Summary Is Not Enough: Source-Blind LLM Judges Mistake Faithful Citation for Hallucination

Maryam Fooladi[1,*], Federico Bottino[1]

[1]*Kakashi Ventures Accelerator (KVA), Turin, Italy*

Abstract

LLM-as-judge hallucination scoring is standard practice, yet when the judge sees only a brief ground-truth summary rather than the source the candidate model read, it cannot verify source-grounded citations and systematically flags faithful ones as fabricated. A four-route audit — a deterministic identifier-membership check over all 252 outputs (zero identifier-level fabrications), a sampled rationale review, and two human raters (one blind) — establishes that the judge's hallucination dimension tracks output citation style, not identifier-level fabrication; we release the deterministic check (≈40 LOC) as a routine supplement for source-grounded LLM-judge pipelines. The structured-context ablation that surfaced the artifact is an informative negative: with the contaminated dimension removed, no composite-level effect of representation survives, and the one dimension where structure does help — typed-binding fidelity — is supported by the source-access audit: human raters diverge sharply from the judge on hallucination but not on typed-binding. An external benchmark arm on FiQA (450 answers, judge given full source access) supports the verification-oriented lesson: source-tagged context makes answers mechanically checkable even when answer quality is unchanged. The lesson is twofold: for evaluation, give the judge source access or a deterministic membership check before trusting its hallucination scores; for system design, structured representation is a per-task switch — typed-binding and multi-hop reasoning — not an architectural default.

Keywords

LLM-as-judge evaluation, hallucination measurement, context representation, structured context, ablation study, marketing reasoning

1. Introduction

LLM-as-judge evaluation is standard practice for scoring open-ended outputs, including for hallucination: a judge reads a candidate output and a brief ground-truth summary and rates whether the output fabricates content. This paper reports a failure mode in that practice, and the structured-context experiment that surfaced it.

We set out to test a common architectural premise in LLM-based marketing automation: that context, given in sufficient quantity, is *enough* — that its shape (prose, JSON, a typed object graph) is presentation detail. We tested it with a four-condition ablation over fixed data, each output blind-judged on a four-dimension rubric (correctness, typed-binding fidelity, reasoning chain, hallucination). The hallucination dimension carried most of the apparent between-condition signal; on audit, it was not measuring fabrication.

This paper makes two contributions, one methodological and one empirical.

An LLM-judge measurement artifact (primary). We show, through a four-route audit, that the hallucination dimension as scored by an LLM judge without source access tracks output *citation style*, not model fabrication. The audit triangulates a deterministic identifier-membership check across the full 252-output corpus (zero fabrications), a sampled judge-rationale review, and two human-rater passes (one blind to the hypothesis), all converging. The mechanism is structural: when the candidate model cites source content present in the source but absent from the judge's brief ground-truth summary, the judge cannot verify it and flags it as unsourced — penalizing faithful citation at a rate governed by output

CLiC-it 2026: Twelfth Italian Conference on Computational Linguistics, September 14–16, 2026, Palermo, Italy

*Corresponding author.

maryam@kakashi.ventures (M. Fooladi); federico@kakashi.ventures (F. Bottino)

style. We argue this is a source-access failure mode alongside the known position- and verbosity-biases, give the conditions under which it arises, and release a deterministic source-membership check ($\approx$40 LOC) as a routine supplement to LLM-judge pipelines on dataset-grounded tasks (§4.4, §5.1).

A structured-context ablation as an informative negative (secondary). The experiment that surfaced the artifact is reported in full because its corrected result is useful to system designers. With the contaminated dimension removed and canonical q21–q23 re-run, no composite-level contrast against prose survives at $p < 0.05$, and the schema preamble adds nothing detectable — the composite effect of representation is best read as null. The surviving effects are dimension- and task-specific: typed-binding fidelity — the dimension the GO hypothesis pre-specifies — clears zero on a paired test ($\Delta = +0.26$, $p = 0.029$), while the task-level patterns (paired d up to $+0.87$) remain exploratory. The takeaway is narrower than "structure wins": structured representation is a per-task switch, not an architectural default (§4, §5).

2. Organizational Belief Systems and Growth Objects

We use *Organizational Belief System* (OBS) to refer to the structured set of claims an organization treats as operationally true, together with the entities, evidence, provenance, and relations that support those claims. Growth Objects are our concrete representation of an OBS in the marketing domain; in this view, the problem is not merely how to serialize context, but how to preserve the identity and evidential structure of organizational beliefs when they are passed to an LLM.

A Growth Object (GO) is a canonical typed unit of marketing knowledge. The canonical form decomposes any marketing situation into five top-level object classes:

- **Channels** — distribution and communication surfaces.
- **Product** — what is offered; axes: lifecycle (new/old), provenance (candidate/real/synthetic).
- **Audience** — the targeted population; provenance: real (observed) or synthetic (generated).
- **Action** — interventions on an Audience through a Channel for a Product (campaigns, structured tests; real or synthetic). Status: candidate, executed, or ongoing.
- **Evidence** — observed data resulting from Actions.

A downstream Observation layer aggregates Evidence into Insights and Sales Leads. Instances are registered in a typed **Library**: for each class, a registry of named identifiers carrying typed attributes, accessed through typed traversals over the schema (e.g. "for Audience A, return all real Actions on Channel C whose Evidence supports a lifecycle transition for Product P"). The pattern is analogous to typed graph-query traversals such as Cypher or SPARQL property paths; we do not claim formal equivalence in expressive power.

Four structural properties are relevant to the experimental claim and are progressively exposed or stripped across the conditions. *Class disambiguation is explicit*: a synthetic Audience is not silently equated with a real one, nor a candidate Action with an executed one. *Relations are typed*: an Action carries `audience_id`, `channel_id`, `product_id`; Evidence carries `action_id`; Sales Leads carry `audience_id` and `product_id`. *Display-name uniqueness is not enforced*: two instances of a class may share a name while differing on id, lifecycle, or provenance, so the model must resolve entities by typed identifier — a property real marketing data has. *Queries compose*: compound questions resolve as traversals over the Library rather than free-form retrieval over prose. The other conditions strip subsets of these properties and contrast which components carry the effect.

3. Experimental Setup

Model. Claude Sonnet (`claude-sonnet-4-20250514`), no fine-tuning, no retrieval augmentation beyond the input context. Temperature 0.3, `max_tokens` 1500.

Table 1
The four input representations.

Condition	Description
`prose`	The situation serialized as natural-language prose. No JSON, no schema.
`prose_plus_json`	Prose serialization with the raw JSON Library appended. No GO framing block.
`go_no_preamble`	The raw JSON Library alone, no prose, no GO framing block.
`go`	The full GO representation: GO framing block (schema preamble + typed-traversal instruction) + JSON Library.

Underlying data. Ten marketing situations, hand-authored by the first author, covering Channels, Products, Audiences, Actions, and Evidence. Each situation is a JSON document conforming to a published schema (in the data release), with explicit typed bindings and ground-truth notes for scoring. The dataset is synthetic by design: the experiment tests representation under fixed content, so realism is not a confound; synthetic authorship eliminates confidentiality concerns and improves reproducibility, leaving external validity to replication on real data (§7, §4.5). The full dataset, scoring and analysis scripts, scored outputs, audit scripts, and the FiQA-arm pipeline (preparation, serializers, generation, scoring, analysis) are available to reviewers on request and will be released publicly upon publication.

Task battery. Eighteen questions across five task types:

- Audience–Action coherence (4 questions)
- Real vs synthetic discrimination (4 questions)
- Multi-hop attribution (3 questions)
- Commercial reasoning under a proprietary private-banking client-rating framework (4 questions). We treat it as a domain-specific reasoning task class; the rating attributes are documented in the data release.
- Lead qualification from Observation data (3 questions)

Each question is mapped to one or more situations and carries a ground-truth summary used by the judge. Question IDs in `question_map.json` are non-contiguous (q01–q23 with gaps); the gaps reflect pilot items removed during dataset construction, not part of the reported battery.

Conditions. Four input representations of the same underlying situation data:

Character counts for a representative question were 1,933 / 4,041 / 2,638 / 4,062 across the four conditions; token budgets are not artificially matched. The design contrasts (a) JSON access, (b) JSON without prose, and (c) the GO framing block — contrasts rather than isolations, since token count and prose framing are not held constant across all comparisons.

Runs. 252 scored runs total. The base battery is 18 questions $\times$ 4 conditions $\times$ 3 runs $= 216$; three questions (q06, q08, q11) received an additional 3 runs per condition (36 total) from earlier preliminary subsets, kept because the data is sound, giving per-condition $n = 63$. Marginal-mean analyses use all 252 runs. Paired analyses use the 54 base question-run sets (18 questions $\times$ 3 runs, matched within (`question_id`, `run_index`) across all four conditions); the 9 extra q06/q08/q11 runs per condition are excluded from paired tests as they were collected outside the original matched-pair plan. The task-type breakdown in §4.3 is reported question-balanced (each question one vote) as the primary analysis and run-weighted in the data release for v1.1 comparability; the qualitative pattern is the same.

Dataset corrections. We identified and corrected one inconsistency in `sit_10` (Lead D score 0.71 → 0.78); q21–q23 were re-run under canonical ground truth, with q22/q23 being the correction-sensitive items (q21 was unaffected), question text, *and* judge ground-truth summaries. The headline composite contrast `go_no_preamble` − `prose` moves from $\Delta = +0.12$, $p = 0.010$ (pre-correction) to $\Delta = +0.12$, $p = 0.060$ (canonical) — i.e. it no longer clears zero. We report the canonical corpus throughout; the pre-correction values are retained only where noted, as a documented example of how a small ground-truth misalignment moved a composite-level significance verdict, which is itself an instance of the judge-side fragility this paper documents (§5.1). Full mechanics are in Appendix A.

Scoring. Each output is blind-judged by Claude Opus 4.7 (`claude-opus-4-7`) on a four-dimension 1–5 rubric: **correctness** (does the answer match the ground truth?), **typed-binding fidelity** (does it respect object identity, provenance, and status distinctions?), **reasoning chain** (is the reasoning exposed or merely asserted?), and **hallucination** (does it fabricate entities, metrics, or relations not in the situation? — reverse-coded, higher = less hallucination). The judge sees the question, the ground-truth summary, and the answer, but not the underlying situation data — an intentional v1.1 choice to keep the judge prompt small; §4.4 documents its consequences for the hallucination dimension. The blind protocol prevents explicit condition-label anchoring, though output style (typed identifiers like `act_*`, `ev_*` in structured-condition outputs) may partially reveal the representation. Our **primary composite** is the unweighted mean of correctness, typed-binding fidelity, and reasoning chain; the hallucination dimension is reported alongside but excluded for reasons developed in §4.4. Controls: identical task wording across conditions, identical model checkpoint, and an append-only run log with full prompts, outputs, per-call timestamps, and exact API identifiers preserved for audit (part of the data release).

4. Results

We flag the reading order up front: the composite and dimension-level analyses (§4.1–§4.3) motivate the paper's primary result, the hallucination-scoring audit in §4.4, with §4.5 as an external benchmark arm. The composite paired contrasts (§4.1) are the confirmatory test; the dimension- and task-level analyses beneath them are exploratory, with one exception — the typed-binding contrast, the dimension the GO hypothesis pre-specifies, carries its own paired test (§4.2). The task-level d values carry no CIs or multiple-comparison correction. All results use the canonical corpus: q22 and q23 were initially scored under since-superseded ground-truth summaries and have been re-run under fully canonical situation data, question text, and judge summaries (§3, Appendix A). The correction sharpened the claims rather than weakening them — the composite-level contrast that was significant pre-correction no longer is (§4.1), so the composite effect is cleanly null and the substantive story is task-specific; that a two-question ground-truth correction flipped a significance verdict is itself an instance of the judge-side fragility this paper documents.

4.1. Headline composite (3-dimension)

The primary composite is the unweighted mean of correctness, typed-binding fidelity, and reasoning chain. Marginal means across all 252 runs, and strict paired contrasts ($n = 54$ matched (`question_id`, `run_index`) tuples, t-based 95% CI), are shown in Tables 2 and 3.

A percentile bootstrap (10,000 resamples, seed = 1202) agrees with the t-based CIs on the sign-of-zero question for all contrasts (see the data release). All reported aggregates use the data in `runs/scored_outputs.jsonl` at git tag `v1.2-results`; the analytic frame is persisted at `runs/composite_3dim.parquet`.

Three findings sit on this table. First, `prose_plus_json` − `prose` is indistinguishable from zero ($\Delta = +0.01$, $p = 0.749$): appending the JSON Library to prose neither helps nor hurts. Second, `go_no_preamble` − `prose` is the largest contrast ($\Delta = +0.12$) but, under canonical ground truth,

Table 2
Marginal means, 3-dimension composite (all 252 runs), canonical (q22/q23-corrected) corpus.

Condition	n	Composite	Std	Δ vs prose
`prose`	63	4.47	0.47	—
`prose_plus_json`	63	4.47	0.45	$+0.00$
`go_no_preamble`	63	4.54	0.39	$+0.07$
`go`	63	4.56	0.54	$+0.09$

Table 3
Paired contrasts, 3-dimension composite ($n = 54$ pairs), canonical (q22/q23-corrected) corpus. The final row directly tests the GO framing block (schema preamble + typed-traversal instruction).

Contrast	Δ	95% CI	d	p
`prose_plus_json` − `prose`	$+0.01$	$[-0.065, +0.089]$	$+0.04$	0.749
`go_no_preamble` − `prose`	$+0.12$	$[-0.005, +0.227]$	$+0.26$	0.060
`go` − `prose`	$+0.12$	$[-0.008, +0.242]$	$+0.26$	0.066
`go` − `go_no_preamble`	$+0.006$	$[-0.098, +0.110]$	$+0.02$	0.906

Table 4
Per-dimension marginal means ($\pm$ std). The hallucination row (‡) is reported for completeness but excluded from the primary composite (see §4.4).

Dimension	`prose`	`prose_plus_json`	`go_no_preamble`	`go`	Δ (`go` − `prose`)
Correctness	4.59 ± 0.61	4.52 ± 0.76	4.59 ± 0.71	4.49 ± 0.80	-0.10
Typed-binding fidelity	4.44 ± 0.76	4.57 ± 0.67	4.68 ± 0.54	4.67 ± 0.74	$+0.23$
Reasoning chain	4.37 ± 0.48	4.30 ± 0.51	4.37 ± 0.51	4.51 ± 0.50	$+0.14$
Hallucination (rev) ‡	3.05 ± 0.95	3.40 ± 1.02	3.22 ± 1.05	3.24 ± 1.03	$+0.19$

does not clear zero ($p = 0.060$); `go` − `prose` likewise does not ($\Delta = +0.12$, $p = 0.066$). Under the canonical corpus, then, *no* composite-level contrast against prose is significant at $p < 0.05$. Third, the direct test of the GO framing block — `go` − `go_no_preamble`, final row of Table 3 — is essentially exactly zero ($\Delta = +0.006$, 95% CI $[-0.098, +0.110]$, $d = +0.02$, $p = 0.906$): holding the JSON Library fixed, adding the schema preamble and typed-traversal instruction has no measurable effect on the composite.

Read together, these indicate that the composite effect of representation is best read as null under canonical scoring at this battery's power: no contrast separates from prose, and the GO framing block adds nothing detectable. The substantive story is therefore entirely below the composite, in the dimension- and task-level analyses (§4.2, §4.3).

4.2. By rubric dimension

‡ *On audit (§4.4), this dimension tracks judge artifact, not identifier-level fabrication: zero identifier-level fabrications across the 252-run corpus, zero source-absent flagged content across 20 sampled rationales. The audit does not establish a zero rate of metric-value or relational fabrication on this battery; those would require separate audits we did not run. Reported here for completeness; excluded from the primary composite in §4.1.*

The per-dimension means in Table 4 reconcile with the composite in Table 2: prose's composite is $(4.59 + 4.44 + 4.37)/3 = 4.47$, matching its composite row. Means and standard deviations should be read from the released per-dimension frame rather than recomputed from rounded table entries.

The composite hides four different stories. **Typed-binding fidelity is where structure pays off**: `go_no_preamble` reaches 4.68 and `go` 4.67 against prose's 4.44 (a $+0.23$–$+0.24$ marginal gap) —

Table 5
Marginal means by task type, question-balanced (3-dim composite, 1 vote per question).

Task type	n q.	prose	prose_plus_json	go_no_preamble	go	Winner
audience_action_coherence	4	4.42	4.42	4.44	4.56	go
commercial_reasoning	4	4.00	4.03	4.25	4.36	go
lead_qualification	3	4.59	4.59	4.59	4.37	3-way tie (excl. go)
multi_hop_attribution	3	4.33	4.44	4.63	4.59	go_no_preamble
real_vs_synthetic	4	4.75	4.69	4.71	4.76	go ($\approx$ prose)

Table 6
Paired Cohen's d by task type, on the 3-dim composite (condition $-$ prose, matched within (question_id, run_index)).

Task type	n pairs	d (p+json $-$ prose)	d (go_np $-$ prose)	d (go $-$ prose)
audience_action_coherence	12	$+0.00$	$+0.08$	$+0.46$
commercial_reasoning	12	$+0.09$	$+0.78$	$+0.87$
lead_qualification	9	$+0.00$	$+0.00$	-0.30
multi_hop_attribution	9	$+0.33$	$+0.76$	$+0.80$
real_vs_synthetic	12	-0.43	$+0.00$	$+0.00$

exactly the dimension the GO hypothesis pre-specifies, where exposed typed bindings should reduce errors that treat a synthetic Audience as real, a candidate Action as executed, or one Product variant as another. Unlike the composite, this contrast clears zero on the paired test: go $-$ prose on typed-binding over the 54 matched sets gives $\Delta = +0.26$, 95% CI $[+0.03, +0.49]$, paired $d = +0.31$, $p = 0.029$ (a percentile bootstrap CI, seed 1202, agrees: $[+0.04, +0.48]$), and the full ladder clears zero on this dimension (prose_plus_json $p = 0.019$, go_no_preamble $p = 0.008$, go $p = 0.029$). The effect is modest but it is the one place structured conditions separate from prose with a confirmatory test behind them. **Correctness is essentially flat**: a 0.10 spread (prose 4.59 to go 4.49), too small for a directional claim; per-run rationales suggest the framing block may occasionally treat typed traversal as sufficient when the ground truth needs further inference, but this is interpretation, not coded error analysis, and we flag it for replication.

Reasoning chain shows no consistent structured-condition pattern (all four conditions 4.30–4.51): what exposes reasoning is the question and the model's defaults, not the representation. **The hallucination dimension is contaminated**: all conditions sit at 3.05–3.40 on the reverse-coded scale, but the four audits in §4.4 establish it is not measuring fabrication on this battery — the between-condition spread tracks output citation style, not fabrication rate — so we report the numbers but do not interpret them and exclude the dimension from the primary composite.

4.3. By task type

The task-level breakdown is reported on the 3-dimension composite (correctness + typed-binding fidelity + reasoning chain). The **question-balanced** table (each question contributes one vote) is the primary analysis: it isolates the per-task-type effect of representation from the n-per-question imbalance introduced by the q06/q08/q11 extras (§3). The **run-weighted** table (each run contributes one vote) is reported in the data release for v1.1 comparability; the qualitative pattern is the same, with two row-level discrepancies flagged in that appendix.

The lead_qualification row reflects the canonical q21–q23 re-run, of which q22/q23 are the correction-sensitive items (§3, Appendix A).

Where structure helps (traversal and provenance tasks). The two strongest structured-condition signals are multi-hop attribution and commercial reasoning — the only task types with $d > 0.5$ across the GO contrasts (paired $d = +0.76$ to $+0.78$ for go_no_preamble $-$ prose and $+0.80$ to $+0.87$ for go $-$ prose). On multi-hop attribution all three structured conditions beat prose (go_no_preamble

4.63 vs prose 4.33); on commercial reasoning the prose-to-go Δ is $+0.36$, the largest in the question-balanced table. When the question requires following Evidence → Action → Audience → Channel → Product or applying a typed rating framework, exposed typed bindings — with or without preamble — reduce the rate at which the chain breaks. (The v1.1 multi-hop d of $+2.13$ was computed on the contaminated 4-dim composite; the 3-dim value of $+0.80$ is what should be cited.) Audience–action coherence sits in the middle: `go` wins (4.56 vs prose 4.42, $d = +0.46$) while `prose_plus_json` ties prose ($d = +0.00$), consistent with the typed-binding-fidelity story of §4.2.

Where it does not (judgment and provenance-tie tasks). On real_vs_synthetic the conditions are effectively tied (prose 4.75, `go_no_preamble` 4.71, `go` 4.76; `prose_plus_json` 4.69, whose paired $d = -0.43$ reverses a v1.1 directional finding that was a hallucination-dimension artifact). On lead_qualification — re-run under canonical ground truth — structure does not help and the GO condition performs worst on the one genuinely hard item: q22, once its premise was corrected (0.78 does satisfy "above 0.75"), sends every condition to ceiling and no longer discriminates, while on the hard chain-depth q23 `go` is worst (3.11 vs prose 4.11, `prose_plus_json` 4.22, `go_no_preamble` 4.00), over-applying the provenance lens at the expense of chain depth. Aggregated over the task type, prose, `prose_plus_json`, and `go_no_preamble` tie exactly (all 4.59, paired $d = 0.00$); only `go` moves, downward ($d = -0.30$, n.s.). Finally, `go` − `go_no_preamble` stays in a tight band across task types (-0.22 to $+0.11$), the one negative being lead_qualification — consistent with the near-zero direct contrast of §4.1.

4.4. The hallucination dimension does not, on audit, measure identifier-level fabrication

In the pre-correction analysis the hallucination column showed the largest dimension-level spread and carried most of the composite-level separation between `go` and prose; the pattern holds on the canonical corpus (means 3.05–3.40, §4.2). To probe the dimension, we ran four audits in increasing order of independence from the Opus judge (the deterministic Audit 2, a membership test over identifiers, is invariant to the q22/q23 score correction). The structural cause is the judge's input: it receives the question, the per-question ground-truth summary, and the answer — but not the underlying situation JSON. When the model cites source content (numbers, identifiers, strings) present in the situation file but absent from the brief summary, the judge cannot verify it and flags it as unsourced.

Audit 1 (judge-rationale audit, $n = 20$). Across 20 sampled rationales from low-scoring runs, every flagged item (113 total) was present in the source the model received; every sampled rationale contained at least one false-positive flag. **Audit 2 (full-corpus deterministic ID check, $n = 252$).** We extracted every identifier-shaped token (`act_*`, `ev_*`, `prod_*`, `aud_*`, `ch_*`, `obs_*`) from every output and matched against the situation Library: **zero outputs across the 252 runs contained an identifier not present in source.** This is a deterministic check over the full corpus, released as `audits/source_membership_check.py`. **Audit 3 (hypothesis-aware human rater, $n = 24$)** and **Audit 4 (blind human rater, $n = 24$).** Two KVA team members (non-authors) independently re-scored a stratified sample with access to the situation files; the second was blind to the hypothesis, condition labels, Opus's scores, and the first rater's scores. Both converged on a strong directional disagreement with Opus, concentrated on the hallucination dimension (Table 7). Full per-dimension statistics for Audits 1 and 3 and the inter-rater agreement analysis are in Appendix B.

Interpretation. Four audits, ordered by independence from the Opus judge, converge: the hallucination dimension as scored by Opus does not measure identifier-level fabrication on this battery. The identifier-level fabrication rate is zero (Audit 2); the substring-level pattern matches for the sampled rationales (Audit 1); and two human raters with source access rated less harshly than Opus with strong directional asymmetry — 21/3/0 and 22/2/0 splits, mean Δ of $+1.6$ to $+1.8$ (Audits 3, 4). The between-condition spread (3.05–3.40) tracks how each representation prompts the model to cite source content, not its fabrication rate. (Audit 2 does not extend to metric-value or relational fabrication, which would need separate audits.) Notably, the same source-access audit supports the typed-binding

Table 7
Audit 4: paired Opus vs blind rater (rater − Opus, t-based 95% CI, paired d, two-sided p, $n = 24$).

Dimension	Δ	95% CI	d	p
correctness	$+0.29$	$[+0.06, +0.52]$	$+0.53$	0.016
typed_binding	$+0.08$	$[-0.19, +0.36]$	$+0.13$	0.539
reasoning_chain	$+0.29$	$[+0.06, +0.52]$	$+0.53$	0.016
hallucination	$\mathbf{+1.62}$	$[\mathbf{+1.22}, \mathbf{+2.03}]$	$\mathbf{+1.68}$	$\mathbf{< 0.001}$

dimension: the blind rater does not significantly disagree with the judge on typed-binding ($\Delta = +0.08$, n.s., Table 7) while diverging sharply on hallucination ($\Delta = +1.62$), so the typed-binding gap of §4.2 is not contradicted by the audit and is unlikely to be a style-leakage artifact. We develop the general lesson — and the remedy — in §5.1.

4.5. External benchmark arm on FiQA

To address the synthetic-data limitation (§7), we ran a public-benchmark analogue of the representation question on FiQA, a financial question-answering retrieval dataset from the BeIR suite [1, 2]. FiQA carries no typed object graph, so this is not a replication of the GO experiment; it tests the form of the thesis the benchmark supports: does the serialization of *retrieved* context change how the model answers and attributes its sources? We sampled 50 test queries, pairing each query's gold (qrels-relevant) passages with corpus distractors (six passages per query), and rendered the same passage set three ways: `concat` (passages glued as prose), `numbered` (delimited and numbered, no source identifiers), and `tagged` (each passage labeled with its `doc_id`). Subject model, temperature, run count, and blind-to-condition judging mirror the main experiment (50 × 3 × 3 = 450 scored answers); attribution fidelity (gold vs. distractor reliance) replaces typed-binding fidelity. Crucially, the judge here receives *all six passages* (gold-labeled) rather than a brief summary — implementing the source-access remedy of §5.1 — so the faithfulness dimension is retained in the composite rather than excluded as in the synthetic arm.

The judge dimensions are flat. Composite means are 4.61 / 4.58 / 4.60 for `concat` / `numbered` / `tagged` (a 0.03 spread); no paired contrast clears zero on the composite or on attribution fidelity (150 pairs per contrast; all $p \geq 0.179$, $|d| \leq 0.11$), and the thesis-predicted attribution gain for `tagged` does not materialize even directionally ($\Delta = -0.05$ vs. `concat`, $p = 0.179$). A ceiling effect contributes (correctness, attribution, and faithfulness all exceed 4.6 on single-hop retrieval), but the substantive pattern matches the synthetic arm: representation does not move the *rate* of judge-scored answer quality. The conditions separate on the deterministic citation check, which needs no judge: in `tagged` the model cited a source `doc_id` in 98.0% of answers and a distractor in only 2.7%; `concat` and `numbered` cite no identifiers — necessarily, since none are exposed. We stress that this is not a behavioral gain to the model's credit; it is an *audit surface* created by the serialization. Tagging does not produce *better* answers on this task; it produces *checkable* ones — a `tagged` answer can be verified by membership-checking the cited `doc_id` against the gold set, exactly the verification primitive of §5.1. The arm also bounds the synthetic findings: on flat single-hop retrieval with no graph to traverse, the task-level gains of §4.3 disappear — structure earns its keep only where the task requires traversal or distinction-preservation.

5. Discussion

5.1. The primary finding: a source-access LLM-judge failure mode

The audit in §4.4 documents a measurement artifact that we believe generalizes beyond this battery. The LLM-as-judge literature catalogues judge biases that distort scores independent of output quality — notably position and verbosity bias [3]. The artifact here is a further failure mode, specific to *source-grounded* scoring: when a judge is asked to score hallucination but is given only a brief ground-truth summary rather than the source the candidate model saw, it cannot verify source-grounded citations

and systematically flags faithful ones as fabricated. The rate of mis-flagging is governed by output *citation style* — outputs that quote longer source strings are penalized more than outputs that cite short typed identifiers — not by any difference in fabrication. The conditions for the failure are precise and common: a closed, inspectable corpus; outputs that may legitimately cite corpus-only identifiers, metrics, or strings; and a judge whose reference text is narrower than that corpus. Any pipeline meeting them is exposed to this failure mode.

The remedy is concrete: give the judge the same source material as the candidate model, or supplement it with a deterministic source-membership check against the corpus. We release `audits/source_membership_check.py` as such a supplement: mechanical (≈40 LOC), fast (under 10s on this corpus), and applicable to any pipeline scoring identifier-grounded outputs against a typed reference; it does not cover metric-value or relational fabrication, which need a typed extension. We propose it as a routine diagnostic before trusting an LLM judge's hallucination scores on dataset-grounded tasks; in our case the unaudited dimension was the single largest driver of a composite-level "finding" (§4.1) that does not survive audit.

The concrete consequence for our own headline, as a cautionary example: in v1.1 the four-dimension composite made the full object-graph condition distinguishable from prose, with roughly half the gap from the hallucination dimension; with the dimension recognised as artifact and excluded, the claim does not survive (§4.1). A practitioner who had trusted the dimension would have shipped a representation choice on the strength of a judge artifact. The deterministic check would have caught it in seconds.

5.2. The secondary finding: representation is a per-task switch

Stripped to its claims, the ablation shows three things about representation (the hallucination dimension is treated in §5.1). (i) **The composite effect is best read as null under canonical scoring and current power**: no paired contrast against prose clears zero (the largest, `go_no_preamble` − `prose`, is $\Delta = +0.12$, $p = 0.060$), and `prose_plus_json` − `prose` is a clean null. (ii) **The schema preamble adds no detectable signal**: the direct `go` − `go_no_preamble` contrast does not separate from zero, so given the typed-bindings access the JSON Library already provides, the preamble adds no measurable composite-level value. (iii) **The effects concentrate where the audit supports the measurement**: typed-binding fidelity, which clears zero on its paired test ($\Delta = +0.26$, $p = 0.029$, §4.2; the blind rater agrees with the judge on this dimension, §4.4), and — exploratorily — the multi-hop-attribution and commercial-reasoning task types (paired $d \approx +0.76$ to $+0.87$).

The pattern across the five task types is consistent: structure helps where the question requires traversal (multi-hop attribution), application of a typed rating framework, or preserving a distinction (typed-binding fidelity); it does not help — and may slightly hurt — where the question rewards semantic inference over the world the data describes (lead qualification). GO is thus a representation switch to be selected per-task, not an architectural default. The cost-benefit is concrete: a $+0.09$ composite lift does not justify a large engineering investment, but a $+0.23$ lift on typed-binding fidelity and paired $d \approx +0.80$ on the two strongest task types are worth deploying where those properties matter. The shortest recommendation: use the JSON Library as the primary representation for typed-bindings or multi-hop tasks; do not merely append JSON to prose, which was a clean null ($\Delta = +0.01$, $p = 0.749$).

5.3. Why prose_plus_json is interesting (and why the v1.1 read of it does not survive)

The hybrid condition was added late as an ablation control for token count. A v1.1 reading favoured it on the strength of its (4-dim) win on real_vs_synthetic; that win does not survive the 3-dim recomputation (§4.3), where it is essentially tied with prose on that task type (paired $d = -0.43$) and a clean null at the composite level ($\Delta = +0.01$, $p = 0.749$). The hybrid remains a plausible hypothesis — prose carries pragmatic cues a pure schema lacks, JSON carries provenance prose lacks — but this experiment provides no positive evidence for it. Practically, systems already operating over prose can attach the structured representation as an extra context block: doing so does not detectably help on average, but

does not detectably hurt either.

6. Related Work

Our contribution sits at the intersection of three lines of work: structured retrieval over knowledge graphs, prompt-format and structured-generation sensitivity in LLMs, and LLM applications in marketing and CRM. We summarise each in turn, then identify the closest prior work.

Structured retrieval and knowledge-graph-grounded LLMs. A substantial literature replaces flat-text retrieval with graph-structured retrieval: GraphRAG builds an LLM-derived entity graph with community summaries [4], G-Retriever formulates subgraph retrieval as Steiner-tree optimization [5], and Zhang et al. survey the area [6]; our predecessor *Retrieval Is Not Enough* argues typed organizational knowledge needs explicit epistemic infrastructure [7]. We ask an orthogonal question: given that the knowledge already fits in context, does the *form* of the in-context representation matter for downstream reasoning?

Prompt format sensitivity and structured generation. Empirical work establishes that LLM performance is highly sensitive to input formatting, independent of content: Sclar et al. show swings of up to 76 accuracy points from meaning-preserving prompt-format perturbations [8]. On the legal-document side, "The Hidden Structure" empirically demonstrates that explicit input formatting (Markdown vs. plain text) lifts exact-match accuracy on the CUAD legal QA benchmark by approximately 20 percentage points on GPT-4.1, with an additional 10–13 points from system-prompt cues that disclose the structured nature of the input [9]. On the output side, typed schemas constrain generation: DSPy abstracts LLM pipelines as typed-signature modules [10], Toolformer trains LLMs to invoke typed APIs [11], and Geng et al. benchmark constrained decoding against real-world JSON schemas [12]. Our work studies the less-explored input side, where the typed schema sits in the prompt rather than the output target.

LLM applications in marketing and CRM. Surveys catalogue LLM use in marketing automation, personalization, and content generation [13], and applied work explores hybrid ML segmentation plus RAG-grounded message generation in financial services [14]. This literature treats business context as text or lightly structured retrieval rather than a typed object graph; we ask whether that assumption is load-bearing, and find it is, but narrowly (typed-binding fidelity, multi-hop attribution, commercial reasoning; absent or slightly negative elsewhere).

Closest prior work. Two papers sit closest. Conceptually, Bottino et al.'s *Retrieval Is Not Enough* [7] is the direct predecessor and shares the premise that typed organizational knowledge needs explicit infrastructure; we differ by holding retrieval fixed (the full Library is in-context) and isolating representation per se. Methodologically, the legal-domain "Hidden Structure" study [9] is the closest empirical cousin (input-format effects on a domain QA battery); we differ in structural granularity — they vary Markdown vs. plain text, we vary prose vs. typed object graph with foreign-key bindings — and in the four-condition ablation isolating the GO framing block specifically.

LLM-as-judge methodology. Zheng et al. [3] and Liu et al. [15] document the human-LLM judge agreement methodology our human-rater validation extends, and surface position- and verbosity-bias failure modes; our §4.4 audit identifies a further one — judges lacking source access systematically penalize faithful citation. The deterministic identifier-membership check we use is, to our knowledge, not yet standard practice; we propose it as a routine supplement.

7. Limitations

Single-model evaluation. The claim is about representation, but the model is fixed at one Claude Sonnet checkpoint; cross-model replication would strengthen the result.

Synthetic dataset and single battery. The main corpus is hand-authored and small (eighteen questions across five task types): the task-type effects are large enough to be visible at this size (multi-hop and commercial-reasoning paired $d > 0.75$), but per-question variance is not, and expanding the battery is the cheapest way to strengthen the paper. Synthetic authorship is appropriate for a representation-effect study (§3) but limits external validity; the FiQA arm (§4.5) partially addresses this on public data, but has no typed object graph, so the typed-binding and traversal tasks remain untested outside the synthetic battery. Replication on anonymized real client data with genuine typed structure is the natural follow-up.

Rubric calibration on commercial-reasoning tasks. The commercial_reasoning task type scored lower than the others across all four conditions (4.00–4.36 on the 3-dim composite); this may reflect genuine task difficulty or rubric items that transfer poorly to the domain, and future work should validate the rubric against domain experts.

Judge stochasticity. The judge (default sampling) is not deterministic and each output was judged once, so judge stochasticity and model-output variability are confounded; a future replication should rescore a fixed subset multiple times.

Human-rater and hallucination-measurement validation. The §4.4 audit uses two raters at $n = 24$ each (Audit 3 hypothesis-aware, Audit 4 blind to hypothesis, condition, Opus's scores, and the first rater); both converge directionally with the deterministic ID check. Sample size and the lack of calibrated raters are the limitations; multi-rater validation at $n \geq 100$ is the priority extension. The audit establishes only that the judge's hallucination scores do not measure *identifier-level* fabrication (zero across 252 outputs), not a zero rate of metric-value or relational fabrication. The FiQA arm (§4.5) gives the judge full source access, but on a different task; a source-access re-run of the synthetic battery remains the cleanest resolution. Scoring sheets and the sampling script are in the data release.

Dataset correction history. One data-quality issue (a misaligned Sales-Lead score in `sit_10`) was identified during analysis; q21–q23 were re-run under canonical ground truth (q22/q23 being the correction-sensitive items) and all results use the canonical corpus (§3, Appendix A). The correction flipped the composite-level significance verdict — evidence of the fragility of small LLM-as-judge batteries, not an unresolved limitation.

Acknowledgments

This paper continues a line of work on what is load-bearing in modern AI systems, following *Retrieval Is Not Enough* [7]. We thank colleagues at KVA for the human-rater audits and discussion of the experimental design.

Declaration on Generative AI

During the preparation of this work, the author(s) used Claude (Anthropic) for LaTeX formatting and grammar/spelling assistance. The LLMs studied as objects of investigation (Claude Sonnet as system under test, Claude Opus 4.7 as judge) are described in Sections 3 and 4. The author(s) reviewed and edited all content and take(s) full responsibility for the publication's content.

References

[1] M. Maia, S. Handschuh, A. Freitas, B. Davis, R. McDermott, M. Zarrouk, A. Balahur, WWW'18 open challenge: Financial opinion mining and question answering, in: Companion Proceedings of The Web Conference 2018, 2018, pp. 1941–1942. doi:10.1145/3184558.3192301.

[2] N. Thakur, N. Reimers, A. Rücklé, A. Srivastava, I. Gurevych, BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models, in: Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks (NeurIPS), 2021.

[3] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, I. Stoica, Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena, in: Advances in Neural Information Processing Systems, 2023. ArXiv:2306.05685.

[4] D. Edge, H. Trinh, N. Cheng, J. Bradley, A. Chao, A. Mody, S. Truitt, D. Metropolitansky, R. O. Ness, J. Larson, From local to global: A graph RAG approach to query-focused summarization, arXiv:2404.16130, 2024. arXiv:2404.16130.

[5] X. He, Y. Tian, Y. Sun, N. V. Chawla, T. Laurent, Y. LeCun, X. Bresson, B. Hooi, G-Retriever: Retrieval-augmented generation for textual graph understanding and question answering, in: Advances in Neural Information Processing Systems, 2024. ArXiv:2402.07630.

[6] Q. Zhang, S. Chen, Y. Bei, Z. Yuan, H. Zhou, Z. Hong, J. Dong, H. Chen, Y. Chang, X. Huang, A survey of graph retrieval-augmented generation for customized large language models, arXiv:2501.13958, 2025. arXiv:2501.13958.

[7] F. Bottino, C. Ferrero, N. Dosio, P. Beneventano, Retrieval is not enough: Why organizational AI needs epistemic infrastructure, 2026. arXiv:2604.11759.

[8] M. Sclar, Y. Choi, Y. Tsvetkov, A. Suhr, Quantifying language models' sensitivity to spurious features in prompt design, or: How I learned to start worrying about prompt formatting, in: International Conference on Learning Representations (ICLR), 2024. ArXiv:2310.11324.

[9] C. Braun, A. Lilienbeck, D. Mentjukov, The hidden structure — improving legal document understanding through explicit text formatting, arXiv:2505.12837, 2025. arXiv:2505.12837.

[10] O. Khattab, A. Singhvi, P. Maheshwari, Z. Zhang, K. Santhanam, S. Vardhamanan, S. Haq, A. Sharma, T. T. Joshi, H. Moazam, H. Miller, M. Zaharia, C. Potts, DSPy: Compiling declarative language model calls into self-improving pipelines, in: International Conference on Learning Representations (ICLR), 2024. ArXiv:2310.03714.

[11] T. Schick, J. Dwivedi-Yu, R. Dessì, R. Raileanu, M. Lomeli, L. Zettlemoyer, N. Cancedda, T. Scialom, Toolformer: Language models can teach themselves to use tools, in: Advances in Neural Information Processing Systems, 2023. ArXiv:2302.04761.

[12] S. Geng, H. Cooper, M. Moskal, S. Jenkins, J. Berman, N. Ranchin, R. West, E. Horvitz, H. Nori, Generating structured outputs from language models: Benchmark and studies, arXiv:2501.10868, 2025. arXiv:2501.10868.

[13] R. Aghaei, A. A. Kiaei, M. Boush, J. Vahidi, M. Zavvar, Z. Barzegar, M. Rofoosheh, Harnessing the potential of large language models in modern marketing management: Applications, future directions, and strategic recommendations, arXiv:2501.10685, 2025. arXiv:2501.10685.

[14] A. C. Shanivendra, Hybrid intent-aware personalization with machine learning and RAG-enabled large language models for financial services marketing, arXiv:2603.14173, 2026. arXiv:2603.14173.

[15] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, C. Zhu, G-Eval: NLG evaluation using GPT-4 with better human alignment, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2023, pp. 2511–2522. doi:10.18653/v1/2023.emnlp-main.153, arXiv:2303.16634.

A. Dataset Corrections and Sensitivity Analysis

The experimental corpus underwent one correction and a canonical re-run during analysis.

Table 8
Audit 3: hypothesis-aware rater vs Opus ($n = 24$). Pearson r, mean absolute error, signed bias (Opus − human), and within-± 1 agreement.

Dimension	r	MAE	Bias	Within ± 1
correctness	$+0.53$	0.29	-0.12	23/24
typed_binding	-0.18	0.46	-0.29	22/24
reasoning_chain	$+0.33$	0.46	-0.29	24/24
hallucination	$+0.21$	1.83	$\mathbf{-1.83}$	9/24
composite (4-dim)	$+0.51$	0.68	-0.64	—

Correction: Lead D score and canonical q21–q23 re-run. q22's question text referenced a Sales Lead scored "above 0.75," but the corresponding lead in `sit_10` was at 0.71. We corrected the situation (Lead D score → 0.78) and re-ran q21–q23 from scratch under fully canonical situation data, question text, *and* judge ground-truth summaries. The pre-correction outputs are preserved in the run log under a superseded tag; all results in this paper use the canonical re-run.

What the canonical re-run changed. The correction has two effects. First, q22 becomes non-discriminating: once the premise is consistent (a lead at 0.78 does satisfy "above 0.75"), every condition reaches ceiling on correctness and typed-binding, so the item no longer separates conditions. Second, q23 — a hard chain-depth question — becomes the cleanest case of the GO framing block hurting: `go` is the worst condition (composite 3.11 vs roughly 4.0–4.2 for the others) because it over-applies the provenance lens at the expense of chain depth, and `prose_plus_json` is the canonical winner. Together these remove the discrimination that had propped up the composite: the headline contrast `go_no_preamble` − `prose` moves from $\Delta = +0.12$, $p = 0.010$ to $\Delta = +0.12$, $p = 0.060$, i.e. from significant to non-significant. We read the flip not as a defect but as a sharper result — the composite effect of representation is null under canonical scoring, and the real signal is task-specific — and as a concrete instance of how a two-question ground-truth misalignment can move a composite-level significance verdict in LLM-as-judge evaluation.

B. Human-Rater Audit Detail

Audit 1 (judge-rationale audit, $n = 20$). We sampled five judge rationales per condition from runs scored ≤ 3 on hallucination. For each item the judge flagged as a fabrication, we checked whether that string, number, or identifier was present in the situation JSON or the prose serialization the model actually received. Across the 20 rationales, every flagged item — 113 in total — was present in the source. The condition-level error pattern was uniform: 5/5 rationales in every condition contained at least one false-positive flag. The 113 figure indexes flagged items per rationale rather than independent observations: items within a rationale share a judge instance and a model output, so the relevant n for inference is 20 rationales, not 113 flagged items.

Audit 3 (hypothesis-aware human rater, $n = 24$). A KVA team member (non-author) scored a stratified random sample of 24 items drawn from `scored_outputs.jsonl` with fixed random seed (6 runs per condition across the five task types; q22 excluded because the human-rater sample was fixed before the canonical q21–q23 re-run). This exclusion applies only to the human-rater subsample; the deterministic identifier-membership check (Audit 2) covers the full canonical 252-output corpus. The rater used the same 1–5 rubric Opus used, with access to the situation files, blind to the condition and to Opus's scores, but not formally blind to the source-access hypothesis (Audit 4 addresses this). Per-dimension agreement statistics against Opus are in Table 8.

The three substantive dimensions show high within-± 1 agreement (92–100% of rows), with Opus mildly more conservative (bias -0.12 to -0.29). The hallucination dimension is the outlier: mean

signed bias $-\mathbf{1.83}$, Opus lower than the human on 21 of 24 rows, human lower than Opus on 0 of 24. The cleanest cases are rows where Opus scored 2 and the rater scored 5, in which the output cited source content (numbers, identifiers, quoted strings) verified present but unverifiable by a judge without the situation file.

Inter-rater agreement (first vs blind). Composite Spearman $\rho = +0.58$ ($p = 0.003$), MAE = 0.22; per-dimension Spearman ρ from $+0.22$ (hallucination) to $+0.67$ (reasoning_chain). Both raters used the top of the scale heavily, which mechanically compresses per-dimension correlations, so these should be read as a lower bound. The relevant cross-rater question for the audit is directional: both raters scored higher than Opus on hallucination on the great majority of items, neither scored lower than Opus on any item, and the disagreement is concentrated on the same dimension for both.

Ceiling-effect caveat. Both raters scored a single 24-row sample with concentrated use of the top of the scale (e.g. 22/24 fives on typed-binding and 21/24 fives on hallucination for the first rater). The hallucination finding is robust to this ceiling effect: the directional splits (21/3/0 and 22/2/0) with mean Δ of $+1.6$ to $+1.8$ cannot be explained by rater leniency, which would produce ties on the other dimensions too — and it does not. Multi-rater validation with calibrated raters at larger n is the priority methodological extension (§7).

C. Supplementary Materials

The full dataset (ten situations and their JSON encodings), the question battery and ground-truth summaries, the scoring script (`score_outputs.py`), the analysis script (`analyze_v1_2.py`), and the deterministic source-membership check (`audits/source_membership_check.py`) are available to reviewers on request and will be released publicly upon publication. Reported aggregates use the data at git tag `v1.2-results`. The FiQA arm uses the BeIR distribution of FiQA (HuggingFace dataset `BeIR/fiqa`); the exact 50-query sample (seed 42), serializers, generation and scoring harnesses, and scored outputs are part of the same release.