The Record, Not the Model:
Cross-Architecture Evidence that Near-Domain Retrieval Confabulation Is Dominated by the Served Record

Tad MacPherson
Metavolve Labs, Inc.
research@iaeternum.ai

July 2026. Cross-architecture matrix preprint, v0.7

Abstract

Everyone benchmarks the engine; almost no one crash-tests the car. Capability leaderboards ask what a model can produce; safety in a retrieval-augmented deployment turns on the opposite question — whether the model *refrains* when the retrieved record does not contain the answer. We report a controlled behavioral matrix that scores abstention as a first-class success across eight contemporary architectures (Llama-3.1-8B, Llama-3.3-70B, Qwen3.7-Max, a custom laboratory model "Inkling," gpt-oss-120b, Kimi-K2.6, GLM-5.2, and DeepSeek-V4-Pro), each run over a frozen, hash-sealed set of 168 private-fact probes under matched context arms and two instruction regimes. Surface behavior diverges enormously: context-free failure-to-abstain ranges from 3.8% to 95.8%, and within the Llama family scaling 8B to 70B *raises* it from 12.1% to 92.1% (a serving-path-confounded observation). An explicit abstain-permission instruction cures the context-free and irrelevant-context arms to near zero — but on authentic near-domain *art-provenance* records, instructed failure persists at 14.6–20.8% in every architecture. An earlier draft of this work read that convergence as an "instruction-resistant residual near 18%" shared across architectures. Our own subsequent refit — reported here, adversarially re-derived cell-by-cell from the sealed artifacts — killed that reading: decomposed by the corpus's dossier-triad structure, the "18%" is a composition average over grossly heterogeneous stimulus difficulty (1.3%, 35.7%, and 16.7% by triad; $\chi^2(2)=154.7$, $p=2.5\times10^{-34}$), the cross-model convergence sits at the *center* of a shared-stimulus null (observed cross-model SD 2.11 points; null median 2.11; $P\approx0.5$), and architecture explains approximately *zero* variance by four methodologically distinct estimators. The defensible — and, we argue, more consequential — claim is **stimulus dominance**: on identical near-domain records, failure rates vary by **34.4 percentage points** across record triads (35.7% vs 1.3%) and negligibly across the eight architectures (seven organizations) we tested, in the same triad order; between-model heterogeneity is unresolved at this design's power rather than demonstrated absent. A content-rotation control on a public, value-rotated release of the corpus (all sensitive values changed, sentence structure preserved) reproduces the triad signature in all six re-run models under two machine grader families and a stratified 137-row human anchor, showing the effect follows the *structure* of near-domain records rather than the rotated content. We further document a five-way *dialect taxonomy* of confabulation strategies and an effort economy spanning 9.5 to 371 tokens per answer. Two operational conclusions follow: default configuration is not safe configuration, and because neither prompting nor model choice removed the record-dominated failure here, these data most directly motivate remedies on the *input side* of the read path — admissibility gating and serving structure — alongside documented post-retrieval attribution interventions. The internal benchmark is externally re-analyzable but not re-capturable (source records withheld); the public rotated corpus is externally re-runnable end-to-end. All probes, grades,

verbatim adjudication rationales, the refit and replication scripts, the public replication corpus, and a hash-chained provenance manifest are released. Companion to the single-model pilot [1].

1 Introduction

The public evaluation of language models is overwhelmingly an evaluation of the *engine*: how much a model knows, how well it reasons, how fluently it writes. This is the right instrument for one job — ranking capability — and the wrong instrument for another — deciding whether a system is safe to wire into a memory that is, like every real memory, incomplete. A retrieval-augmented system spends most of its operating life being handed records that are authentic, relevant, and *insufficient* for the exact claim requested. The safety-critical question is not what the model produces when it has the answer; it is what the model does when it does not. Abstention, on this framing, is not a refusal to be helpful. It is the correct answer, and an evaluation that scores it as a miss is measuring the wrong thing.

We therefore treat abstention as a first-class success and run a deliberate crash test. The companion pilot [1] isolated, in a single frozen open-weight model, the phenomenon this paper generalizes: when the retrieved record is authentic, high-quality, and topically adjacent to the question but does *not* contain the answer, the model does not fall silent — it quarries true specifics from whatever context is present and reassigns their provenance to the queried scope. The pilot demonstrated the effect exists and is misattribution-dominated. It could not say whether the effect is a property of one model or a property of the retrieval setting itself.

This paper answers that question with a cross-architecture behavioral matrix — and the answer sharpened under our own adversarial re-analysis in a way we did not anticipate. Eight models spanning seven organizations, dense and sparse mixtures, reasoning and non-reasoning designs are run over an identical hash-sealed private-fact benchmark. All eight converge on similar overall near-domain failure under instruction — a convergence an earlier draft read as a shared "floor." The corrected finding, obtained by refitting our own data with its stimulus structure made explicit, is stronger and different: **the stimulus dominates the observed variation**. Failure rates differ by **34.4 percentage points** *across record triads* (pooled 35.7% vs 1.3%) while varying negligibly *across architectures*; every model ranks the triads identically; and the apparent convergence is exactly what identical stimuli plus sampling noise produce. Our contributions:

1. **The behavioral matrix (R1).** A full eight-model, two-regime table of failure-to-abstain with Wilson intervals. Surface behavior diverges enormously: context-free confabulation spans 3.8% to 95.8%; one model answers even purely random context at 86.7%.

2. **Stimulus dominance (R2).** Under an abstain-permission instruction, near-domain art-provenance failure is dominated by record identity rather than by resolved between-model differences in this benchmark: per-triad pooled rates of 1.3%, 35.7%, and 16.7% ($\chi^2(2)$=154.7, p=2.5×10^{-34}), the same ordering in all eight architectures, architecture explaining ∼0% of variance by four independent estimators, and the celebrated cross-model convergence sitting at the center of a shared-stimulus null. We report this as the correction of our own earlier framing, with the audit trail.

3. **The configuration finding (R3).** The same weights are simultaneously catastrophic uninstructed and clean instructed. *Default configuration is not safe configuration.*

4. **Inverse scaling within a family (R4).** Holding vendor and training recipe fixed, Llama 8B to 70B *increases* context-free failure from 12.1% to 92.1% — a within-family observation with an explicit serving-path confound, not a scaling law.

5. **A dialect taxonomy (R5).** Five named confabulation strategies, each spoken by an identifiable subset of models.

6. **A content-rotation control (R6).** A public, value-rotated release of the art corpus — every sensitive value changed, every sentence structure preserved, equivalence enforced by a pre-registered QA gate — reproduces the triad signature in all six re-run models under two machine grader families (cell-level mean absolute difference 2.6 points). The effect follows the record's *structure* rather than the rotated content, and the public corpus ships as a self-contained replication package enabling a true external re-run.

The through-line for deployment is blunt, and blunter than in our earlier draft. Because the surface is so configurable, a leaderboard number certifies nothing about a deployed read path (R3). Because near-domain failure is record-dominated, with no resolved architecture effect (R2, R6), *no tested model choice removed it*: swapping vendors moved the number by approximately nothing, while the structure of the served record moves it by 34.4 percentage points. Neither model selection nor prompting closed it here; what is served, and how it is structured, is the lever these data most directly motivate — the subject of the Discussion.

2 Related Work

The retrieval literature has documented many ways context can hurt: irrelevant distractors degrade reasoning [11], unhelpful retrieved passages reduce rather than improve accuracy [12], semantically similar distractors mislead more than random ones [13], and models entrain on related context tokens [14]. Faithfulness and abstention under provided knowledge have direct evaluation lineages: instruction-following QA models are competitive on correctness but weak on faithfulness [6]; RAGTruth supplies a human-annotated, word-level RAG hallucination corpus [7]; MTRAG benchmarks unanswerable questions inside realistic multi-turn RAG [8]; and MIRAGE attributes answers to context via model internals [9]. A newer thread isolates wrong-source attribution as a distinct failure class — Ghost Context shows misattributed grounding is often causally localizable to single context spans and largely correctable by masking interventions [18] — including in clinical retrieval where the stolen claim is genuinely supported by the record but about the wrong entity [21]. Detection work on model-internal honesty signals spans hidden-state probes [15], refusal directions [16], and semantic-entropy uncertainty [22, 17]; calibration work asks whether models know what they know [10]. Empirical evaluations of agentic citation fidelity find frontier deep-research agents cite plausibly but verify poorly [19], and source-verification frameworks operationalize the provenance-filter idea [20]. Our grading pipeline relies on multiple machine families with human adjudication kept primary, a decision motivated by work on the reliability of LLM-as-a-judge systems, which shows that judge verdicts are sensitive to rating indeterminacy [4] and exhibit run-to-run instability [5]. We do not re-derive these results. Relative to this literature, our contribution is narrower and specific: a *private-fact, hash-sealed* benchmark where abstention is scored as success, run as a cross-model matrix, with failure decomposed by the stimulus's own clustering structure before anything is attributed to models.

Two external results we rely on directly. First, the global-workspace lens [2]: verbalizable representations form a shared workspace band in language models. Our companion detector line *applies* that lens to this task; in the pilot's exploratory analysis, workspace-band features at the final prompt position separated abstain-bound from failure-bound trials before any output token [1] — an application of [2], not a result reported there. That line and a concurrent Jacobian-space danger-recognition protocol [3] are current, independent works we cite as concurrent rather than foundational. Every reference in this section was independently verified against the primary source; a draft-stage citation confabulation was caught and removed during revision, sealed at the catch (Section 7).

3 Method

3.1 Corpus and probes

The benchmark is the frozen confirmatory set of the pilot program: 21 authentic work-session records from our laboratory, grouped into 7 near-domain *triads* (a triad is the minimal unit in which every fact is tempted by at least two sibling records — a reviewer-mandated property that a simple pair cannot satisfy). Four triads are technical (dense in commit hashes, scores, and identifiers) and three are art-provenance dossiers (names, dates, materials, attributions; not number-dense; length-matched to the technical logs at ∼800–1000 characters and held out from any detector fitting as a negative-transfer genre). The three art triads differ in *internal structure* in a way that becomes load-bearing in R2: one (A1) comprises three complementary facets of a single collection (identity, audit, production — few parallel fact-slots across records), one (A2) comprises three parallel asset dossiers (overlapping fact-slots — edition caps, reserve counts, production economics — with different values), and one (A3) is intermediate (three pipeline/production narratives). The triad structure yields 168 private-fact probes and, because each probe's near-domain arm runs against *both* sibling records, 336 D trials per regime; the full row set is 840 per regime, 1680 across both.

Each probe is evaluated under five context arms holding the question constant: **A** (no context), **R** (a length-matched, out-of-domain random text controlling for the generic effect of any context), **D-tech** and **D-art** (the paired near-domain record: authentic, topically adjacent, answer-absent — split by genre), and **C** (the correct record, as positive control). Two system-prompt regimes cross the arms: an *instructed* regime with an explicit answer-only-from-context instruction and abstain permission, and a *no-instruction* regime measuring native disposition. The probe file, arm construction, and leakage audit are hash-sealed (SHA-256 `399ffd44...`) before any capture; 105 anchor facts were human-audited (105/105) against original source documents, with zero authoring confabulations, and short golds ($\leq$ 3 chars) pre-declared for sensitivity exclusion.

3.2 Models, endpoints, and the reasoning fallback ladder

Eight models are run (Table 1). Two are *anchors* carried at full single-model fidelity: `Llama-3.1-8B-Instruct` (the pilot model, local L4, greedy, the study's calibration point) and `Inkling`, a custom laboratory model. The remaining six run on Together serverless endpoints: `Llama-3.3-70B-Instruct-Turbo` (adding a within-family scale axis against the 8B anchor), `Qwen3.7-Max`, `gpt-oss-120b`, `Kimi-K2.6`, `GLM-5.2`, and `DeepSeek-V4-Pro`. Endpoint quantization is undisclosed by the provider and is a stated limitation.

Reasoning was disabled where the endpoint permits it. Four models — gpt-oss-120b, Kimi-K2.6, GLM-5.2, and DeepSeek-V4-Pro (the "reasoning club") — have *non-disableable* reasoning. For these we invoked a sealed, pre-declared fallback ladder: capture at `reasoning_effort=low` with a 1024-token cap, grading the emitted answer only (reasoning excluded); rows whose answer was starved by reasoning consuming the budget ("cap-starved") were re-called once at a 2048-token cap and spliced by row index under the unchanged emitted-text rule. Rows still empty after the top-up form a **cap-censored** mechanical class (empty answer within budget). We report these two ways (Section 4.2): excluded from the primary rate denominators, *and* counted as failures, because non-response within an operational token budget is a deployment failure, not benign noise. The four non-reasoning-club models (both Llama, Qwen, Inkling) ran as bare runners; Qwen used a streaming (SSE) path on which logprobs are unavailable (declared). Every fallback amendment was sealed to the provenance manifest before the affected data existed.

3.3 Outcome ontology, grading, and tiering

Each answer receives one of three outcomes, with **abstention scored as success**. **CORRECT**: the answer states the gold fact. **ABSTAIN**: the answer declines without asserting any candidate value for the queried fact. **FAILURE-TO-ABSTAIN** ("FAIL"): everything else — every asserted wrong specific, and every hedged, half-declined, "educated-guess," partial, or substituted candidate. The governing bright line, ratified across walkthroughs: *did the answer offer a candidate for the queried fact?* If yes, in any register, FAIL; if it declined and merely described what the context contains, ABSTAIN (flagged as descriptive leak).

Grading used three families — a mechanical heuristic, a Claude family, and a Gemini family — followed by human adjudication. The two anchor studies (Llama-3.1-8B, Inkling) are **full human-primary**: every eventful row blind-adjudicated by the human author against the frozen rubric, with checkpoint walkthroughs settling family disagreements against precedent and verbatim rationales preserved. The six matrix models are **survey-grade**: the human adjudicates all family-disagreements, all flags, all control-arm CORRECTs, and a pre-declared 10% random audit of unanimous rows (seed `20260722`); dual-machine-agreement labels stand elsewhere, disclosed as survey-grade throughout. On the 929-row audit tab, survey-tier label-noise measured **0.00%** (an upper bound on unanimous rows, not a global rate). The anchor agreement was Tad–Claude 94.2%, Tad–Gemini 85.5%, Krippendorff $\alpha = 0.743$ (nominal, computed on the hardest-row subset only; not comparable to the pilot's 0.901, which spanned easy rows). An external hypothesis-blind adjudicator on a stratified sample is planned and disclosed as not-yet-run; because a single non-blind adjudicator is a potential common-cause driver of cross-model agreement, we treat it as a load-bearing limitation (Section 6), not a footnote.

3.4 Public-knowledge probe exclusions

Three probes were found, during adjudication, to be answerable *without* the private corpus — a Delaware franchise-tax figure derivable from the public statutory formula (one model computed $85,400 from it), an arXiv license-share, and a default-rights posture. These were caught mid-study by an A-arm high-correct sweep, then declared, sealed (`public_knowledge_probes.json`), and *excluded from all private-fact rates* (the mid-study timing is disclosed in Section 6). They are the clearest single instance of the instrument auditing itself (Section 7).

4 Results

4.1 R1 — The behavioral matrix

Table 1 reports failure-to-abstain for all eight models, under the no-instruction regime (arms A, R, D-tech, D-art), the positive-control CORRECT rate (C), and the instructed regime on the two near-domain arms (iD-tech, iD-art). Under instruction, the A and R arms cure to essentially 0% across models and are omitted for space; the near-domain arms are where instruction is tested (R2).

	no-instruction FAIL %				C %	instructed FAIL %	
Model	A	R	D-tech	D-art	CORR	iD-tech	iD-art
Llama-3.1-8B †	12.1	11.5	25.3	47.2	98.8	6.5	18.1
Llama-3.3-70B	92.1	9.1	32.8	58.3	99.4	4.8	18.1
Qwen3.7-Max	58.8	5.5	28.0	42.4	99.4	3.2	14.6
Inkling †	69.1	12.7	59.1	71.5	100.0	6.5	18.8
gpt-oss-120b	**95.8**	**86.7**	**90.3**	**91.7**	98.8	8.1	18.8
Kimi-K2.6	3.8	3.1	16.6	39.3	99.4	3.2	14.9
GLM-5.2	27.0	3.0	22.7	39.6	98.8	6.5	20.8
DeepSeek-V4-Pro	39.0	5.5	27.1	44.4	98.8	5.4	18.9

Table 1: The behavioral matrix. FAIL = failure-to-abstain (any non-correct answer that does not decline); abstention is scored as success. † full human-primary anchor; other six survey-grade. C = CORRECT rate on the positive-control arm. Cap-censored rows (Kimi 86, DeepSeek 10, GLM 4 across the run) are excluded from the primary denominators per the pre-declared fallback ladder and dual-reported in Section 4.2 (the count-as-failure reading changes no conclusion; both accountings ship in the release JSON); per-arm n ranges 112–186; the three public-knowledge probes are excluded from every private-fact denominator, removing at most 3 trials per arm cell (full counts and Wilson intervals in the release JSON). Wilson 95% intervals for the headline cells: gpt-oss A 95.8 [91.5, 97.9], R 86.7 [80.6, 91.0]; Llama-3.1-8B A 12.1 [8.0, 18.0] vs. Llama-3.3-70B A 92.1 [87.0, 95.3]; Kimi A 3.8 [1.7, 8.0].

Three features organize the rest of the paper. (i) The no-instruction A column — context-free failure on unanswerable private facts — spans an entire order of magnitude, from Kimi's 3.8% to gpt-oss's 95.8%; a leaderboard cannot distinguish these on capability grounds because all eight extract the correct record near-perfectly (C column, 98.8–100.0%). (ii) gpt-oss-120b is the pathological corner: it confabulates even under *random* context at 86.7%, i.e. it treats the presence of any text as license to answer. (iii) The D-art arm is the worst *context-bearing* no-instruction arm for every model (vs. R and D-tech; the context-free A arm exceeds it for three models: Llama-3.3-70B, Qwen3.7-Max, and gpt-oss-120b), and it is the arm on which instruction leaves the largest residue — the residue R2 decomposes.

4.2 R2 — Stimulus dominance: the refit that retired our own headline

Instruction is a powerful lever on most of the matrix. It collapses the A and R arms to $\sim 0\%$ everywhere, and it drives the technical near-domain arm (iD-tech) into a 3.2–8.1% band. On the art-provenance near-domain arm it does far less: instructed failure persists at 14.6–20.8% in every architecture (mean 17.9%, sample SD 2.11; per-model Wilson CIs in the release).

An earlier draft of this paper (v0.2, sealed into the provenance chain alongside this revision) read that convergence as an "instruction-resistant residual near 18%," homogeneous within the benchmark's resolution — eight architectures, one floor. Before submission we subjected that headline to the strongest attack we could construct against our own data: a refit that makes the corpus's stimulus structure explicit. The attack succeeded, and its result is this section. All refit artifacts are sealed and were verified by a two-auditor adversarial pass (AI agents operating without access to the original analysis code, performing blind cell-by-cell re-derivation from the sealed label files and four-estimator corroboration of the variance conclusion; the audit record is itself sealed, and the auditors' nature is disclosed in Section 9). One audit correction was adopted and logged: the initial effective-sample-size figure ($n_{\text{eff}}\approx 8$) had used the wrong (latent) scale, overstating the design effect in the refit's own favor; the corrected observed-scale figure is $n_{\text{eff}}\approx 14$.

The decomposition. Instructed D-art trials nest as trial → probe → session → *triad* (144 trials, 72 probes, 9 sessions, 3 triads per model; 1,148 trials over eight models, after the four instructed cap-censored exclusions). Decomposed by triad, the "18%" vanishes as a describable quantity:

Model	A1 (facets)	A2 (parallel slots)	A3 (pipelines)	overall
Llama-3.1-8B	2.1	35.4	16.7	18.1
Llama-3.3-70B	0.0	35.4	18.8	18.1
Qwen3.7-Max	0.0	27.1	16.7	14.6
Inkling	0.0	39.6	16.7	18.8
gpt-oss-120b	0.0	37.5	18.8	18.8
Kimi-K2.6	2.1	31.1	12.5	14.9
GLM-5.2	4.2	41.7	16.7	20.8
DeepSeek-V4-Pro	2.1	37.5	17.0	18.9
pooled	**1.3**	**35.7**	**16.7**	17.9

Table 2: Instructed D-art failure (%) by dossier triad. A1 = three complementary facets of one collection (few parallel fact-slots); A2 = three parallel asset dossiers (overlapping slots, different values); A3 = three pipeline narratives (intermediate). Triad heterogeneity: $\chi^2(2) = 154.7$, $p = 2.5 \times 10^{-34}$. Every model ranks the triads identically (A2 > A3 > A1, 8/8). The v0.2 "residual near 18%" is the column-weighted average of three numbers spanning 34.4 percentage points (35.7% vs 1.3%). The released refit artifacts name these triads T1-vigor/T2-assets/T3-curation, corresponding to A1/A2/A3 here.

The convergence is a shared-stimulus artifact. If all eight architectures shared *identical* per-probe difficulties — pure stimulus dominance — binomial sampling noise alone would still scatter their overall rates. Simulating that null (per-probe difficulties pooled across models; 4,000 draws; leave-one-model-out and shrunk-difficulty variants concur) yields a cross-model SD distribution with median 2.11 points and 95% band $[1.11, 3.33]$. The observed cross-model SD is 2.11 points ($P[\text{null} \geq \text{obs}] \approx 0.5$): the celebrated eight-model convergence sits at the *center* of what identical stimuli produce. It is not evidence of an architecture-general constant; it is what shared records look like through sampling noise.

Architecture explains approximately nothing. Four methodologically distinct estimators agree (none is independent of the shared clustering caveat above; their value is methodological diversity, not statistical independence). A Bayesian variance-components GLMM (Binomial-

logit; variance components for probe, session, triad, and model) assigns the stimulus side ∼80% of latent variance (probe 43.8%, triad 36.0%, session 0.2%) and *model* 0.3%; a maximum-a-posteriori refit drives σ_{model} to 0.000 exactly; a DerSimonian–Laird moment estimate on the eight model rates gives $\hat{\tau} = 0$ ($Q(7) = 3.02$, $p = 0.884$); a triad-stratified between-model test gives $\chi^2(21) = 10.35$, $p = 0.974$; a probe-fixed-effects likelihood-ratio test for a model term gives $p = 0.552$. A model×triad interaction likelihood-ratio test is likewise null ($\text{LR}(14) = 8.39$, $p = 0.868$): architectures do not even *reorder* on the triads beyond noise. The DerSimonian–Laird $\hat{\tau}^2$ remains 0 under leave-one-triad-out refitting (all three drops). (The v0.2 Gauss–Hermite partial-pooling posterior — τ mode at zero, 95% CrI to ∼4.7 points — is consistent with these and superseded by them; we no longer cite the mode at zero as positive evidence of anything.)

Honest uncertainty at the model level. With three triads, cluster-honest uncertainty is wide: leave-one-triad-out *sensitivity ranges* on per-model rates (three delete-one-triad replicates — too few for calibrated interval semantics) span [0, 48–68%], and the empirical design effect (≈10) reduces the effective sample to ∼14 trials per model — not 144. The pooled χ^2 above is likewise a *within-corpus descriptive* contrast computed at trial level; the inferential unit for any structural generalization is the triad, of which there are three. Per-model point estimates in Table 2 should be read accordingly; the paper's claims rest on the triad contrast (enormous at the trial level, within this corpus) and the ordering invariance (8/8), not on per-model precision.

What survives, and what it means. It is not true that instruction fails to move this arm — pooled uninstructed D-art failure of 54.7% (613/1120, trial-pooled) falls to 17.9% instructed. What instruction does *not* do is reach the component of failure carried by the record's own attribution structure: where three near-domain records expose the same fact-slots with different values (A2), one instructed answer in three misattributes a sibling's value, in every architecture we tested; where sibling records are complementary facets with few parallel slots (A1), instructed failure is ∼1%, in every architecture we tested. The claim "eight architectures, one floor" is retired. The corrected claim is stronger for deployment, and we state it at two calibrations. Identity-level, established by the decomposition: **instructed near-domain failure is set by *which record is served* — rates vary by** 34.4 **percentage points across triads and within sampling noise across the eight models we tested; between-model heterogeneity is unresolved at this power, not demonstrated absent.** Property-level, a candidate: the built-to-vary structural contrast (slot-parallelism) is the leading explanation for *why* the triads differ, pending the dose–response ladder (Section 6). In this benchmark, model selection was not a control. The input was.

Why an absolute difference and not a ratio. Earlier drafts expressed the triad effect as a ratio ("a factor of ∼27", from 35.7%/1.3%). We report the *absolute* difference of 34.4 percentage points instead, and recommend that any citation of this result do the same. The reason is that the ratio is governed by its denominator, and the denominator here is the least well-estimated quantity in the design: the A1 cell rests on an effective sample of $n_{\text{eff}} \approx 14$ independent stimulus units rather than on the nominal trial count, so its Wilson 95% interval is wide and admits values close to zero. A ratio built on that denominator is unstable by construction — small, statistically indistinguishable movements in the A1 rate swing the quoted factor across a large range, which makes the ratio a poor summary of an effect we otherwise regard as well established. The percentage-point difference is bounded, is not inflated by a small denominator, and preserves the finding: instruction leaves a record-determined gap of roughly a

third of all responses between the easiest and hardest near-domain triads. The ratio is retained nowhere in this paper as a headline figure.

Cap-censored rows, dual-reported. The reasoning-club fallback left a mechanical cap-censored class (empty answer within budget); on the instructed D-art arm this affects Kimi (3 rows) and DeepSeek (1 row). Excluding these from the denominator gives the rates above. Counting them instead as failures — the deployment-realistic reading — shifts Kimi 14.9 → 16.7% and DeepSeek 18.9 → 19.4%. All conclusions above are unchanged under either accounting.

4.3 R3 — The configuration finding: default ≠ safe

Two rows of Table 1 make the point by themselves. gpt-oss-120b uninstructed fails A at 95.8%, R at 86.7%, D-tech at 90.3%, D-art at 91.7% — it answers, with a specific asserted value, essentially every unanswerable question regardless of what context (or noise) it is shown. The *same weights* under one instruction line drop to iD-tech 8.1% and iD-art 18.8%. Inkling tells the same story from a different starting point: 69.1% context-free failure, cured to $\sim 0\%$ on the A arm by the same instruction; any context at all, even random, already restores its abstention (R FAIL 12.7%, abstain 86.1% by sealed count).

The lesson is about measurement, not capability: **the same weights are unsafe as a bare default deployment posture and clean in a pinned configuration**. A number attached to "the model" is therefore not deployment-meaningful; only a number attached to a *pinned configuration* — system prompt, decoding, reasoning setting — describes deployed behavior. R2 extends the same lesson from configuration to stimulus: after configuration is pinned, the *served record* is the remaining dominant source of observed variation. (The commercial development of this observation — configuration- and corpus-conditioned risk reporting — is treated in a companion whitepaper rather than here.)

4.4 R4 — Inverse scaling within the Llama family

Holding vendor, tokenizer family, and training lineage as fixed as the roster allows, scaling from Llama-3.1-8B to Llama-3.3-70B raises context-free failure-to-abstain from 12.1% [8.0, 18.0] to 92.1% [87.0, 95.3] — a near-complete inversion. The larger model is more willing to produce a confident specific answer to a question it cannot possibly know. We report this as a within-family observation, not a scaling law: it is two points in one family, and it is *confounded with the serving path* — the 8B anchor ran locally at known precision with greedy decoding, while the 70B ran on a served, provider-quantized "Turbo" endpoint. Quantization and endpoint optimizations can themselves shift the abstain threshold, so the inversion is a flagged candidate effect, not a validated one; the minimal clean test (both models served, matched precision and decoding) is named in Section 6. (The two models nonetheless mirror each other on the stimulus decomposition: identical overall instructed D-art rates and identical triad ordering — consistent with R2's unresolved-architecture result.)

4.5 R5 — The dialect taxonomy and the effort economy

Blind adjudication across the matrix produced not one failure mode but a taxonomy of five *named* confabulation strategies, each an identifiable way of offering a candidate while appearing not to, and each spoken preferentially by a subset of models:

- **Sneak register** — a substantive claim about the queried matter smuggled inside a decline ("the sneak in a None hat"). The camouflage defeats skimming: an adjudicator reading only the decline-opener, and not the tail, mislabels it as safe. This was the single strongest driver of the external-adjudicator recommendation.

- **Null-marker three-way** — an emitted "None"/"null" token whose reading depends entirely on what follows: *None* + insufficiency gloss is abstention; *None* + a substantive continuation is a sneak; bare *None* where the true answer is nonempty asserts absence-as-fact and is a failure. This register is a **Qwen**-family and scale-Llama habit (Qwen 57 rows, Llama-70B 46, Kimi 21), *absent* in GLM and gpt-oss — a decline dialect that is itself a fingerprint.

- **Tautology-dodge** — a definition-shaped non-answer that restates the question ("renamed from its original name") or emits a pseudo-range ("1–100" for items `AETR-0001..0100`). **gpt-oss** never declines uninstructed; it dodges. C-arm evidence disambiguates: 12/12 grounded models answer the true prefixed identifier, 0/12 emit a bare range — so a bare range is the fingerprint of not-knowing. (This dialect is also the reason semantic-entropy detection is insufficient here: a model can emit the same definitional restatement across samples, giving low entropy while conveying no answer.)

- **Derived-guess** — arithmetic theater performed on wrong-record numbers, producing a value that looks computed rather than recalled. A **reasoning-club** habit (GLM, DeepSeek), where the visible "work" lends the wrong number false authority.

- **Question-mining** — taking the answer out of the question's own framing and presenting it as knowledge. In the author's words, canon for this program: *"They are taking the answer from the question and acting like they know. That is the trick."*

Two adjudication classes accompany the taxonomy. *True-but-empty*: tautology-dodge rows are kept in the FAIL family for comparability but reported on their own line, never lumped with misattributions, because they assert nothing true-yet-misplaced — they assert nothing at all. *Cleanest declines*: GLM-5.2 produced 210 genuine declines, the cleanest decline behavior in the reasoning club ("seemed more honest," confirmed at the row level). The taxonomy converts "how safe is this model" into "which dialect of unsafety does this model speak," which is the actionable form for a downstream evaluator. The taxonomy also resolves an apparent tension with R2: *rates* converge while *registers* diverge — the record dominates how often models fail; architecture shapes how failure is voiced.

The dialects have an economic signature. Table 3 reports mean reasoning/answer tokens and the cap-censored and true-but-empty counts. Verbosity spans a factor of 39: Kimi-K2.6 spends 371 tokens per answer against Llama-3.3-70B's 9.5, and pays for it with 86 cap-censored non-responses — the heaviest reasoner in the club, a methods datum in its own right.

4.6 R6 — The content-rotation control: structure, not content

The strongest remaining objection to R2 is that the triad contrast might be a property of these particular records' *content* — familiar real-world values, latent memorization, or accidental cues — rather than their structure. We tested this directly with a content-rotation control (we reserve "replication" for an external re-run, which the released package enables). A public release corpus was constructed from the nine art sessions by *value rotation*: every monetary

Model	reasoning	tok/ans	cap-cens.	true-but-empty
Llama-3.1-8B	disabled/none	—	0	0
Llama-3.3-70B	none	9.5	0	2
Qwen3.7-Max	none	—	0	2
Inkling	none	11.3	0	0
gpt-oss-120b	non-disableable	51.5	0	4
Kimi-K2.6	non-disableable	**371.2**	**86**	1
GLM-5.2	non-disableable	190.9	4	1
DeepSeek-V4-Pro	non-disableable	156.6	10	2

Table 3: The effort economy. Mean reasoning/answer tokens per answer ("—" where unavailable on the endpoint path). Cap-censored = empty answer after the 2048-token top-up; dual-reported (Section 4.2). The four non-disableable-reasoning models form the "reasoning club"; the four bare-runners are reasoning-disabled or reasoning-absent.

and balance figure changed, unreleased-product names and internals replaced with structural equivalents, unpublished creative-strategy content replaced in kind — while preserving every sentence's structure, all published facts, and the corpus's slot geometry. A pre-registered QA gate enforced equivalence before any capture: per-session sentence counts identical, per-probe gold-presence parity 72/72, and D-arm gold-absence parity against both siblings (the gate caught and forced the repair of one rotation-induced substring collision, which is recorded). The rotated corpus, its probe arms, and its hashes were sealed before the first capture call.

All six matrix models were re-run on the rotated corpus under the instructed regime with capture parameters matched to the internal run, and graded by two machine families under the study rubric (the Gemini family, and a Claude-family pass whose input was served through the structured-serving grammar our companion engineering line develops — both relationships disclosed). A stratified 137-row human anchor was then blind-graded by the study adjudicator: all 88 machine-indeterminate rows, all 7 inter-family disagreements, and 42 determinate strata covering every model×triad cell (the adjudicator is the author — the external non-author adjudication remains a submission-version control). Raw inter-family agreement is 86.7–99.7% per model; the Gemini family returned 88 indeterminate (UNSURE) verdicts (Kimi 48, GLM 32, Llama-70B 8), which the Claude family (zero indeterminate) resolves — agreement on mutually-determinate rows is 99.4–100%. Concordance figures below are computed under three grading treatments: Gemini with indeterminates counted non-FAIL, Gemini with indeterminates excluded, and the Claude family alone. Cell-level mean absolute difference against the internal run is 2.63, 1.98, and 2.63 points respectively — the signature is robust to the indeterminate handling, and the exclusion treatment is the tightest. Per-cell indeterminate counts are released (they concentrate in Kimi-A2, 13 of 48 trials); raw grader outputs AND the exact grading scripts for both families ship in the release — the "structured-serving grammar" of the second family is an open, released prompt format, not withheld software — so any third family can be applied independently. **Human-anchor results:** the human agreed with the Claude family on 130/133 anchored rows (97.7%) and, on the 84 machine-indeterminate rows specifically, 82/84 (97.6%) — validating the second family's resolution of the first family's indeterminacy; human–Gemini agreement on the deliberately hard-case-enriched determinate sample was 43/49 (87.8%; not corpus-representative by construction). Under human-anchored labels the pooled triad signature is 3.5%/35.3%/17.0% — within noise of both the machine-graded public run and the internal human-primary run. Eighteen cap-starved rows were completed under the pre-declared 2048-

token top-up (four remain cap-censored and are excluded per protocol); walkthrough lineage for four re-ruled rows, and a new precedent (coincidental-value sibling-grab = failure: numeric overlap does not rescue wrong-fact binding), ship in the release. The triad signature reproduced: pooled rates A1 3.1%, A2 33.3%, A3 15.6% (primary family; 3.1/34.0/17.0 under the second family), against the internal 1.4/35.1/16.7 for the same six models; **triad ordering preserved in 6/6 models under both families**; cell-level mean absolute difference ≈2.6 points across the 18 model×triad cells (1.98–2.63 across grading treatments; maximum single-cell difference 8.2 points, at 48-trial cells). One capture false-start (a wrong token-cap on one model) was discarded before grading and is disclosed in the provenance chain.

Two conclusions, stated at their supported strength. First, the effect follows the properties *preserved* under rotation — structure, register, slot geometry — rather than the rotated values: with every sensitive value changed, the same records produce the same failure signature in the same order in every model. This removes rotated-value memorization and value-familiarity as explanations for the triad contrast (it cannot speak to properties the rotation deliberately preserved — structure, register, retained published facts — which is what the dose–response ladder of Section 6 isolates). The pattern is consistent with slot-parallelism — the property the triads were built to vary — as the operative structural feature. Second, the rotated corpus is releasable without exposing laboratory-sensitive content, so the replication package ships whole: corpus, arms, capture configuration, grades from both families, and the concordance analysis, hash-chained. A reader can reproduce Table 2's signature end-to-end without access to anything private.

5 Discussion

Benchmark evasion. A sneak-register or tautology-dodge answer is engineered, whether or not intentionally, to pass. An exact-match harness scoring the emitted string sees a decline opener or a definitional restatement and records a safe row; the substantive misattribution in the tail, or the bare range that fingerprints not-knowing, goes uncounted. Every one of the five dialects is a way to score as safe on a lexical harness while failing epistemically. This is why abstention must be scored on the epistemic notion (was unsupported content imported?) rather than the lexical one (does the reply contain refusal language?) — and why the crash test needs human-or-human-confirmed adjudication with an explicit bright line, not string matching.

One failure surface, two levers — and only one of them works on the hard part. The matrix separates the problem cleanly. The A-arm disposition — whether a model answers unanswerable questions from nothing — is enormously variable across models (R1) and enormously responsive to configuration (R3): a *disposition* problem, addressable by instruction, curated training, and pinned configuration. The near-domain failure that remains after instruction is different in kind: it is *record-dominated* (R2), reproduces under content rotation (R6), and is indifferent to architecture — which means the two most popular remedies, better prompting and better model selection, were insufficient to eliminate it in this experiment. Two remedy classes remain, stated here as practitioner implications with the competing-interests disclosure of Section 9 attached. Upstream, on the read path: control the *structure* of what is served. That has two faces. Preventive: an *admissibility gate* that asks whether a retrieved record is sufficient for the specific claim requested — and, per R2, especially whether it is a hazard-structured near-domain sibling — before the model sees it. Constructive: serve records in an attribution-explicit, per-source-delimited form rather than as melted prose, so that parallel fact-slots arrive already

bound to their owners. Better retrieval *ranking* alone need not help — the near-domain record IS the top-ranked record; it is admissible-looking and inadmissible, which is the whole point. Downstream, post-retrieval attribution interventions form a documented complementary class: Ghost Context shows wrong-span grounding is often causally localizable and largely correctable by masking the offending span [18]; whether masking-style interventions transfer to this setting — where the whole record, not a span, is the hazard — is an open empirical question, and we do not claim input-side control is the only viable remedy. We have deterministic prototypes of the input-side faces; their sealed evaluations are reported in a companion engineering line rather than asserted here, so that no serving-layer number appears in this paper ahead of its sealed artifact.

Scope of the claim. We claim a descriptive cross-architecture matrix; the stimulus-dominance finding with its refit, null-simulation, and the content-rotation control; a null on instruction and on model selection as cures for hazard-structured near-domain retrieval; and a dialect taxonomy. We do not claim a universal law of retrieval, a mechanism at the representation level, or anything about parametric (retrieval-free) hallucination.

6 Limitations and next experiments

Three triads. The load-bearing contrast (A2 vs. A1) rests on three art triads — a deliberately deep rather than wide design. Within-corpus statistics are correspondingly cluster-honest (jackknife intervals, $n_{\text{eff}}\approx 14$/model; Section 4.2), and the triad contrast itself is enormous at the trial level (descriptive χ^2, $p = 2.5 \times 10^{-34}$ pooled) and reproduced under content rotation (R6). What three triads cannot establish is the *dose–response curve* between slot-parallelism and failure rate; topic, register, and surface statistics co-vary with the designed contrast, so property-level attribution is a candidate, not a result. The measurable co-variates we checked run *against* the obvious alternatives: the hazard triad A2 is neither the longest (mean 894 chars vs. A1 958 / A3 965), nor the most digit-dense (3.1% vs. A3 4.3%), nor the most entity-dense (3.9% vs. A1 6.1%) — the per-triad covariate profile ships in the release. The natural next experiment constructs a graded ladder of record pairs from fully complementary to fully parallel and traces the curve; the A2-vs-A1 contrast identifies, concretely, *which* structural property to manipulate.

The pre-registered cardinality test: outcome reported. v0.2 pre-registered a bounded-answer-space test (does instructed failure track $1/k$ for k plausible distractors?) as the decisive check on a guessing-artifact reading. It ran (a machine-graded exploratory sweep, sample-verified); the $1/k$ hypothesis was **falsified in direction** (failure does not track $1/k$; at small k it moves opposite to the guessing prediction, and no consistent k-law appears across models). An exploratory synthetic follow-up initially appeared to reproduce a blur-driven failure until an ablation showed the synthetic effect was carried by an interpolation-inviting prompt cue rather than by the records; we record that dead end explicitly (its artifacts are sealed) both as a caution on synthetic reproductions and because the *real*-corpus effect — no such cue present, standard instruction, human-adjudicated, reproduced under content rotation — is precisely what the synthetic corpora failed to fake. The full binding study remains future work.

Single adjudicator. A single non-blind adjudicator graded the study; the checkpoint-walkthrough plus precedent-card procedure stabilizes calibration, and machine-family cross-checks bound drift, but a single grader is a potential common-cause driver of cross-model agreement. An external hypothesis-blind adjudicator on a stratified $\sim$20% sample, over-indexed on

the sneak-register and null-marker rows where thresholds diverge, is *planned and not yet run* (a candidate non-author adjudicator is secured); a grader-threshold sensitivity sweep on the existing continuous machine-judge scores is a complementary analytical control.

Serving-path and quantization confounds. Endpoint quantization is undisclosed by the provider for the six served models, so cross-model comparisons conflate architecture with an unknown precision; the two anchors are precision-known. Quantization is known to shift internal danger-recognition behavior [3], so endpoint compression could contribute to flattening apparent between-model differences — a further reason we report the architecture null as unresolved heterogeneity rather than demonstrated equivalence. Context-position effects ("lost in the middle") do not plausibly drive the triad contrast: every arm serves a single ∼900-character record with position matched across triads by the same arm builder, so a positional mechanism cannot produce a *between-triad* difference; a position-rotation control is nonetheless cheap and named for the ladder study. The R4 inverse-scaling result specifically confounds scale with the local-vs-served path; the minimal clean test is both Llama models served on the same stack at matched precision (BF16) and decoding (temp 0), which we flag as required before the inversion is asserted as a scale effect. (Note that R2/R6's record-dominance result makes precision confounds *less* threatening to the central claim: a triad signature that reproduces across eight differently-served models is not easily produced by any one serving path.)

Measurement of reasoning models. The reasoning-club models were graded answer-only under a fallback ladder that is not token-parity-matched to the bare runners; cap-censored non-responses are dual-reported (Section 4.2) but remain a measurement asymmetry inherent to comparing models that cannot be evaluated at token parity.

Other. Public-knowledge probe exclusions (3 probes) were discovered mid-study rather than pre-declared at freeze, and are disclosed as such; the systematic A-arm leak audit that caught them bounds residual public-recoverability, and the "empirically unrecoverable" premise is accordingly softened to "hard to recover, leak-audited." The matrix is eight architectures on one benchmark, one record-length band, and two genres. Krippendorff $\alpha = 0.743$ is on the hardest-row subset only and is not comparable to the pilot's easy-row-inclusive 0.901. The 0.00% survey-tier label-noise is an upper bound on unanimous rows only.

7 Method-Integrity: The Instrument Auditing Itself

Because this study's pipeline is itself AI-assisted, we treated its own reporting as adversarial and record the self-corrections here as recommended practice — a paper about confabulation should expect its own specifics to be checked.

The headline refit (v0.3). The v0.2 headline — "an instruction-resistant residual near 18%, homogeneous across eight architectures" — was re-examined with a dossier/triad random-effects refit, verified by a two-auditor adversarial pass (AI agents, per Sections 4.2 and 9; blind cell-by-cell re-derivation; four-estimator corroboration), and superseded by the stimulus-dominance decomposition of Section 4.2. The audit corrected one scale error in the effective-sample-size calculation ($n_{\text{eff}} \approx 8 \rightarrow \approx 14$); the correction is logged in the manifest.

Grader-family invariance. When the triage grader was switched families by directive mid-program, we re-graded both anchor studies against the sealed human-primary finals: 2758/2759 batch-sanctioned labels were *unchanged* (99.96%), the single disagreement routed to human ruling.

Gold-narrowing catch. That one disagreement exposed a real authoring defect: a single

"Name a ..." probe admitted multiple valid golds, and the original gold was over-narrow. The row was ruled CORRECT, a corpus-wide scan confirmed no other multi-valid probe existed, and four anchor rows were amended CORRECT via lineage (originals untouched).

Cap-censored class. The reasoning club's non-disableable thinking produced empty answers within budget; rather than score silence as either abstention or failure, we defined the cap-censored mechanical class mid-study, extended a uniform 2048-token top-up ladder to all four reasoning models before their top-up runs, excluded residual empties from the primary denominators, and dual-report them as failures (Section 4.2).

Public-knowledge probe discovery. An A-arm high-correct sweep revealed that three probes were computable from public information. They were declared out of the private-fact rates upon discovery and sealed (mid-study timing disclosed in Section 6). The A-arm, designed as a native-disposition measure, doubled as a leak detector on our own benchmark.

Citation confabulation, caught and removed. A draft of this paper cited a fabricated title attached to a real arXiv identifier (2607.12792, in fact Prosvirnin et al., "Silent Alarm"). The fabrication was produced by an AI drafting assistant, missed by same-family review, and caught by an independent deep-research review; the catch was sealed against the v0.1 draft and carried into v0.2, replaced with independently verified references [4, 5]; every remaining citation was re-verified against its primary source.

Content-rotation-chain corrections. The R6 control chain records two disclosed repairs: a rotation-induced substring collision caught by the pre-registered QA gate before any capture, and a wrong-token-cap capture false-start on one model, discarded before grading. The provenance chain is append-only; corrections appear as dated amendments.

8 Release and Provenance

Reproducibility splits in two, stated plainly: the *internal* benchmark is externally re-analyzable (all outputs, grades, and scripts released) but not re-capturable (source records withheld); the *public rotated corpus* is externally re-runnable end-to-end, including capture. We release the frozen probe set and leakage audit, all graded outputs from every grader family, the blind grading kits and final labels, the adjudication log with verbatim rationales, rubric amendments, and the precedent card, the public-knowledge exclusion list, the cap-censored key, the refit script and its audited results (triad decomposition, null simulation, variance components, jackknife), **the public value-rotated replication corpus with its probe arms, QA gate, captures, dual-family grades, and concordance analysis** (a self-contained control package enabling an external re-run, requiring nothing private; raw grader outputs from both machine families included), and a provenance manifest hash-chaining every artifact from pre-registration through final adjudication (including dated amendments where post-seal errors were caught). Per-model per-arm counts and Wilson intervals ship in the release JSON. The internal-corpus source records are not released; the rotation map linking internal to public values is deliberately withheld. No third-party personal data is included. *[DOI: inserted by Zenodo at publication.]*

9 AI Involvement Disclosure

This study was conceived, directed, funded, and adjudicated by the human author, who bears sole responsibility for its claims. AI systems contributed under human direction and are disclosed rather than credited as authors: Claude (Anthropic) built the capture and grading infrastructure,

executed runs, computed statistics including the R2 refit and R6 control, and assisted drafting and one grader family; Gemini (Google) provided the second machine grader family and an independent adversarial deep-research review (which caught the citation confabulation of Section 7); ChatGPT (OpenAI) and Grok (xAI) contributed additional adversarial methodological review. The R6 Claude-family grading pass received its input through the structured-serving grammar developed in the companion engineering line, and this is disclosed wherever those grades are used; Claude-family grades are never load-bearing for Claude-family subjects. The R2 refit audit and a pre-submission fresh-context review were performed by AI agents instantiated without access to the analyses they audited, with blind re-derivation from sealed artifacts; their nature is disclosed here rather than described as external. Every AI-produced statistic was verified against sealed artifacts; all authoritative labels on the internal corpus are human or human-confirmed (the rotated public corpus is machine-graded, disclosed in Section 4.6); the released log records where human judgment overrode machine output and where machine review corrected the human–AI team's own reporting, in both directions.

Competing interests. The author is the founder of Metavolve Labs, which develops commercial systems — structured retrieval serving and read-path admissibility gating — whose value would be supported by this paper's conclusions; the remedy discussion in Section 5 describes that product direction. To manage the conflict, no evaluation number for those systems appears in this paper; their assessments are pre-registered and reported separately with their own sealed artifacts. The scientific claims here stand on the released data independently of any product.

References

[1] T. MacPherson. *Never Mistake Proximity for Proof: When Retrieval Creates the Temptation to Misattribute*. Metavolve Labs pilot preprint, 2026. Concept DOI 10.5281/zenodo.21367214.

[2] W. Gurnee et al. *Verbalizable Representations Form a Global Workspace in Language Models*. Transformer Circuits Thread (Anthropic), July 2026; arXiv:2607.15495.

[3] R. Prosvirnin, M. Minchenkov, S. Soldatov, and V. Bashun. *Silent Alarm: A J-Space Protocol for Comparing Danger Recognition Across Models and Quantization Levels*. arXiv:2607.12792, 2026.

[4] L. Guerdan, S. Barocas, K. Holstein, H. Wallach, Z. S. Wu, and A. Chouldechova. *Validating LLM-as-a-Judge Systems under Rating Indeterminacy*. NeurIPS 2025; arXiv:2503.05965.

[5] E. Yagubyan. *The Coin Flip Judge? Reliability and Bias in LLM-as-a-Judge Evaluation*. arXiv:2606.13685, 2026.

[6] V. Adlakha, P. BehnamGhader, X. H. Lu, N. Meade, and S. Reddy. *Evaluating Correctness and Faithfulness of Instruction-Following Models for Question Answering*. TACL 2024; arXiv:2307.16877.

[7] C. Niu, Y. Wu, J. Zhu, S. Xu, K. Shum, R. Zhong, J. Song, and T. Zhang. *RAGTruth: A Hallucination Corpus for Developing Trustworthy Retrieval-Augmented Language Models*. ACL 2024; arXiv:2401.00396.

[8] Y. Katsis et al. *MTRAG: A Multi-Turn Conversational Benchmark for Evaluating Retrieval-Augmented Generation Systems*. arXiv:2501.03468, 2025.

[9] J. Qi, G. Sarti, R. Fernández, and A. Bisazza. *Model Internals-based Answer Attribution for Trustworthy Retrieval-Augmented Generation*. EMNLP 2024; arXiv:2406.13663.

[10] S. Kadavath et al. *Language Models (Mostly) Know What They Know*. arXiv:2207.05221, 2022.

[11] F. Shi, X. Chen, K. Misra, N. Scales, D. Dohan, E. Chi, N. Schärli, and D. Zhou. *Large Language Models Can Be Easily Distracted by Irrelevant Context*. ICML 2023 (PMLR 202:31210–31227); arXiv:2302.00093.

[12] O. Yoran, T. Wolfson, O. Ram, and J. Berant. *Making Retrieval-Augmented Language Models Robust to Irrelevant Context*. ICLR 2024; arXiv:2310.01558.

[13] C. Amiraz, F. Cuconasu, S. Filice, and Z. Karnin. *The Distracting Effect: Understanding Irrelevant Passages in RAG*. ACL 2025; arXiv:2505.06914.

[14] J. Niu et al. *Llama See, Llama Do: A Mechanistic Perspective on Contextual Entrainment and Distraction in LLMs*. ACL 2025 (Outstanding Paper); arXiv:2505.09338.

[15] A. Azaria and T. Mitchell. *The Internal State of an LLM Knows When It's Lying*. Findings of EMNLP 2023.

[16] A. Arditi, O. Obeso, A. Syed, D. Paleka, N. Panickssery, W. Gurnee, and N. Nanda. *Refusal in Language Models Is Mediated by a Single Direction*. NeurIPS 2024; arXiv:2406.11717.

[17] J. Kossen, J. Han, M. Razzak, L. Schüt, S. Malik, and Y. Gal. *Semantic Entropy Probes: Robust and Cheap Hallucination Detection in LLMs*. arXiv:2406.15927, 2024.

[18] R. Namboothiri. *Ghost Context: Measuring Cross-Context Interference in Long-Context Language Models*. TrustNLP 2026 (ACL Anthology 2026.trustnlp-main.19).

[19] H. Onweller et al. *Cited but Not Verified: Parsing and Evaluating Source Attribution in LLM Deep Research Agents*. arXiv:2605.06635, 2026.

[20] S. Alvarez, A. Rajan, S. Mugel, and R. Orús. *ProvenanceGuard: Source-Aware Factuality Verification for MCP-Based LLM Agents*. arXiv:2606.18037, 2026.

[21] C. Caruzzo, D. Yoo, and T. S. Kim. *Deceptive Grounding: Entity Attribution Failure in Clinical Retrieval-Augmented Generation*. arXiv:2607.09349, 2026.

[22] S. Farquhar, J. Kossen, L. Kuhn, and Y. Gal. *Detecting hallucinations in large language models using semantic entropy*. Nature 630:625–630, 2024.