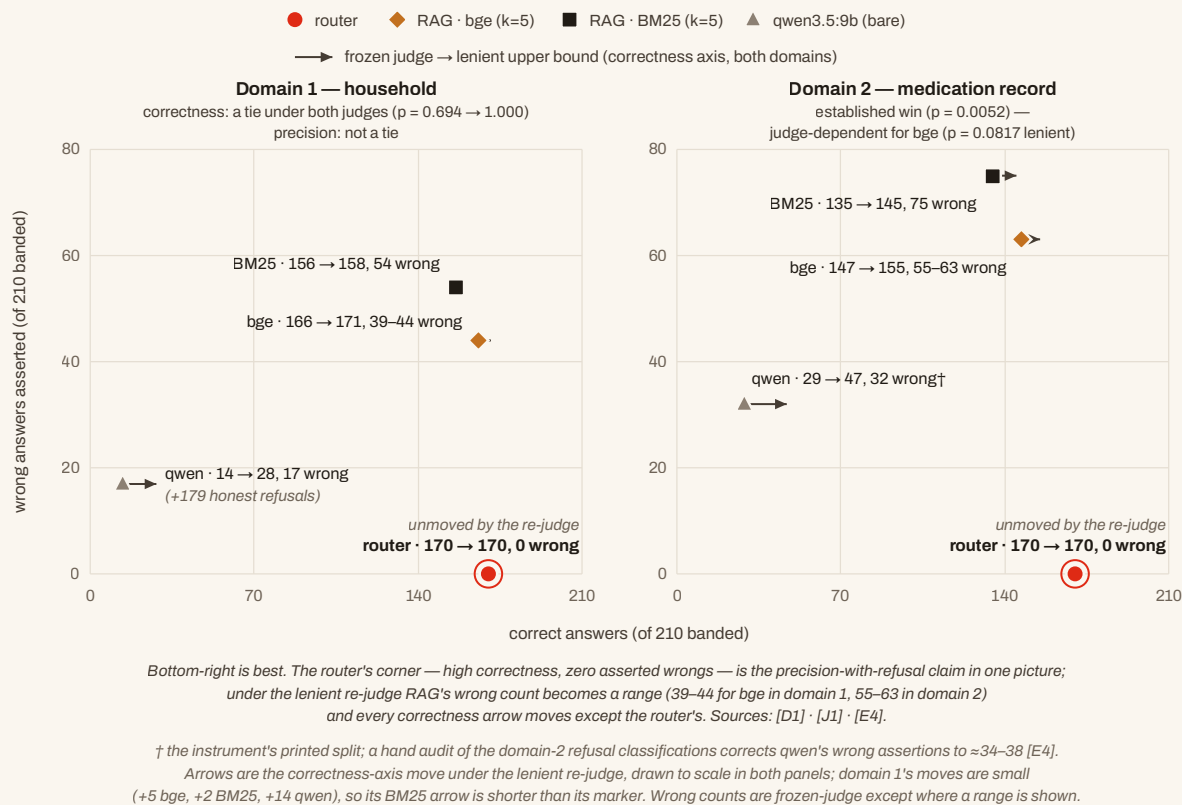


Figure 6 — each arm by correctness (x) and asserted wrong answers (y); bottom-right is best. The router occupies the zero-wrong line in both domains. The arrows on the domain-2 panel are §4.6's judge-sensitivity audit: the lenient re-judge moves every baseline and leaves the router unmoved. Domain 1's audit is tabulated in §4.6 and moves the router by zero there too.

Domain 1 — correctness is a statistical tie; precision is not [D1] — Table 11:

Table 11 — domain 1: the retrieval arm against the router, under the frozen judge.

band	bge (headline)	BM25	router
asserted-wrong	44/210 (21.0% [16–27%])	54/210	0/170 wrong served answers
correct	166/210 [73–84%]	156/210	170/210
pairs separated	40/52 [64–86%]	39/52	52/52

On correctness the pre-named “head-to-head win” fires ($166 < 170$) — and the criterion makes quoting p mandatory precisely here: $p = 0.694$, a statistical tie. The difference lives in the decomposition: on the 170 turns the router served, RAG was **wrong 31 times where the router was wrong none** (139/170 vs 170/170); on the 40 turns the router declined, RAG recovered 27 and asserted 13 wrongs. That is the whole trade, stated without spin and with its two halves kept apart: serving avoids 31 asserted errors on the served turns, and refusing forgoes 27 correct answers on the declined turns while avoiding 13 more there — 27 answers forgone against 44 errors avoided across the exam, at this bar. Of the 12 broken pairs, 9 broke by genuine conflation (a reply naming the paired member's gold).

Domain 2 — the tie does not survive transfer [E4] — Table 12:

Table 12 — domain 2: the same arms after transfer, under the frozen judge.

band	bge (headline)	BM25	router
asserted-wrong	63/210 (30.0% [24–37%])	75/210	0/170 wrong served answers
correct	147/210 (70.0% [63–76%])	135/210	170/210
exact McNemar vs router	p = 0.0052	p = 0.0001	—
pairs separated	25/32 [61–90%] (5 conflations)	18/32	32/32

The cause is structural rather than mysterious: the corpus grew 70% (210 → 357 chunks) while k stayed pinned at 5, so the right chunks compete harder for the same slots. On the 170 served turns RAG is 127/170 (BM25 115/170) against the router’s 170/170.

Where retrieval’s loss concentrates — and where it does not. Single-fact lookup is 30/30 for RAG, a tie with the router; the router’s advantage is concentrated in **aggregation** (`private_inventory`: RAG 15/30 vs 30/30 — a counting question’s members are scattered across the corpus and have to be assembled inside the prompt, and domain 1 had already exhibited the failure mode in a case where all four correct chunks were retrieved and one of them was simply missing from the answer), **event records** (12/30 vs 30/30), and **refusal** (the 63-vs-0 above). The advantage is aggregation, events, and refusal — *not everything* — and saying so is what makes the rest credible.

Near-neighbors, both domains. Across the two exams the in-house evidence is now **84 deliberately confusable pairs: the router separated 84/84** (Wilson lower bound 95.6%), **and retrieval conflated 14** (9 in domain 1, 5 in domain 2 — tartrate for succinate, one levothyroxine strength for the other, overlapping medication classes). This retires the external structural-ceiling citation the project previously leaned on: the claim now stands in-house in both domains (§6).

Unbanded. Retrieval helps enormously over the bare model — +91 turns over the 7B baseline in domain 1, where that model scores 75/210, and +77 in domain 2, where it scores 70/210 — and pays embed + search + a full LLM call on every turn, with zero displacement by construction. The k-sweep (domain 2: bge 114/134/147/155, BM25 98/118/135/148 at k = 1/3/5/10) shows the banded k = 5 is not the peak in either domain.

How far the budget explanation goes. Per-case rows persist only for the pre-committed k = 5, so an exact paired test at other budgets is not computable from the frozen artifacts, and we do not re-run a closed leg to obtain one. The marginals still constrain it: since the router is correct on exactly the turns it served, the exact McNemar p at k = 10 is a function of a single unknown — how many of the router’s forty declines retrieval answers

correctly — and sweeping that unknown across its whole feasible range bounds the test. In domain 2 the router’s lead over BM25 at $k = 10$ remains significant under every feasible pairing ($p \leq 0.037$); the dense retriever’s $k = 10$ margin is indeterminate offline (p between 0.0001 and 0.151) and we make no claim about it [K10]. The budget-under-growth reading is therefore supported at $k = 10$ for the sparse retriever and left open for the dense one — stated rather than resolved, because resolving it would mean re-running an arm after seeing its aggregate.

4.6 The judge-sensitivity audit: the finding that cuts against us

The frozen judge requires exact morphology, and the router’s replies are rendered from the same ids the golds were written in. That is fine for comparing arms under one fixed rule — and it is suspiciously kind to the router. The audit therefore re-judged **every arm of both exams, mechanically and identically**, under a deliberately lenient rule: every content word of the gold present in any morphology and any order — an explicit **upper bound**, not a score, since it ignores negation, attribution, and endorsement.

Table 13 — domain 2 under both judges [E4].

arm (domain 2)	frozen judge (banded)	lenient upper bound	Δ
router (serve path)	170/210	170/210	0
RAG bge (CLS), $k=5$	147/210	155/210	+8
RAG BM25, $k=5$	135/210	145/210	+10
qwen3.5:9b	29/210	47/210	+18

Table 14 — domain 1 under both judges [J1]. The lenient column is the primary reconstruction of the domain-2 rule; across the 32 admissible variants described below, the dense retriever’s bound runs 168–171 and no other arm’s conclusion changes.

arm (domain 1)	frozen judge (banded)	lenient upper bound	Δ
router (serve path)	170/210	170/210	0
RAG bge (CLS), $k=5$	166/210	171/210	+5
RAG BM25, $k=5$	156/210	158/210	+2
qwen3.5:9b	14/210	28/210	+14

The router moves by zero *by construction* in both domains — the unmoved marker in Figure 6: the causal golds are KB state ids (*cardiac event risk*, *poor lung deposition*; *high energy usage*, *pipe freeze*) and the renderer builds replies out of exactly those ids, so the rendered surface is the gold surface. Every other arm paraphrases and pays for it. This is the home-game advantage of a closed-loop judge, quantified rather than hidden.

Domain 1’s correctness comparison is a tie under both judges, and under leniency the point estimate straddles the router. Exact two-sided McNemar moves from $p = 0.694$ (166 vs 170) to $p = 1.000$ (171 vs 170) — thirty-one turns where retrieval is right and the layer is not, against thirty the other way — and across every admissible morphology fold the lenient dense-retriever bound falls between 168 and 171, at p between 0.90 and 1.00. We report this as measured, because it cuts against us. Domain 1 was never a correctness win; the criterion made quoting p mandatory precisely so that it could not be read as one, and leniency deepens the tie rather than overturning it. What survives and what bends:

- **Zero-wrong survives untouched, in both domains.** A wrong answer is wrong under either judge, and the router asserted none. Retrieval’s wrong-assertion count becomes a range in each domain: 39–44 in domain 1 and 55–63 in domain 2, against 0. The asserted-wrong asset is judge-independent.
- **The stronger-model verdict survives with room in both domains:** even at its upper bound the bare 9B sits at 47/210 (domain 2) and 28/210 (domain 1) against 170/210.
- **The domain-2 correctness win over the dense retriever is judge-dependent:** $p = 0.0052$ under the frozen judge (the criterion’s, and the only banded one), $p = 0.0817$ under the lenient bound. BM25 stays established at both ends ($p = 0.0001 / 0.0066$). The honest sentence, which this paper adopts as its statement of the comparison: **the router’s correctness lead over retrieval in domain 2 is robust in direction and, for the dense retriever, sits at the edge of significance once the judge’s morphology strictness is relaxed.**
- **The strictness bites on the causal surface only.** Of the rows the lenient rule moves for the bare 9B, all 14 of 14 in domain 1 and 16 of 18 in domain 2 are in the two chain cells; the lookup, inventory, polarity, and event cells move zero rows in either domain. The surface the private-record claim lives on is not judge-sensitive; the causal surface, where a bare model paraphrases a published mechanism, is.

One asymmetry in how the two audits were produced, disclosed. Domain 2’s lenient figures are the frozen measurement document’s own. Domain 1’s are a later, additive re-scoring of the same persisted replies, produced by a **reconstruction** of the domain-2 rule: the script that produced the domain-2 table was never persisted, so the rule was rebuilt from its stated definition and pinned three ways — it reproduces the frozen judge on all six arm-by-domain cells with zero mismatches, reproduces all four frozen McNemar p -values exactly, and reproduces the published domain-2 lenient table exactly on three of its four arms, with a one-row residual on the dense retriever [J1]. That residual is why domain 1 is reported as a family of 32 admissible rules rather than as a point. The two domains’ lenient numbers are therefore not the output of the same executed code, and no sentence here should be read as though they were; the domain-1 instrument is persisted with the evidence.

The judge itself was not changed: changing it would silently move every frozen artifact, the router’s own numbers included, and void byte-comparability. Both audits sit beside the frozen judge, never inside it. Disclosed sensitivity beats retroactive re-scoring.

4.7 The stronger-model baseline: honesty is not knowledge

Domain 1 [D1]: qwen3.5:9b (9.7B, Q4_K_M, Apache-2.0, thinking off) scores **14/210** [4–11%] on displacement and **10/170** [3–10%] on the turns the router serves exactly (7B comparator: 75/210 and 50/170). The instrument emits the decomposition inline, and it must never be collapsed into a single “wrong” count: **14 correct · 179 honest refusals · 17 wrong assertions**. The 179 are the model correctly declining household facts it cannot know — which is *evidence for* the thesis, not against the model: the facts are not in its weights, and no parameter count puts them there. It is also why the 9B scores below the guessier 7B (14 vs 75 of 210). Only the 17 wrong assertions are comparable to the router’s 0.

Domain 2 [E4]: **29/210** (13.8% [10–19%]) displacement, **28/170** (16.5% [12–23%]) on served turns. The instrument’s printed split — 29 correct · 149 refusals · 32 wrong assertions — is itself wrong in a quantified way, and we publish the correction with the number. A hand audit of all 149 machine-classified refusals flagged **19 replies the label misdescribes, of which 17 are not refusals at all**: two of the nineteen survive the audit as genuine refusals of the record question; two are outright wrong assertions (both answering “yes, widely available as a generic” where the gold is **no** — the model answered a question about the market rather than about the record); and fifteen are **hybrid replies**, a substantive pharmacological answer followed by a record-access disclaimer, which a binary refusal/assertion detector cannot represent. Table 15 is the corrected ledger — 149 – 17 = 132 genuine refusals — and unlike the printed split it closes on 210.

Table 15 — the stronger model’s 210 domain-2 replies, as the instrument printed them and as the hand audit reclassified them. The hybrid column is the category the detector has no representation for.

category	printed	corrected
correct under the frozen judge	29	29
genuine refusals	149	132
wrong assertions	32	34
hybrid: substantive answer + record-access disclaimer	—	15
total	210	210

The hybrid column is where the judgment call sits, which is why the corrected reading is published as a range — **≈132 genuine refusals and ≈34–38 wrong assertions**: the upper end of that range counts four of the fifteen hybrids as assertions, and the remainder are substantively right about the pharmacology while unmatched by a judge looking for the

record's gold. The model's substantive knowledge is therefore materially higher than 29 — 16 of its 18 lenient-bound gains sit in the two pharmacology chain cells (§4.6). The raw counter made the model look more honest and less knowledgeable than it is; the error direction is stated because it matters.

Both pre-named verdicts fired in both domains: *“the correctness advantage is not a 7B artifact”*, and *“even a stronger model errs on turns the router serves exactly.”*

4.8 Grounding bite: established twice, from independent populations

The §3.7 probe ran on each domain's own fall-through population — domain 1: N = 38; domain 2: N = 40; **zero skips in either, so both bite legs are complete** [D1][E4]. Domain 1's population is 38 rather than 40 because two of its forty banded fall-throughs, one per chain cell, fall at resolution and carry no near-miss block for the probe to remove (§4.1). Figure 7 shows the three arms per domain — Table 16:

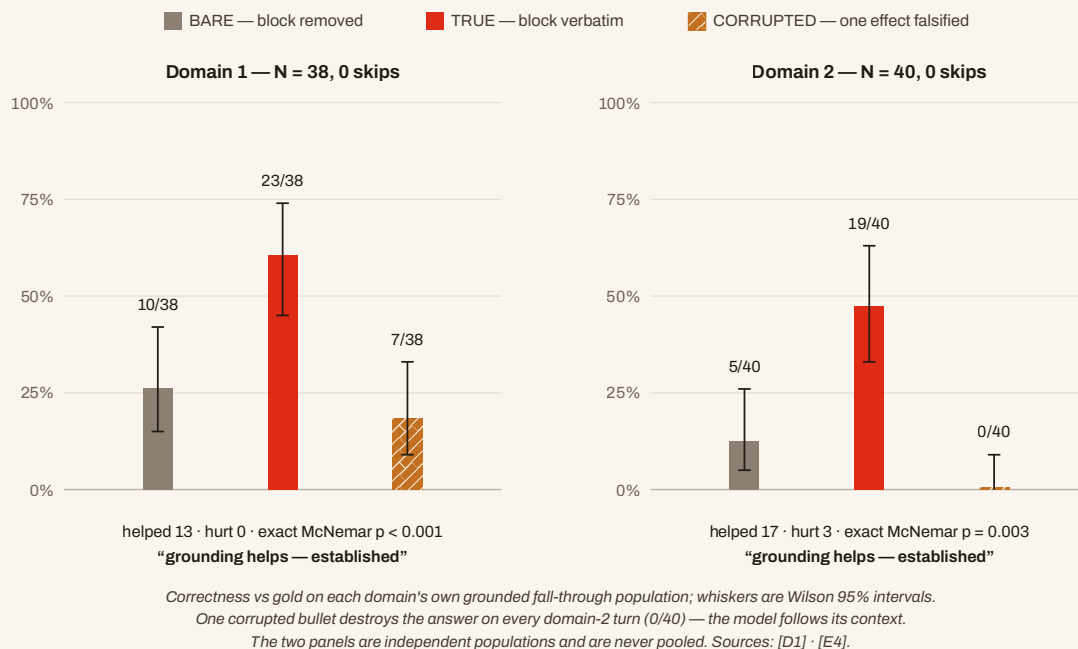


Figure 7 — correctness on each domain's own grounded fall-through population, with Wilson 95% intervals. The verbatim block lifts both domains (established, twice); one corrupted bullet collapses domain 2 to 0/40. The panels are independent populations and are never pooled.

Table 16 — the grounding-bite probe, three arms on each domain's own fall-through population.

arm	domain 1 (N = 38)	domain 2 (N = 40)
BARE (block removed)	10/38 (26.3% [15–42%])	5/40 (12.5% [5–26%])
TRUE (block verbatim)	23/38 (60.5% [45–74%])	19/40 (47.5% [33–63%])
CORRUPTED (one effect falsified)	7/38 (18.4% [9–33%])	0/40 (0.0% [0–9%])
helped / hurt	13 / 0	17 / 3
exact McNemar	p < 0.001	p = 0.003
verdict (pre-committed)	“grounding helps — established”	“grounding helps — established”
follow-rate (CORRUPTED)	15/38 = 39.5% [26–55%]	14/40 = 35.0% [22–50%]

Two domains, two independently derived populations — one with a strong bare baseline, one with a weak one — same established verdict, which is a stronger statement than either alone. Both follow-rates sit between the pre-committed 25% and 50% thresholds, so per the frozen rule both legs are reported separately with no synthesized bite verdict. Domain 2’s CORRUPTED arm is 0/40: falsifying a single bullet destroyed the answer on every turn — the model follows its context, which is precisely why choosing that context deterministically is the design. (Domain 2’s follow-rate was predicted before the run — numerator, denominator, and percentage, exactly.) The two follow-rates are **not like-for-like** and are never compared directly: §5 records the co-residency asymmetry and the probe’s run-to-run noise.

4.9 What a served answer costs

Observational, never banded; magnitudes only, referents always named. Under a pinned performance power profile with strictly sequential metered batches [E4]: symbolic serve arm median **1169 μ s** at 0.0544 J/query (RAPL net); LLM arm median **616 ms** at 112.0 J/query (NVML) — **$\sim 10^3$ on the full serve path** (extraction + two PostgreSQL reads + routing + render). The unpinned Leg-B run gave the same magnitude — both ratios land in the low 10^3 — which is the evidence the magnitude is not a machine-state artifact; the ratio is never quoted as a multiplier with digits. Benchmarking only the bare symbolic inference core against a bare model call yields **$\sim 10^5$** — a different referent, quoted only for what it measures. The stronger model costs 100.7 J/query at 1467 ms median (9.7B, Q4_K_M). Every served turn is a GPU call that never happens; displacement is zero for every baseline by construction.

The referent is the call a serve actually displaces. A served turn does not displace a context-free model call; it displaces a *grounded* fall-through call. The metered LLM arm already is that call: it prompts through the same near-miss grounding surface the real fall-through path uses, so its 250 metered calls are a mixture in which every fall-through turn carries the grounding block, and 112.0 J is the average cost of a call under the real fall-through prompt distribution rather than of a bare-prompt one [EN]. The k-sweep bounds what remains: raising the injected context from one chunk to ten moves median

latency by 0.5% in domain 1 and 20% in domain 2 — generation dominates, prefill of a short grounding block is noise — and a referent change bounded at a fifth of the latency cannot move a magnitude.

The retrieval arm’s per-query joules are withdrawn as a metering artifact. That arm is the only one in the harness whose idle baseline is captured immediately after an LLM call rather than on a quiet machine; the over-stated idle draw is then subtracted across the whole batch window, and the residual — ≈ 5.6 J/query in domain 2, 14.0 J in domain 1, internally inconsistent across k and across domains, and implying a net draw a 7B model cannot generate at — is not a measurement of a `mistral:7b` call [EN]. Its LLM leg costs the same order as the bare arm’s, $\sim 10^1$ – 10^2 J, and its retrieval stage, which the metering window excluded entirely, is additional. The 447 ms median latency is real, is owned by the same frozen document as the rest of the arm, and is retained. Figure 8 puts the arms on one log axis.

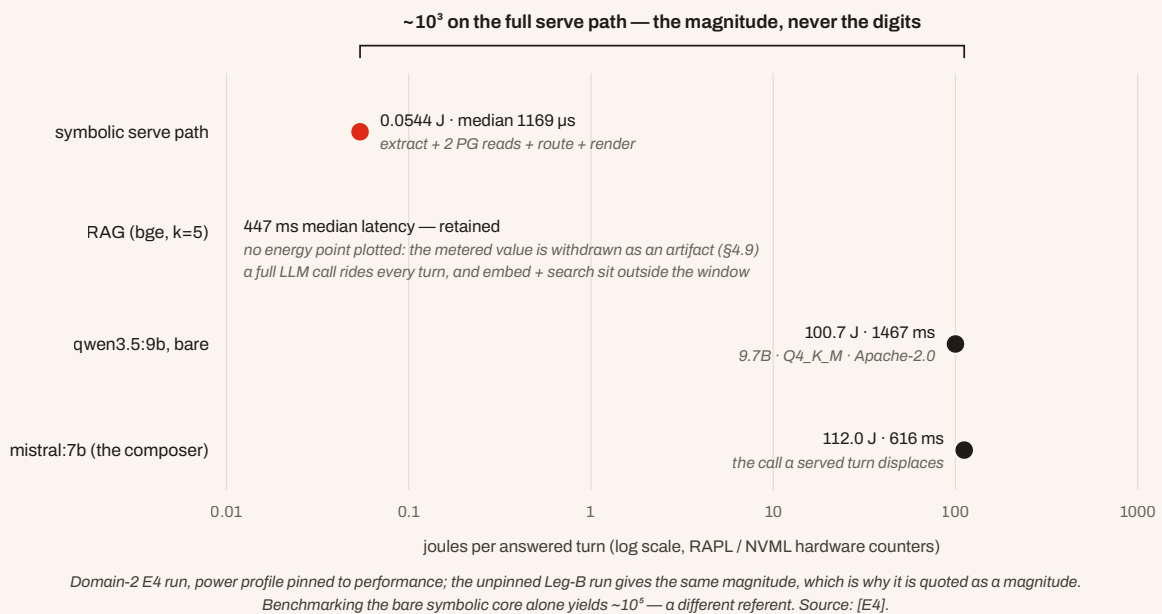


Figure 8 — energy per answered turn on a log axis, hardware-metered under the pinned power profile. The gap is quoted as a magnitude ($\sim 10^3$ on the full serve path), never as a multiplier with digits. The retrieval arm carries no valid energy figure — its metered value is withdrawn (§4.9) — and only its 447 ms median latency is retained.

5 Limitations

Stated at the same altitude as the results, because several are findings.

1. “Tables-as-data” today means data a human writes, not data the system derives. Every one of Leg C’s 159 vocabulary rows was mechanically derivable from the seeded ontology, but they are hand-authored beside the KB, not generated from it. The router did not learn domain 2’s words; a human typed them. Deriving them at seed time

would be a new capability — and it would invalidate the very number Leg C exists to produce, so it was recorded and not built (§7).

2. **The judge is morphology-strict and home-game for the router.** §4.6 quantifies it in both domains: the lenient bound moves every baseline (+8/+10/+18 of 210 in domain 2, +5/+2/+14 in domain 1) and the router by zero; the domain-2 correctness margin over the dense retriever is judge-dependent ($p = 0.0052$ frozen, $p = 0.0817$ lenient), and domain 1’s tie deepens into a straddle ($p = 0.694$ frozen, $p = 1.000$ lenient). The accept-list mechanism cures morphology at the data layer for the lookup-family cells; the causal cells’ golds are KB state ids by design, and that is where the asymmetry concentrates. The judge errs in the opposite direction too, and that error is bounded rather than assumed: a blind hand-audit of thirty judge-correct retrieval replies found one false positive — 3.3% [0.6%, 16.7%], or 13.3% under the strict-hedge reading — an error that can inflate a baseline and can never inflate the router (§3.5).
3. **One judge gap runs against the baselines and could not be closed:** polar YES-golds cannot be given tolerant accept entries, because every candidate surface appears verbatim inside the corresponding negative render — an accept entry would judge a NO reply correct on a YES gold. Disclosed; it biases *against* the arms that paraphrase.
4. **The refusal detector cannot represent hybrid replies.** A hand audit found 19 of the stronger model’s 149 machine-classified domain-2 “refusals” mislabeled — 17 of them not refusals at all, and 15 of those substantive-answer-plus-disclaimer hybrids (§4.7). Reported, not fixed — tuning a detector against the replies it just classified would be fitting. The corrected range ships beside the raw split wherever it is quoted.
5. **assistant_anchored is deliberately empty in domain 2.** Bentley is smart-speaker-specific; no patient-side medication-record traffic anchor survived the bibliography’s standard, and a weak anchor is worse than a named absence. Domain 2’s ecological scenario (med_record) is the mechanical mean of its two sourced scenarios only.
6. **The exams pin resolution to the deterministic by-id path.** The live system adds an embedding-based resolution stage the frozen exams deliberately exclude, so that every serve is attributable to a named vocabulary row, claim, and rule. Every recognition number here is therefore a **conservative lower bound** on what the shipped system resolves.
7. **The two §4.8 follow-rates are not like-for-like, and the probe is not run-reproducible.** Domain 1’s 39.5% was measured against a single-domain state vocabulary (zero cross-domain injections); domain 2’s 35.0% was measured with both domains’ vocabularies co-resident (30% cross-domain injection share). Both stand as measured and frozen; neither may be re-run expecting its frozen value. The corruption arms also carry run-to-run noise at $t = 0$: on today’s co-resident knowledge base two runs of the same binary scored domain 1’s CORRUPTED arm 11/38 and 10/38 correct, against the 7/38 the frozen artifact recorded before co-residency. The main LLM leg, by contrast, reproduced exactly [SH][E4].

8. **A cross-domain-injection *observation*, named as such and not a finding:** within domain 2's CORRUPTED arm, cross-domain (implausible) injections were followed more often than in-domain (plausible) ones — 8/12 vs 6/28, Fisher $p = 0.011$. It is post-hoc, on $n = 12$ vs 28, opposite in direction to the pre-run expectation, and partly a detector artifact (an in-domain follow is more easily masked by the reply also naming the gold). It is a candidate pre-registered hypothesis for a future probe, nothing more [E4].
9. **Scope, verbatim from the frozen scope statement.** Generality beyond the two measured domains is not claimed — two provisioned domains are not open-domain, and the KBs are curated at hundreds-of-facts scale. Routing is not learned or adaptive (the mode inputs are hardcoded literals). Attention/focus is out (unbuilt). Production-traffic coverage is not claimed — fixture-set rates are never production rates. Confidence weights are hand-set by ratified design, not calibrated probabilities; conformal fitting is deferred and the frozen exams are never the calibration set. Temporal expressions are out of both exams (a new capability, not vocabulary growth). No live device state (the event cell serves past-tense records). Local models up to 9.7B; a larger, frontier-class model was named in the arm criterion as an escalation and was not run. Exam traffic, not production traffic.
10. **Author-adjacency, mitigated structurally — and the vocabulary was written with the frozen fixtures in view.** The domain-2 growth was performed by the same author who wrote the exam, and the exam's questions were visible while the 159 rows were written. We state it because it is the fact, and because a reader would otherwise have to assume one way or the other. What it permits is bounded by the protocol rather than by trust: the exam was hash-frozen before a single row was written, so no fixture could be edited to match a row; every cell's count was predicted from the shipped grammar before the measurement ran, so growth could not be steered by a result; the domain-1 gate was measured byte-identical, three times over, rather than argued; and the falsifier fired and was printed. What it does not rule out is the narrower risk that the rows were written toward the phrasings the exam happens to use. Every fixture is authored in-grammar by the exam's author, and the 96-probe paraphrase audit (§3.5) tests the *judge's* tolerance, not the router's parsing of phrasings it has never seen. §7 pre-commits the leg that would settle it.
11. **The retrieval arm's inability to refuse is partly prompt design.** The arm was built to be losable in everything we controlled — same facts, same words, same composer, same judge, better-of-two retrievers — but it was never *instructed* to abstain, and a pipeline told to decline when its retrieved context is insufficient is a different arm from the one measured here. That variant is one of three competitors §6 names and this paper does not measure. They are unmeasured because the submission gate closed before they were named, and adding arms after seeing the results is precisely what the pre-registration exists to prevent — not because we think they would lose.

6 Related work

Retrieval-augmented generation. RAG (Lewis et al. 2020; surveyed in Gao et al. 2024) is the standard architecture for grounding a language model in external knowledge. The present work differs in kind, not degree: retrieval selects context for a model that still composes the answer; the serve seam answers *without* the model, and falls through to (grounded) generation on doubt. The RAG arm in §4.5 is constructed to isolate exactly that difference — same knowledge, same words, same composer, same judge.

KG question answering and neuro-symbolic systems. Answering questions directly over structured knowledge is the KGQA line (surveyed in Lan et al. 2021), and coupling symbolic substrates with neural models is the neuro-symbolic program (Garcez & Lamb 2023), whose LLM-era form is the LLM-plus-KG roadmap of Pan et al. (2024). Our contribution to this line is not the pairing but the *contract* (§2) — in particular fall-through-on-doubt as a stated substrate property with measured consequences — and the pre-registered transfer methodology (Leg A/B/C with a designed falsifier).

Abstention and selective prediction. Refusing to answer under uncertainty is the classic reject option (Chow 1970), revived for deep models as selective prediction (Geifman & El-Yaniv 2017) and for QA under distribution shift (Kamath et al. 2020); Wen et al. (2025) survey abstention in LLMs, where it is increasingly positioned as a hallucination mitigation (Ji et al. 2023). The refusal here is deterministic and priced (a frozen confidence bar over hand-set weights), not a learned reject option; the shadow evidence (§7) shows the refusals are calibration policy over derivations that were correct — correct largely **by construction**, since the fixtures were authored from the same graph the router walks, which is the guard §7 states and this sentence carries with it. The asymmetry that motivates the posture is not ours alone: the Neo4j-funded study of Wedge et al. (2026) independently scores its systems **+1 for a correct answer, 0 for a refusal, and -1 for a hallucination**, on the stated ground that a wrong answer is far worse than a missing one. That is convergent support for the doctrine from a party with no incentive to favor refusal — for the doctrine, not for any measurement in this paper.

Machine understanding, and what this paper does not claim. Chen, Grimm, Russakovsky & Lombrozo (2026) give claims of the form “*does system S understand target T?*” a working taxonomy: understanding constituted by internal models or by measured abilities (optionally with causal-history requirements), asserted at one of four system levels — formal, trained, embedded, distributed — with a load-bearing distinction between what *constitutes* understanding and what is merely *evidence* of it. This paper claims no understanding on any of those accounts. In their terms its results live at the embedded level — a trained model plus deterministic resources — and its evidence is ability-evidence about a mechanism the reader can inspect (§2). Their running boundary case, a lookup table versus a system that derives products it never stored, reads here as a design commitment rather than a criticism: for stored record facts the serve layer is

deliberately lookup-shaped, because on record-shaped domains an auditable lookup that refuses on doubt beats fluent composition (§4.5) — and the bounded rule chaining of C7 is the smallest measured step past lookup, taken under the same zero-wrong discipline.

Structural limits of embedding retrieval. This project previously borrowed an external structural-ceiling result (AssetOpsBench, Patel et al. 2025) for the near-neighbor argument; §4.5 replaces it with in-house evidence in both domains (84/84 vs 14 conflations).

Competitors we did not measure. Three alternatives could contest these results and were not built; the closed submission gate is the reason, and it is a reason rather than an argument. (a) An **abstention-prompted retrieval arm**: our arm’s inability to refuse is partly a property of its prompt, not of retrieval as such (§5), and this is the cheapest of the three to concede later. (b) An **LLM-as-semantic-parser** arm, generating queries against the same knowledge graph, which attacks the vocabulary-table cost of §4.4 directly. (c) **GraphRAG-style structured retrieval**, which sits between our retrieval arm and our layer. For (c) there is now a published referent rather than an abstraction: the Neo4j-funded study of Wedge et al. (2026) builds a lightweight navigational graph over an encyclopedic corpus behind a tool-using agent, and reports its graph arm ahead of a vector-RAG arm on a hallucination-penalizing composite. That is what the arm looks like when it is built properly. It is not a comparison with ours and cannot be made into one: the task is open-domain multi-entity aggregation over tens of thousands of articles rather than curated personal records at two hops, their graph is a navigational index over text where ours is a semantic ground truth whose served answers cite claim ids, and they measure neither latency nor energy. For (b) the same study contributes an independent difficulty report: they attempted text-to-query generation and **abandoned it** — agent reluctance plus injection risk — falling back to hand-written queries hidden behind tools. That does not excuse our not measuring the arm. It records that the alternative has a documented failure mode of its own, and that a second team arrived independently at an authored rather than generated access path into structure.

Rule induction over KGs (the “does this scale past a hand-built KB?” future-work answer): AMIE with its partial-completeness confidence (Galárraga et al. 2013; Lajus et al. 2020), AnyBURL (Meilicke et al. 2019), and LLM-assisted mining in ChatRule (Luo et al. 2023).

Energy measurement. The metering protocol (hardware counters, batch metering, warm-up discipline, idle-baseline guards) follows the position that TDP- or latency-derived energy figures overstate or understate real inference energy, argued and instrumented by the ML.ENERGY benchmark (Chung et al. 2025).

Workload anchors (verified in the project bibliography, each with a does-not-license note): Graesser & Person (1994) for the question taxonomy; Stivers (2010) and Enfield, Stivers & Levinson (2010, the ten-language comparison) for conversational question form; Bentley et al. (2018) for smart-speaker traffic; Arnold & Kaske (2005) and Radford &

Connaway (2013) for reference-desk shares. Design-theory framing: Gregor & Jones (2007).

7 Discussion and future work

What the refusals are made of. After the submission evidence froze, a pre-registered observational probe re-invoked the router’s own decision core on every banded refusal with the serve bar moved to zero — reading the router’s own derivation, not a re-implementation — and corroborated each result against the near-miss block already materialized on the wire (78/78 agree) [SH]. Across both domains, **all 38 two-hop conclusions the router derived, priced, and declined to assert were correct** (18/18 and 20/20; the one-hop refusals were 40/40 beside them). The denominator is 38, not the 40 two-hop questions the two exams contain between them: two domain-1 fixtures fall at resolution, where the entity-peel ladder never yields a declared state, so no closure is ever attempted. They are not failed derivations — and they are the same two turns whose absence makes §4.8’s domain-1 population 38 [SH]. Read honestly, this is correct largely **by construction** — the fixtures were authored from the same graph the router walks — so it establishes only that the refusal is *calibration policy over a correct derivation*, removing one alternative explanation (“the engine cannot get there”) and nothing more. It moved no band, and it sharpens exactly one future question: whether a *derived* conclusion deserves its own, separately calibrated serve bar. That question is deliberately post-paper: the 0.80 bar predates both exams, and moving it together with the exams would make any movement unattributable. (One mechanical footnote the probe surfaced: the transitive render form — “*can ultimately lead to*” — has never been exercised on the serve path in any frozen artifact; every measured serve is depth-1, and all 20 transitive-form fixtures per domain are among the designed refusals.)

The loop this is step one of. In the wider system, the seam is the first stage of a designed loop (Figure 9): **serve** what the graph proves (this paper); **ground** what it cannot (measured here: §4.8); **record** every miss as an explicit knowledge gap that drives the assistant to ask; **gate** every write; **distill**, eventually, with the LLM as one more registered, gated source of structured claims.

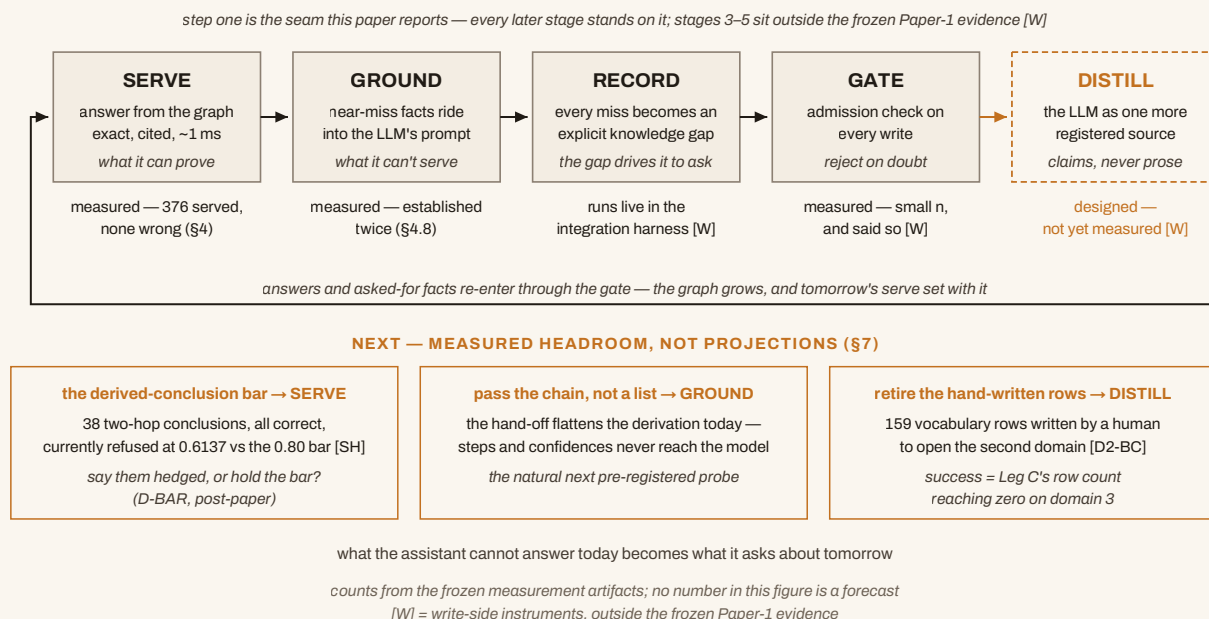


Figure 9 — the loop, with each stage's measurement status and the measured headroom the next steps aim at. Stages 3–5 sit outside this paper's frozen evidence and are labeled [W]; no number in the figure is a forecast.

The write side sits outside this paper's frozen evidence and is quoted here with its own instruments' numbers and scope caveat [W]: the admission gate measures recall 100% [72–100%] (n = 10) with 0 false positives (n = 5), the read-side consistency gate 100/100/0 on its core set (small n — the Wilson lower bounds sit near 50% and are quoted deliberately), and the gate's coverage is symbolic-fact writes only — ontology and memory-internal writes bypass it today. A live integration harness demonstrates the gap-driven posture qualitatively: holding nothing about a nonsense entity, the gated assistant asks about it; the ungated reply beside it hallucinates. None of this loop machinery was allowed to touch the frozen exams.

Future work, in order of how directly it follows from the measurements. (1) The derived-conclusion bar, now backed by 38 correct withheld conclusions. (2) A richer fall-through hand-off: the grounding block today is a flattened fact list; passing the derivation chain with its confidences is the natural next probe. (3) **Widening the exercised inference set** — the implication of §1's scope statement. The substrate already declares relational axioms (inverse, symmetric, transitive) that no live serve path consumes today; one-hop inverse and symmetric derivations clear the existing, unmoved serve bar for high-confidence edges, and a bounded subtype closure would let enumeration questions name supertypes honestly. Each widening is a *reach* change, not a *licensing* change — the bar, the refusal discipline, and the audit trail are held fixed — and each carries its own pre-committed measurement. This is the designed trajectory from the one inference family measured here toward the engine's full registered set of 49 primitives, with the precision asset carried forward at every step rather than re-earned at the end. (4) The distillation leg that begins to retire limitation #1 — vocabulary rows derived from the seeded ontology through the same admission gate as every other write, measured by whether Leg C's row

count goes to zero on a third domain. (5) A pre-registered version of the cross-domain-injection observation (limitation #8). (6) **A held-out paraphrase leg, pre-committed here.** Every fixture is authored in-grammar by the exam’s author, and the vocabulary was grown with those fixtures in view (§5), so robustness to unseen phrasing is untested. The leg is a rewrite of each frozen question by a party that did not author the grammar — another person, or a model given the record but not the tables — scored against the same golds under the same frozen judge, with both outcomes named before it runs. It is deliberately not run here: it is a new measurement, and the submission gate is closed. (7) Learned/adaptive routing and attention are later papers by explicit sequencing, and nothing here implies them.

8 Conclusion

Placed in front of a local LLM, a deterministic serve layer satisfying the seven-point contract of §2 answered bounded-domain record questions with a property none of the measured alternatives has: **it was never wrong when it answered, in any measured leg, in either domain.** That is 376 served answers — a count over two domains that are never pooled — zero over-routes on every control leg, and 84/84 near-neighbor pairs separated. It paid for the discipline in refusals whose cost the same exams price: on domain 1, 27 answers forgone on the turns it declined, against 44 errors retrieval asserted and it did not — 31 of them on the turns it served. And the boundary of its competence proved to sit exactly where its vocabulary tables end — near-total refusal zero-shot, cell-for-cell parity after 159 human-written rows, and a pre-committed falsifier that fired on the two code changes adaptation actually required, a sentence we print rather than soften. Retrieval over the same facts in the same words matched it on single-fact lookup, tied it on domain-1 correctness ($p = 0.694$, and $p = 1.000$ under a lenient judge bound whose point estimate straddles the layer), and lost the tie under transfer ($p = 0.0052$, judge-dependent for the dense retriever at that bound) — while asserting between 39 and 44 wrong answers in domain 1 and between 55 and 63 in domain 2, each a judge range (§4.6), and conflating 14 of the 84 pairs. A stronger bare model mostly — honestly — declined. Supplying the declined facts as grounding helped: established twice, from independent populations. A served answer costs about a millisecond and about a thousandth of the displaced call’s energy.

The evaluation protocol is, we think, as much of a contribution as the layer: bands frozen before fixtures, both outcomes pre-named, hash-pinned workloads, stage-level predictions from shipped code before every run, a falsifier designed into the claim most likely to be oversold — and fired. The claims stop where the measurements stop, and §5 is the map of where that is.

Reproducibility statement

All criteria, workloads, evidence artifacts, and analysis documents are version-controlled in the project’s research repository; the measurement harness is an eval crate in the implementation repository. Frozen data is git-hash-pinned and verified before and after every session — Table 17.

Table 17 — the pins.

frozen artifact	kind	hash
domain-1 v1 workload	git tree	651585ed9744
domain-1 KG (concepts / rules / facts)	git blobs	b650f00a8450 / 368c4b5b79ff / e8117d07d9aa
domain-1 v2 workload (250 fixtures + KG extension)	git tree	a08bc144838e
domain-2 workload	git tree	5ec66945c272
domain-2 KG	git subtree	63ab8f4834cc
Leg-A vocabulary state (per grammar file)	git blobs	96e62d51bcb7 · c799ea37ef40 · 4d489f11fc63 · aa0d806280e7
RAG corpus, domain 1 (210 chunks)	SHA-256	aa0c9d8d3091...
RAG corpus, domain 2 (357 chunks)	SHA-256	f5a294daa20c...

Judges, bars, and weights are byte-frozen; additive instruments carry byte-identity regression contracts against the frozen comparators; the domain-2 headline run was independently reproduced with 0 of 250 per-case rows differing. Energy is metered by hardware counters (Intel RAPL, NVML), batch-metered with mandatory warm-up, under a recorded pinned power profile for the final runs. Hardware: Ryzen-class 16C/32T CPU, RTX 4070 Laptop GPU (8 GB), 88 GB RAM; models via Ollama (mistral:7b-instruct; qwen3.5:9b Q4_K_M, thinking off), embedder bge-base-en-v1.5 via ONNX.

Throughout the paper, bracketed tags cite the frozen document that owns each number; every number is quoted from its owning document without adjustment. The tags resolve as follows (criteria [C-∗] precede their measurements by construction, per Table 2) — Table 18:

Table 18 — the citation tags and the documents that own them.

tag	frozen artifact
[C-D1]	the domain-1 v2 criterion (+ traffic profiles, workload freeze)
[C-D2]	the domain-2 criterion (+ workload freeze, Leg-B and E4 predictions)
[C-RAG] · [C-SM] · [C-F]	the RAG-arm, stronger-model-arm, and grounding-probe criteria
[D1]	the domain-1 v2 measurement report
[D2-A] · [D2-BC]	the domain-2 Leg-A and Legs-B/C measurement reports
[E4]	the domain-2 arms + judge-sensitivity report, the last measurement the evidence freeze admitted
[SH]	the two-hop shadow report — observational: pre-registered, reported beside the bands, entering none
[J1] · [EN] · [K10] · [FP]	the four additive offline analyses run after the freeze — the domain-1 lenient re-judge, the energy-metering diagnosis, the k = 10 feasibility bounds, and the judge false-positive hand audit. All re-score or re-read persisted artifacts, make no model call, move no banded number, and enter no band
[W]	the write-side instruments, outside the frozen Paper-1 evidence

The contract of §2 is the reproduction path: the claims can be checked on a substrate a reader builds — any system satisfying C1–C7 — by testing whether the predictions hold.

The measurement artifacts live in the project’s research repository.

Ethics statement

All data in both domains is synthetic and was authored for this evaluation: no real person, household, prescription, or clinician exists anywhere in the evidence. Domain 1 is an invented domotic household (`maison_jasper`); domain 2 is an invented medication and care record for one invented individual, and its exam asks about that record and about published mechanisms — never what a person should do; advice and deontic forms are out-of-grammar and refuse by construction. No personal data was collected, processed, or archived at any point. Everything runs locally; the per-assistant knowledge graph is isolated structurally. The refusal-first posture evaluated here is itself a safety stance: on record-shaped domains, a wrong served answer is worse than a refusal, and the measured design treats it that way.

Appendix A — the traffic-scenario weight matrices

Tables 19 and 20 give the complete cell weights (percent) as frozen in each criterion’s `traffic_profiles.yaml`; the evaluation reader recomputes every ceiling and floor from these weights and refuses a file that disagrees. Derivations: `conversation_anchored` from Stivers 2010 (polar 230 / content 90 / alternative 8; alternative dropped and renormalized); `assistant_anchored` from Bentley et al. 2018 under the engagement-proportional reading (information 17 : automation 9); `reference_anchored` from the reference-desk studies with the ready-reference share pinned at the harsh-to-thesis 33%; `uniform` is the null; each domain’s final scenario is the **mechanical** element-wise mean of its sourced scenarios — no editorial profile exists. Mapping law: uniform-within-mapped-group; `out_of_domain` weighted only by topic-axis sources; form-authoring pins.

Table 19 — domain 1, the five scenarios [C-D1].

cell	conversation	assistant	reference	uniform	jasper_house
<code>private_polarity</code>	35.94	5.77	6.60	11.11	16.10
<code>event_history</code>	35.94	5.77	6.60	11.11	16.10
<code>private_lookup</code>	4.69	5.77	6.60	11.11	5.69
<code>private_relational</code>	4.69	5.77	6.60	11.11	5.69
<code>private_inventory</code>	4.69	5.77	6.60	11.11	5.69
<code>grounded_chain</code>	4.69	5.77	22.33	11.11	10.93
<code>generic_chain</code>	4.69	21.79	22.33	11.11	16.27
<code>llm_required</code>	4.69	21.79	22.33	11.11	16.27
<code>out_of_domain</code>	0	21.79	0	11.11	7.27
weighted ceiling	95.3%	56.4%	77.7%	77.8%	76.5%
weighted guess floor	35.9%	5.8%	6.6%	11.1%	16.1%

Table 20 — domain 2, the four scenarios [C-D2]. The three shared columns are numerically identical to domain 1’s — a pinned instrument property, and the license for cross-domain comparison; `assistant_anchored` is empty with the reason stated in §3.4.

cell	conversation	reference	uniform	med_record
private_polarity	35.94	6.60	11.11	21.27
event_history	35.94	6.60	11.11	21.27
private_lookup	4.69	6.60	11.11	5.64
private_relational	4.69	6.60	11.11	5.64
private_inventory	4.69	6.60	11.11	5.64
grounded_chain	4.69	22.33	11.11	13.51
generic_chain	4.69	22.33	11.11	13.51
llm_required	4.69	22.33	11.11	13.51
out_of_domain	0	0	11.11	0
weighted ceiling	95.3%	77.7%	77.8%	86.5%
weighted guess floor	35.9%	6.6%	11.1%	21.3%

(Weights shown to two decimals; the frozen YAML files carry exact fractions, and the reader recomputes from those.)

Appendix B — the domain-2 rule base and its mechanism sources

The 16 curated causal rules of §3.3, each with the published source its mechanism is curated from. The record the rules are applied to is synthetic (Ethics statement); the mechanisms are general, published pharmacology — the only part of the domain-2 KG that refers to the real world. The sources document the mechanisms; they do not license the rules’ confidence values, which are hand-set design parameters mirroring domain 1’s multiset (§3.3), nor any clinical advice — the exam asks what a mechanism causes, never what anyone should do.

Table 21 — the 16 domain-2 causal rules and the published mechanism each is curated from. Causes and effects are quoted as the knowledge base labels them, which is why one entry keeps a spelling the rest of the paper does not use.

#	rule (cause → effect)	mechanism	source
1	missed dose → subtherapeutic drug level	trough falls below the therapeutic window on an omitted dose	Osterberg & Blaschke 2005
2	subtherapeutic drug level → symptom recurrence	efficacy is lost below the minimum effective concentration	Osterberg & Blaschke 2005; Muller & Milton 2012
3	grapefruit intake → raised plasma concentration	furanocoumarins irreversibly inhibit intestinal CYP3A4	Bailey, Dresser & Arnold 2013
4	raised plasma concentration → increased side-effect risk	dose-related adverse effects scale with exposure	Muller & Milton 2012
5	NSAID use → gastric mucosa irritation	COX-1 inhibition suppresses protective mucosal prostaglandins	Wallace 2008
6	gastric mucosa irritation → stomach ulcer risk	impaired mucosal defense is the pathway to NSAID peptic ulceration	Lanas & Chan 2017
7	antibiotic course → gut flora disruption	broad-spectrum activity depletes commensal flora	Bartlett 2002
8	gut flora disruption → antibiotic diarrhoea	loss of colonic flora and its fermentation function; C. difficile as the severe form	Bartlett 2002
9	abrupt beta-blocker stop → rebound tachycardia	up-regulated adrenoceptors exposed on withdrawal	Miller et al. 1975
10	rebound tachycardia → cardiac event risk	raised rate-pressure product; withdrawal precipitates coronary events	Miller et al. 1975
11	expired medication → reduced potency	assay guaranteed only within the stability-tested shelf life	ICH Q1A(R2) 2003
12	heat exposure → insulin degradation	thermal stress degrades the protein and its activity	Kaufmann et al. 2021
13	double dose → overdose risk	dose-exposure proportionality crosses the toxic threshold for narrow-margin drugs	Muller & Milton 2012
14	warfarin with NSAID → bleeding risk	additive hemostatic impairment plus NSAID mucosal injury	Battistella et al. 2005

#	rule (cause → effect)	mechanism	source
15	alcohol with sedative → excessive drowsiness	additive CNS depression	Weathermon & Crabb 1999
16	inhaler without spacer → poor lung deposition	oropharyngeal impaction without a spacer reduces the respirable fraction	Newman 2004

References

- Arnold, J., & Kaske, N. (2005). Evaluating the quality of a chat service. *portal: Libraries and the Academy* 5(2), 177–193.
- Bailey, D. G., Dresser, G., & Arnold, J. M. O. (2013). Grapefruit-medication interactions: forbidden fruit or avoidable consequences? *CMAJ* 185(4), 309–316.
- Bartlett, J. G. (2002). Antibiotic-associated diarrhea. *New England Journal of Medicine* 346(5), 334–339.
- Battistella, M., Mamdani, M. M., Juurlink, D. N., Rabeneck, L., & Laupacis, A. (2005). Risk of upper gastrointestinal hemorrhage in warfarin users treated with nonselective NSAIDs or COX-2 inhibitors. *Archives of Internal Medicine* 165(2), 189–192.
- Bentley, F., Luvogt, C., Silverman, M., Wirasinghe, R., White, B., & Lottridge, D. (2018). Understanding the long-term use of smart speaker assistants. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2(3), Article 91. doi:10.1145/3264901.
- Chen, H., Grimm, S. R., Russakovsky, O., & Lombrozo, T. (2026). Machine understanding. *Trends in Cognitive Sciences*. doi:10.1016/j.tics.2026.04.003.
- Chow, C. K. (1970). On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory* 16(1), 41–46.
- Chung, J.-W., Ma, J. J., Wu, R., Liu, J., Kweon, O. J., Xia, Y., et al. (2025). The ML.ENERGY benchmark: toward automated inference energy measurement and optimization. *NeurIPS 2025*. arXiv:2505.06371.
- Enfield, N. J., Stivers, T., & Levinson, S. C. (2010). Question–response sequences in conversation across ten languages: an introduction. *Journal of Pragmatics* 42(10), 2615–2619.
- Galárraga, L., Teflioudi, C., Hose, K., & Suchanek, F. (2013). AMIE: association rule mining under incomplete evidence in ontological knowledge bases. *WWW 2013*. doi:10.1145/2488388.2488425.
- Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., et al. (2024). Retrieval-augmented generation for large language models: a survey. arXiv:2312.10997.
- Garcez, A. d’Avila, & Lamb, L. C. (2023). Neurosymbolic AI: the 3rd wave. *Artificial Intelligence Review* 56(11), 12387–12406.

- Geifman, Y., & El-Yaniv, R. (2017). Selective classification for deep neural networks. *NeurIPS 2017*, 4878–4887.
- Graesser, A. C., & Person, N. K. (1994). Question asking during tutoring. *American Educational Research Journal* 31(1), 104–137.
- Gregor, S., & Jones, D. (2007). The anatomy of a design theory. *Journal of the Association for Information Systems* 8(5), 312–335.
- ICH (2003). *Q1A(R2): Stability testing of new drug substances and products*. International Council for Harmonisation, Step-4 guideline.
- Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., et al. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys* 55(12), Article 248.
- Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., et al. (2023). Mistral 7B. arXiv:2310.06825.
- Kamath, A., Jia, R., & Liang, P. (2020). Selective question answering under domain shift. *ACL 2020*, 5684–5696.
- Kaufmann, B., Boule, P., Berthou, F., Fournier, M., Beran, D., Ciglenecki, I., et al. (2021). Heat-stability study of various insulin types in tropical temperature conditions. *PLOS ONE* 16(2), e0245372.
- Lajus, J., Galárraga, L., & Suchanek, F. (2020). Fast and exact rule mining with AMIE 3. *ESWC 2020*. doi:10.1007/978-3-030-49461-2_3.
- Lan, Y., He, G., Jiang, J., Jiang, J., Zhao, W. X., & Wen, J.-R. (2021). A survey on complex knowledge base question answering: methods, challenges and solutions. *IJCAI 2021*. doi:10.24963/ijcai.2021/611.
- Lanas, A., & Chan, F. K. L. (2017). Peptic ulcer disease. *The Lancet* 390(10094), 613–624.
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *NeurIPS 2020*. arXiv:2005.11401.
- Luo, L., Ju, J., Xiong, B., Li, Y.-F., Haffari, G., & Pan, S. (2023). ChatRule: mining logical rules with large language models for knowledge graph reasoning. arXiv:2309.01538.
- Meilicke, C., Chekol, M. W., Ruffinelli, D., & Stuckenschmidt, H. (2019). Anytime bottom-up rule learning for knowledge graph completion. *IJCAI 2019*, 3137–3143. doi:10.24963/ijcai.2019/435.
- Miller, R. R., Olson, H. G., Amsterdam, E. A., & Mason, D. T. (1975). Propranolol-withdrawal rebound phenomenon: exacerbation of coronary events after abrupt cessation of antianginal therapy. *New England Journal of Medicine* 293(9), 416–418.
- Muller, P. Y., & Milton, M. N. (2012). The determination and interpretation of the therapeutic index in drug development. *Nature Reviews Drug Discovery* 11(10), 751–761.
- Newman, S. P. (2004). Spacer devices for metered dose inhalers. *Clinical Pharmacokinetics* 43(6), 349–360.

- Osterberg, L., & Blaschke, T. (2005). Adherence to medication. *New England Journal of Medicine* 353(5), 487–497.
- Pan, S., Luo, L., Wang, Y., Chen, C., Wang, J., & Wu, X. (2024). Unifying large language models and knowledge graphs: a roadmap. *IEEE Transactions on Knowledge and Data Engineering* 36(7), 3580–3599.
- Patel, D., Lin, S., Rayfield, J., Zhou, N., Shyalika, C., Yarrabothula, S. R., et al. (2025). AssetOpsBench: benchmarking AI agents for task automation in industrial asset operations and maintenance. arXiv:2506.03828.
- Qwen Team, Alibaba Group (2026). *Qwen3.5 open-weight model series* (Qwen3.5-9B, Apache-2.0). <https://huggingface.co/Qwen>.
- Radford, M. L., & Connaway, L. S. (2013). Not dead yet! A longitudinal study of query type and ready reference accuracy in live chat and IM reference. *Library & Information Science Research* 35(1), 2–13.
- Stivers, T. (2010). An overview of the question-response system in American English conversation. *Journal of Pragmatics* 42(10), 2772–2781.
- Wallace, J. L. (2008). Prostaglandins, NSAIDs, and gastric mucosal protection: why doesn't the stomach digest itself? *Physiological Reviews* 88(4), 1547–1565.
- Weathermon, R., & Crabb, D. W. (1999). Alcohol and medication interactions. *Alcohol Research & Health* 23(1), 40–54.
- Wedge, C. J., Stutter, J., Dixon, D., & Cała, J. (2026). Reducing hallucinations in complex question answering using simple graph-based retrieval-augmented generation (long version). Technical report, National Innovation Centre for Data, Newcastle University. Funded by Neo4j; CC BY-NC-SA 4.0.
- Wen, B., Yao, J., Feng, S., Xu, C., Tsvetkov, Y., Howe, B., et al. (2025). Know your limits: a survey of abstention in large language models. *Transactions of the Association for Computational Linguistics* 13, 529–556.
- Xiao, S., Liu, Z., Zhang, P., Muennighoff, N., Lian, D., & Nie, J.-Y. (2024). C-Pack: packed resources for general Chinese embeddings. *SIGIR 2024*. doi:10.1145/3626772.3657878.