Ghost Context
Measuring Cross-Context Interference in Long-Context Language Models

Rohith Namboothiri
Independent Researcher
rohithneaasm05@alumni.iimk.ac.in

Abstract

Long-context language models assemble prompts from heterogeneous sources, and every deployed system trusts the model to use the *right* span. We show that this trust is routinely violated. Irrelevant spans silently shape outputs and produce errors that are neither fabrication nor omission but *misattributed grounding*, a claim supported by the wrong part of the input. Unlike intrinsic hallucination, which contradicts the source, or extrinsic hallucination, which introduces unsupported claims, misattributed grounding uses real evidence from the wrong source, and standard faithfulness metrics are structurally unable to detect it. We call this phenomenon *Ghost Context* and present an empirical study of its prevalence, structure, and remediation.

We evaluate three widely used models across a 272-case corpus using a mask-and-rerun causal attribution protocol with dual-judge calibration. We measure two complementary signals. Strict Ghost Context Rate (GCR) captures verifiable factual misattribution. Open-ended influence captures any detectable shaping, including non-factual shifts. Under realistic contextual conflict, strict GCR spikes. Temporal contradictions trigger misattributed grounding in **38.3%** of cases. Open-ended influence reaches **20.4%** across all cases. Crucially, Ghost Context is not only detectable but *correctable*. Masking the single highest-attributed distractor span resolves 95.5% of detected errors (Fix@1), with collateral damage of only 2.4% and zero false positives on negative controls. We release the full corpus, pipeline, and results.

1 Introduction

Production language model deployments assemble prompts from multiple sources. A retrieval-augmented generation (RAG) system (Lewis et al., 2020) appends retrieved document chunks to a user query. An agent framework preserves tool outputs and past conversation turns (Park et al., 2023a; Sumers et al., 2024). A multi-step reasoning pipeline concatenates intermediate results with original instructions. In each case, the model receives a context C composed of spans from disparate origins, and the system trusts the model to attend selectively to the spans that matter for the current query. Ghost Context undermines that trust by producing errors that appear well-supported yet rely on irrelevant spans, invisible to existing faithfulness metrics.

The breakdown happens under specific, realistic conditions. When a RAG system retrieves both the current remote work policy and a superseded 2022 draft, the model draws on the outdated draft 38.3% of the time on the most susceptible model we tested. The answer sounds authoritative. It is two years out of date. The error is invisible to any faithfulness metric that checks whether the claim is supported by *some* span in the context (Honovich et al., 2022; Min et al., 2023), because it is. It is supported by the wrong span.

We call this phenomenon *Ghost Context*, the causal influence of an irrelevant context span on model generation. Ghost Context differs from conventional hallucination (Ji et al., 2023) in a way that has been overlooked. A hallucinating model fabricates claims. A model exhibiting Ghost Context uses real evidence from the wrong source, a failure mode we term *misattributed grounding*.

Prior work has identified related phenomena. Liu et al. (2024b) document positional attention decay; Tang et al. (2024b) show distractor degradation in multi-hop QA; Shi et al. (2023) show reasoning failures from irrelevant context. These studies establish that irrelevant context matters, but none provide a unified formalization, causal detection protocol, and empirical measurement of this failure mode. We close that gap. We define Ghost Context and misattributed grounding as a failure mode distinct from fabrication and omission, introduce a causal mask-and-rerun protocol to measure and

Proceedings of the 6th Workshop on Trustworthy NLP (TrustNLP 2026), pages 316–329

attribute it, and show that under realistic conflict it is both frequent and largely correctable (Fix@1 = 95.5%).

The key finding of this paper is that Ghost Context is not uniformly distributed. It is rare in benign scenarios but spikes dramatically under realistic contextual conflicts. The overall rate across all categories and models is a modest 3.2%. Temporal contradictions push that figure to 38.3%, and open-ended distractor influence reaches 20.4%. These are not edge cases. They are the everyday conditions of RAG systems serving versioned documents, agentic systems accumulating memory over time, and multi-source contexts assembled by retrieval.

Equally important, Ghost Context is not intractable. Unlike many LLM failures, it is not only measurable but *causally localizable* to a single span and *largely correctable* with one masking intervention. Masking a single identified distractor span resolves 95.5% of errors with only 2.4% collateral damage (Table 2). This combination of high detectability, causal localizability, and cheap remediation is rare among LLM failure modes and points toward a practical mitigation path.

This paper makes five contributions. (1) An empirical measurement of Ghost Context under controlled conditions, revealing that contextual conflict produces failure rates an order of magnitude above baseline (§6). (2) A formal causal framework defining Ghost Context and its diagnostic metrics (GCR, Fix@k, CDR), plus Contextual Invariance Rate (CIR) as a system-level robustness objective, and a demonstration that source-blind faithfulness metrics are structurally unable to detect it (§2). (3) A taxonomy of four interference patterns (§3). (4) A mask-and-rerun protocol with dual-judge calibration, together with practical guidance for deployment when ground-truth relevance labels are not available (§4). (5) Analysis of Ghost Context as a security vulnerability in authorized contexts (§7).

2 Defining Ghost Context

2.1 Formal Framework

Consider a language model M that receives a prompt consisting of a query q and a context $C = \{c_1, c_2, \ldots, c_n\}$ of n spans. Each span c_i originates from a distinct source. The model generates an output $o = M(q, C)$.

Not all spans in C are relevant to q. Let $R(q, C) \subseteq C$ denote the set of spans that a correct answer should draw upon, and let $\bar{R}(q, C) = C \setminus R(q, C)$ denote the irrelevant spans.

Definition 1 (Ghost Context)**.** *An output $o = M(q, C)$ exhibits Ghost Context with respect to span $c_j \in \bar{R}(q, C)$ if there exists a claim $\phi \in$ Claims(o) such that (1) ϕ appears in $M(q, C)$, and (2) ϕ does not appear in $M(q, C^{(j \to mask)})$, where $C^{(j \to mask)}$ replaces span c_j with a neutral placeholder preserving approximate token count, and Claims(o) denotes the set of atomic factual assertions extracted from o. Two claims are equivalent if they assert the same factual proposition (entity, relation, value) regardless of surface phrasing. Stylistic variation, verbosity changes, and paraphrase do not constitute distinct claims.*

The placeholder is a fixed token pattern (repeated `[MASKED CONTENT]` tokens), chosen to avoid introducing new semantic content while preserving positional context for surrounding spans. This operationalization treats masking as an intervention that approximates causal influence (Pearl, 2009) under fixed decoding settings (temperature zero, greedy decoding).

We note here that the masking operation is methodologically a span-level counterfactual. This sits in the same family as leave-one-out attribution (Koh and Liang, 2017) but is not the contribution of this paper. The contribution is the conceptual category (misattributed grounding), the metrics that follow from it (Definitions 2–4), and the structural argument that source-blind faithfulness cannot detect this category (Proposition 2). Span counterfactuals are the verification mechanism, not the definition of the phenomenon.

Definition 2 (Ghost Context Rate)**.** *For a dataset $\mathcal{D} = \{(q_i, C_i)\}_{i=1}^{N}$ and model M, let $\bar{R}_i = \bar{R}(q_i, C_i)$.*

$$GCR(M, \mathcal{D}) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\big[\exists\, c_j \in \bar{R}_i : GC(M, q_i, C_i, c_j)\big] \tag{1}$$

Definition 3 (Fix@k and Collateral Damage Rate)**.** *Fix@k is the fraction of Ghost Context cases resolved by masking the top-k attributed distractor spans. The Collateral Damage Rate (CDR) measures how often masking degrades the answer to the original query.*

These three metrics together characterize prevalence, remediability, and cost. A model with high GCR but also high Fix@1 and low CDR is one

where Ghost Context is frequent yet cheaply correctable, a property we find empirically across all tested models.

Definition 4 (Contextual Invariance and CIR)**.** *A model M is* context-invariant *for query q and context C if the set of query-relevant factual claims in $M(q, C)$ is equivalent to that in $M(q, R(q, C))$ under deterministic decoding and a task-specific equivalence function $\equiv_q$. The* Contextual Invariance Rate *(CIR) measures the fraction of cases satisfying this property.*

$$\mathit{CIR}(M, \mathcal{D}) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\big[M(q_i, C_i) \equiv_{q_i} M(q_i, R(q_i, C_i))\big]. \tag{2}$$

In our experiments, $R(q, C)$ is instantiated using the corpus relevance labels (relevant vs. distractor vs. filler), which makes CIR directly measurable. We operationalize $\equiv_q$ via the same deterministic claim extraction and judge rubric used in strict detection, restricted to query-target facts. Under labeled relevance, CIR complements GCR. GCR measures distractor-attributed failure frequency, while CIR measures invariance to irrelevant context at the system level. We recommend reporting both, GCR for diagnostic granularity and CIR for system-level robustness assessment. In our setting, CIR is empirically correlated with conflict-conditional GCR under the same strict equivalence and corpus labeling assumptions, but the two are not complements in general.

2.2 Formal Properties

Proposition 1 (Masking Monotonicity, Diagnostic Expectation)**.** *If masking span c_j eliminates a span-attributed claim ϕ from the output, then masking any superset $S \supseteq \{c_j\}$ removes the span-based evidence for ϕ. Any reappearance of ϕ must originate from other context spans or from parametric knowledge, not from c_j. Monotonicity thus holds for span-attributed claims but not for claims recoverable from parametric memory, which makes it a useful diagnostic expectation rather than an absolute guarantee.*

Proposition 2 (Incompleteness of Source-Blind Faithfulness)**.** *Let $F : \mathcal{O} \times 2^{\mathcal{C}} \to \{0, 1\}$ be a faithfulness metric returning 1 iff every claim in o is entailed by at least one span in C. Then F cannot distinguish Ghost Context from correct attribution. For any output exhibiting Ghost Context, $F(o, C) = 1$, because the ghost claim is entailed by $c_j \in C$.*

This structural limitation follows directly from the standard definition of source-blind faithfulness metrics (proof in Appendix G). Metrics such as FActScore (Min et al., 2023), TRUE (Honovich et al., 2022), SummaC (Laban et al., 2022), and MiniCheck (Tang et al., 2024a) check whether claims are supported by *some* span but not whether they are supported by the *correct* span, and are therefore unable to detect Ghost Context.

We propose *misattributed grounding* as a complementary category to intrinsic and extrinsic hallucination (Maynez et al., 2020). Intrinsic hallucination contradicts the source. Extrinsic hallucination introduces unsupported claims. Misattributed grounding introduces claims that *are* supported, by the wrong source. The evidence is real and present in the context, so it is not extrinsic. The model does not contradict its source, so it is not intrinsic. Existing taxonomies miss the category entirely because they treat the context as an undifferentiated pool (see Appendix F for a detailed comparison).

3 Interference Taxonomy

Ghost Context arises through four interference patterns, each corresponding to a different mechanism by which an irrelevant span captures model attention (Vaswani et al., 2017). These patterns are not mutually exclusive.

Temporal Bleed. The model draws on an outdated span when a newer span contains the correct information. This occurs when contexts contain multiple versions of a fact, as in conversational and agent memory settings (Park et al., 2023a). Our experiments show this is the most potent interference pattern, with GCR reaching 38.3%.

Cross-Document Fusion. The model synthesizes information from unrelated spans, producing composite claims that appear in neither source. This pattern is amplified by semantic retrieval systems (Gao et al., 2024) that surface chunks by topic similarity without regard for shared provenance.

Distractor Dominance. A span that is semantically similar to the query but factually irrelevant overrides a less similar but more relevant span. Surface-level token overlap or richer detail in the distractor biases attention. This interacts with the "lost in the middle" effect (Liu et al., 2024b).

Instruction-Data Interference. An instruction span (system prompt, safety directive) fails to sup-

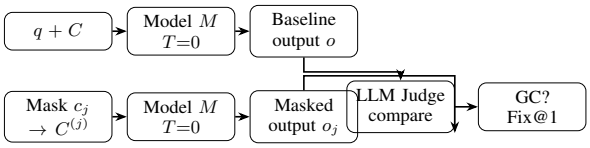

Figure 1: Mask-and-rerun detection pipeline. For each distractor span c_j, the protocol generates a masked variant and an LLM judge compares outputs to identify ghost claims.

press the causal influence of a data span. The model may acknowledge the instruction while generating content shaped by the data span, a behavior related to negation blindness (Nadeem et al., 2024; Kim et al., 2025; Truong et al., 2025) and instruction-following failures (Jang et al., 2023).

4 Mask-and-Rerun Detection

The causal definition in Definition 1 suggests a natural detection methodology. We systematically mask each distractor span and observe whether suspect claims persist. The approach is analogous to counterfactual explanation methods (Molnar, 2022), causal mediation analysis (Vig et al., 2020), and leave-one-out influence estimation (Koh and Liang, 2017), applied at the span level. Figure 1 illustrates the protocol.

Step 1, baseline generation. Generate $o = M(q, C)$ at temperature $= 0$.

Step 2, claim extraction. Decompose o into atomic claims using an LLM extractor.

Step 3, span masking. For each distractor span $c_j \in \bar{R}(q, C)$, generate $o_j = M(q, C^{(j \to \text{mask})})$. The mask replaces each distractor with repeated `[MASKED CONTENT]` tokens at approximately half the original word count, preserving positional context.

Step 4, differential attribution. An LLM judge (Zheng et al., 2024) compares o with each o_j, identifies ghost claims that appeared in the baseline but vanished after masking, and assesses collateral damage.

Step 5, aggregation. A case exhibits Ghost Context if any masking intervention reveals causal attribution to a distractor span.

For a context with d distractor spans, the protocol requires $d + 1$ inference calls. In our experiments, the average case contains 2.5 distractor spans, which makes single-span attribution tractable.

4.1 Operationalization Without Relevance Labels

In our evaluation corpus, relevance labels are known by construction, so $R(q, C)$ and $\bar{R}(q, C)$ are explicit. A natural question, raised in review, is how the protocol applies when those labels are not available, since most production deployments lack hand-annotated distractor labels.

We see three deployment patterns that preserve the spirit of the protocol without requiring labels.

Retrieval-score gating. Most RAG systems already produce retrieval scores or rerank scores for each chunk. A practical surrogate for $\bar{R}$ is the bottom-k chunks under that score, or chunks below a confidence threshold. The mask-and-rerun protocol then runs only on candidates the retriever itself was least sure about. This trades a perfect relevance partition for a cheap and deployable proxy.

All-span counterfactual sweep. When chunk counts are small (typical of conversational agents and short RAG contexts), the protocol can mask each chunk in turn and look for chunks whose removal materially changes the output. A span whose removal eliminates a claim but does not degrade the answer to the original query is, by Definition 1, behaving like a distractor under the current query, independent of any prior relevance label. This converts the protocol from a labeled diagnostic into an unsupervised attribution tool.

Audit mode rather than runtime mode. The protocol is not meant for runtime use on every query. We expect deployment as a periodic audit applied to logged query-context pairs, sampled from production traffic, with the goal of estimating GCR for a given pipeline and locating systematic interference patterns (specific source pairs, specific time-gap profiles, specific document templates). In audit mode, the absence of R is less constraining, because identification rather than online correction is the objective.

We view these patterns as starting points rather than completed methods. A full study of unsupervised mask-and-rerun, including its sensitivity to retriever quality and its calibration against labeled GCR, is left to future work.

5 Experimental Design

5.1 Evaluation Corpus

We construct a 272-case evaluation corpus spanning five categories (Table 1).

Table 1: Corpus composition. Adapted cases use MuSiQue (Trivedi et al., 2022), HotpotQA (Yang et al., 2018), and SituatedQA (Zhang and Choi, 2021).

Category	Cases	Adapted	Custom
Cross-document fusion	65	45	20
Distractor dominance	67	49	18
Temporal bleed	60	35	25
Instruction-data interf.	60	0	60
Negative control	20	13	7
Total	**272**	**149**	**123**

Adapted cases (149) preserve original queries and gold answers from established multi-hop QA benchmarks while restructuring documents into labeled multi-span contexts with annotated ghost claims.

Custom cases (123) provide controlled interference with precisely specified gold answers, distractor answers, and 3–5 ghost claims per case. Custom cases are essential for instruction-data interference, for which no benchmark exists, and for enterprise-domain scenarios (HR policies, medical protocols, financial guidance) underrepresented in academic datasets.

Negative controls (20) contain only relevant and filler spans with no distractor content. Each case averages 4.5 spans (1.3 relevant, 2.5 distractor, 0.7 filler), totaling approximately 479 words.

All cases underwent automated structural validation, LLM-assisted semantic validation, and cross-contamination checking (Appendix A).

5.2 Models Under Test

We evaluate three widely used models. **Claude Haiku 4.5** (Anthropic, 200K context). **DeepSeek V3.2** (DeepSeek, 128K context). **Llama 3.3 70B Instruct** (Meta, 128K context, open-weight). All accessed via AWS Bedrock at temperature $= 0$.

5.3 Judge Model and Dual-Judge Calibration

Claude Sonnet 4.6 serves as primary LLM judge, used *exclusively* as evaluator and never as test subject, which prevents direct circularity. The reliability of LLM judges for structured evaluation has been validated in prior work (Zheng et al., 2024; Chiang and yi Lee, 2023). DeepSeek V3.2 serves as secondary judge on a 30% stratified sample of detection judgments ($n = 206$ judgments per model, drawn from the 688 total judgments per model).

Same-family judge. We note explicitly that the primary judge and one of the test models, Claude Haiku 4.5, belong to the same model family. This raises a legitimate concern about family-aligned scoring, even though the primary judge is never asked to evaluate its own output. Three pieces of evidence speak to it. First, the dual-judge agreement rate between Claude Sonnet 4.6 and DeepSeek V3.2 is consistent across all three test models (97.1%, 98.1%, 99.0%), with no spike in disagreement specifically on Haiku judgments. Second, the strict GCR ranking the protocol produces (Haiku $>$ DeepSeek $>$ Llama) is not the ranking a family-aligned judge would be expected to favor for its own family. Third, the protocol detects misattribution by comparing two outputs from the same test model, baseline and masked, rather than scoring an output in isolation, so any uniform family-aligned bias would cancel between the two outputs. We treat this as mitigation, not elimination, and flag it again in Limitations.

5.4 Experimental Conditions

Over 8,000 inference calls across four conditions.

Main detection. 688 judgments per model (272 cases $\times$ $\sim$2.5 distractor spans per case) $\times$ 3 models = 2,064 detection judgments.

Positional variants. 3 positional variants for 67 distractor-dominance cases plus 1 shuffled variant for 252 non-control cases, across 3 models = 2,880 variant judgments.

Open-ended detection (Suite E). Independent rubric-based protocol without predefined ghost claims = 2,064 evaluations.

Context scaling. 15 cases $\times$ 3 models $\times$ 3 lengths (2K, 8K, 32K tokens) = 135 runs. We report this as a sanity check rather than as a primary result. At $n = 15$ per condition it is underpowered for strong scaling conclusions and we treat it as such throughout.

6 Results

Table 2 summarizes the key findings.

6.1 Ghost Context under Contextual Conflict

The overall strict GCR of 3.2% is a mixture. It averages benign scenarios (near 0%) with high-conflict scenarios (up to 38.3%). The policy-relevant number is the conflict-conditional rate, because production systems routinely encounter temporal contradictions and overlapping retrievals. The central

Table 2: Key findings at a glance.

Metric	Value
Peak GCR (temporal bleed, Haiku)	38.3%
Open-ended influence (all models)	20.4%
Overall strict GCR	3.2%
CIR (overall)	94.1%
Fix@1 (single-span remediation)	95.5%
Negative control false positive rate	0.0%

Table 3: Ghost Context under contextual conflict. Temporal contradictions and open-ended distractor influence represent the most realistic and highest-impact failure scenarios.

Scenario	Haiku 4.5	DeepSeek	Llama 70B
Strict detection (binary GCR)			
Temporal bleed	**38.3%**	13.3%	3.3%
Open-ended influence (any detectable change)			
Any influence	22.2%	20.5%	18.6%
Factual contam.	9.9%	14.8%	6.7%
Framing shift	4.8%	2.6%	5.2%

finding is that Ghost Context concentrates in specific, realistic contextual conflict scenarios. Table 3 presents GCR for the two highest-interference categories.

Temporal contradictions are the most dangerous. When the context contains both an outdated and a current version of a fact, Claude Haiku 4.5 draws on the outdated version in 38.3% of cases (95% Wilson CI: 26.5%, 51.4%), more than one in three queries. For RAG systems serving versioned documents (Gao et al., 2024), this suggests that temporal conflicts can frequently trigger contamination on susceptible models. Even the most resilient model (Llama 3.3 70B) exhibits 3.3%, a non-negligible rate at scale.

An illustrative case. The query "*What is the current population of the United Kingdom? Answer with a single current estimate.*" explicitly requires one value. The context contains a 2023 source (gold, 67.7 million) and a 2019 distractor (66.4 million). The baseline output reports *both* figures, "approximately 67.7 million people (as of 2023)" and "approximately 66.4 million people (as of 2019)." The task requires a single value, which makes the inclusion of the outdated figure incorrect under the task specification. This is misattributed grounding. The model drew on the wrong span to produce a claim that violates the query constraint. After masking the 2019 distractor, the model reports only the correct 2023 figure. The ghost claim is verifiably supported by a span in the context, which makes it invisible to source-blind faithfulness metrics while being causally attributable to the wrong span.

Table 4: GCR for baseline interference categories.

Category	Haiku 4.5	DeepSeek	Llama 70B
Distractor dom.	5.2%	3.4%	1.7%
Cross-doc fusion	2.4%	1.9%	1.3%
Instruction-data	0.0%	1.7%	0.0%
Negative control	0.0%	0.0%	0.0%

Open-ended influence is a companion metric. Strict GCR measures factual misattributed grounding, verifiable claims causally attributed to distractor spans. Open-ended influence measures non-factual but detectable shaping, framing shifts, emphasis changes, and hedging. It serves as a companion metric characterizing the broader interference spectrum. The gap between strict GCR (3.2%) and open-ended influence (20.4%) represents cases where distractor spans do not introduce verifiable factual errors but shift the output's framing, emphasis, or hedging. Temporal bleed shows the highest open-ended influence (63–70% across models; see Appendix C). Instruction-data cases reveal a notable divergence. Strict GCR is near-zero, but open-ended influence spans 5–43%, which suggests that explicit instructions suppress factual contamination but not subtler forms of contextual shaping.

Open-ended influence rates must be interpreted relative to each model's negative-control baseline.[1]

6.2 Baseline Interference Rates

Table 4 presents GCR for the remaining interference categories.

Distractor dominance and cross-document fusion produce lower but consistent GCR across models (1.3–5.2%). Instruction-data interference is rare under strict detection (1 instance across 180 judgments), though open-ended analysis reveals broader influence. Critically, **negative controls achieve 0% GCR across all 228 control detections**, establishing zero false positives when no distractor content is present.

[1]DeepSeek shows 18.4% influence on negative controls due to parametric knowledge leakage (the model adding biographical details from training data), compared to 6.6% for Haiku and 9.2% for Llama. Net influence above baseline: Haiku 15.6%, Llama 9.4%, DeepSeek 2.1%. Full breakdown in Appendix C.

Table 5: Remediation metrics. Fix@1 = resolved by masking top-1 span; CDR = collateral damage rate. Confidence = fraction of detections rated "high" by the LLM judge.

Model	GCR	Fix@1	CDR	Confid.
Haiku 4.5	5.5%	94.7%	1.3%	98.4%
DeepSeek V3.2	2.9%	95.0%	2.8%	99.6%
Llama 3.3 70B	1.3%	100.0%	3.2%	98.4%
Overall	**3.2%**	**95.5%**	**2.4%**	**98.8%**

6.3 Remediation, Detection, Attribution, and Correction

A failure mode that is frequent but intractable is an interesting finding. A failure mode that is frequent, precisely localizable, and cheaply correctable is an *actionable* one. Table 5 presents the remediation metrics.

Three properties make Ghost Context unusually tractable among LLM failure modes.

Detectable. 98.8% of detections were rated "high confidence" by the LLM judge. No detected Ghost Context cases received a "low" confidence rating, which means Ghost Context produces output differences that are unambiguous to the judge.

Causally attributable. Each Ghost Context error is traceable to a specific distractor span. Unlike parametric hallucination, where the offending input is distributed across the training corpus, Ghost Context errors can be localized to a single span in the runtime context.

Correctable. Masking the single highest-attributed distractor resolves **95.5%** of all detected errors. Collateral damage from masking is only 2.4%, which means removing the offending span rarely degrades the correct answer. This suggests a practical mitigation path. A post-generation verification step that masks low-relevance spans and compares outputs could catch and correct most Ghost Context errors before they reach users.

6.4 Dual-Judge Calibration

Table 6 presents inter-judge reliability.

Raw agreement exceeds 97% across all models. The $\kappa = 0.0$ for Llama is an instance of the well-documented prevalence paradox (Feinstein and Cicchetti, 1990). With only 1.3% GCR, the 206-judgment sample contains so few positive instances that expected chance agreement is already 99%, which makes κ uninformative. For Claude Haiku, where the positive class is better represented, $\kappa = 0.652$ indicates substantial agreement (Landis and Koch, 1977).

Table 6: Dual-judge agreement. Primary judge is Claude Sonnet 4.6; secondary judge is DeepSeek V3.2. 30% stratified sample of detection judgments ($n = 206$ per model). †Uninformative under low prevalence (prevalence paradox).

Test Model	Agreement	κ	n
Claude Haiku 4.5	97.1%	0.652	206
DeepSeek V3.2	98.1%	0.325	206
Llama 3.3 70B	99.0%	0.000†	206

Table 7: GCR by positional variant. Positional, $n = 116$ per model per position. Shuffled, $n = 612$ per model.

Model	Begin	Middle	End	Shuffled
Haiku 4.5	5.2%	1.7%	4.3%	4.4%
DeepSeek V3.2	3.4%	3.4%	3.4%	2.9%
Llama 70B	1.7%	2.6%	3.4%	2.8%

6.5 Positional Robustness

Table 7 tests whether Ghost Context depends on span position.

Position has bounded influence. Haiku shows a 3× range (1.7–5.2%) while DeepSeek is uniform, consistent with prior findings on position sensitivity (Liu et al., 2024b; Levy et al., 2024; Hsieh et al., 2024). Shuffled variant GCR (2.8–4.4%) is comparable to positional averages, which confirms that span ordering has limited impact. That is an important result for RAG systems where chunk order is arbitrary (Barnett et al., 2024). The preliminary context scaling check (Appendix D, $n = 15$ per condition at 2K, 8K, and 32K tokens) finds no significant GCR variation. It does not establish absence of length effects at production scale and is reported as a sanity check rather than as a finding.

7 Security Implications

Ghost Context exposes three security surfaces that existing defenses do not address. These are framed as hypotheses derived from our causal results rather than as evaluated attacks, and we say so explicitly.

Authorized context leakage. Even when every span is individually authorized, cross-context interference can produce outputs disclosing information no single document intended to convey. An employee viewing both a headcount report and individual project assignments could receive an LLM-generated salary estimate that neither document contains. Authorization-first retrieval frameworks

prevent unauthorized spans from entering context, but Ghost Context arises from *authorized* spans interacting in unintended ways.

Agent memory contamination. In multi-user agent deployments (Park et al., 2023a; Sumers et al., 2024), a memory entry from one user's session can shape responses to another user, which creates a cross-user information flow channel that no per-session access control prevents.

Manipulation surface. An adversary who can inject a span into a RAG corpus can exploit Ghost Context to subtly bias generation. Unlike prompt injection (Greshake et al., 2023; Perez and Ribeiro, 2022; Zhan et al., 2024), which aims to override instructions, Ghost Context manipulation shifts the output without visible traces. The injected span contains no adversarial instructions. It contains information that, blended with legitimate spans, biases the output in the attacker's desired direction (Yi et al., 2023).

We do not empirically evaluate these scenarios in this paper, and we list this as a limitation in the next section. The causal mechanism we demonstrate, irrelevant spans shaping outputs, is what makes the surfaces plausible, and the manipulation surface in particular is increasingly timely as RAG and agent systems assemble contexts from partially trusted sources.

We identify the need for *Generative Contextual Fidelity* (GCF), a guarantee that irrelevant spans do not causally influence generation, analogous to Nissenbaum's Contextual Integrity (Nissenbaum, 2004) applied to generation. Formally, a system satisfies GCF if for all queries q and contexts C, the output $M(q, C)$ is functionally equivalent to $M(q, R(q, C))$, which means irrelevant spans have zero causal influence. We treat this as an idealized guarantee; stochastic or long-context models may not achieve strict equivalence. Practical systems can target *approximate GCF*, outputs equivalent up to factual claims about the query target, under deterministic decoding and task-specific equivalence. This is operationalized by the Contextual Invariance Rate (Definition 4). Targeting high CIR on conflict-stress suites is a concrete way to pursue approximate GCF as an audit criterion or design goal. Our Fix@1 results suggest that approximate GCF is achievable in practice through post-hoc verification, even though it is not guaranteed by current model architectures.

8 Related Work

Hallucination. Surveys by Ji et al. (2023), Huang et al. (2023), and Zhang et al. (2023) organize hallucination into intrinsic and extrinsic categories but do not address misattributed grounding. FActScore (Min et al., 2023) and MiniCheck (Tang et al., 2024a) verify claims against context but treat it as an undifferentiated pool. SelfCheckGPT (Manakul et al., 2023) detects hallucination through sampling consistency but cannot localize the causal span. Our Proposition 2 demonstrates that all such source-blind metrics are structurally unable to detect Ghost Context.

Long-context evaluation. Liu et al. (2024b) demonstrate positional attention decay; Shi et al. (2023) show distractor degradation of reasoning. RULER (Hsieh et al., 2024) and NeedleBench (Li et al., 2024) benchmark retrieval within long contexts. Our positional results (Table 7) show that positional effects on Ghost Context are measurable but bounded, and that distractor content matters more than position.

RAG. Lewis et al. (2020) introduce RAG; Self-RAG (Asai et al., 2024) adds self-reflection; RAGAS (Es et al., 2024) automates evaluation. Barnett et al. (2024) identify seven RAG failure points; Gao et al. (2024) survey the landscape. None operationalize the causal link between retrieval errors and generative contamination; our protocol does.

Attribution and ablation. Influence functions (Koh and Liang, 2017), TRAK (Park et al., 2023b), and AttriBoT (Liu et al., 2024a) trace predictions to training examples or context elements. Attributable QA (Bohnet et al., 2022) and Gao et al. (2023) require models to cite sources. These treat attribution as a generation task. We provide causal verification at the span level. We acknowledge that the mask-and-rerun mechanism shares its mechanical structure with leave-one-out ablation; the contribution here is the framing (misattributed grounding as a distinct failure category), the formal incompleteness result for source-blind faithfulness, the metric suite (GCR, Fix@k, CDR, CIR), and the empirical demonstration that this framing yields actionable remediation in deployed model settings.

Prompt injection. Indirect injection (Greshake et al., 2023), instruction-override attacks (Perez and Ribeiro, 2022), and tool-integrated vulnerabilities (Zhan et al., 2024) exploit instruction-data confusion. Ghost Context manipulation differs. The injected content need not contain instructions.

Attention and multi-document QA. Attention weights do not capture span-level causal attribution (Jain and Wallace, 2019; Wiegreffe and Pinter, 2019; Vig et al., 2020). MuSiQue (Trivedi et al., 2022), HotpotQA (Yang et al., 2018), and SituatedQA (Zhang and Choi, 2021) provide datasets we adapt, adding distractor and ghost claim annotations the originals lack.

9 Conclusion

Ghost Context is not a uniform failure. It spikes under realistic contextual conflict. Temporal contradictions trigger misattributed grounding in 38.3% of cases on the most susceptible model, open-ended influence reaches 20.4%, and the overall strict rate is 3.2%. Crucially, it is detectable (98.8% high-confidence), causally localizable to specific spans, and correctable. Masking a single distractor resolves 95.5% of errors with 2.4% collateral damage. These results challenge the assumption that more context is uniformly better. RAG and agent builders can use the mask-and-rerun protocol to audit deployed systems. Evaluation teams can report GCR and Contextual Invariance Rate as standard robustness metrics. Future directions include attention-weight analysis, multi-span joint masking, production-scale context lengths (50K–200K tokens), training-time mitigations such as context-source tagging, and human-validated calibration of the detection judge.

Limitations

Corpus scale and length. The 272-case corpus establishes prevalence and identifies patterns but limits fine-grained statistical comparisons at low prevalence. Category-level comparisons, particularly for instruction-data interference (1 detected instance under strict detection), require larger corpora. Primary evaluation also uses short contexts of roughly 479 words. Production RAG systems assemble 5K–50K token contexts. Our scaling check at $n = 15$ per condition is a sanity check rather than a production-scale result, and we treat it as such. A scaled-up evaluation in the 5K–50K token range remains the most important follow-up.

Judge-based detection and human validation. Detection relies on LLM judges. Dual-judge calibration shows high raw agreement (97–99%), but κ values are difficult to interpret at low prevalence because of the prevalence paradox (Feinstein and Cicchetti, 1990). We did not run a human annotation study on a subset of judgments. That is the cleanest way to validate the LLM-as-judge setup we use, and we identify it as the highest-priority next step for strengthening these claims.

Same-family judge. The primary judge (Claude Sonnet 4.6) and one test model (Claude Haiku 4.5) share a model family. We mitigate this through dual-judge calibration with an unrelated judge and through differential rather than absolute scoring, and the agreement rates and rankings reported in Section 5 do not indicate family-aligned bias. Mitigation is not elimination, however, and cross-family judge replication is a useful follow-up.

Single-span attribution. The protocol tests single-span masking. Cross-document fusion can involve jointly causal spans, and single-span masking may underestimate GCR for this category. A multi-span joint-masking variant is straightforward in principle but quadratic in inference cost and is left to future work.

Operationalization without relevance labels. Our headline experiments use labeled relevance. Section 4.1 sketches three deployment patterns that do not require labels, but none is empirically validated in this paper. A calibrated unsupervised version of the protocol is an open problem.

Security claims are unevaluated. The three attack surfaces discussed in Section 7 are derived from our causal results. They are not separately tested. Empirical evaluation of authorized context leakage, agent memory contamination, and manipulation attacks is the natural next step for the security implications section.

Model coverage. We evaluate three models from three families. Frontier models with larger parameter counts may produce different interference profiles.

Ethical Considerations and Broader Impact

This work measures a failure mode that could be exploited for adversarial manipulation (Section 7). We disclose the Ghost Context manipulation surface to enable defensive research; the mask-and-rerun protocol is itself a defensive tool.

Data and AI use. Our corpus contains no sensitive personal data; all custom cases use fictional entities. Adapted cases are drawn from publicly available Wikipedia content via MuSiQue, HotpotQA, and SituatedQA. Corpus construction employed LLM-assisted generation for custom cases and an-

notation enrichment; all generated content was validated through three independent layers (Section 5, Appendix A). The evaluation pipeline uses an LLM judge as an integral methodological component, consistent with established LLM-as-judge evaluation practices (Zheng et al., 2024).

Broader impact. Ghost Context detection can improve the trustworthiness of deployed RAG systems, agent frameworks, and multi-source LLM applications by identifying a previously invisible failure mode. While the taxonomy could inform adversarial strategies, the detection and remediation capabilities we provide substantially outweigh that risk. Our protocol enables defenders to identify and correct Ghost Context errors before they reach users.

Acknowledgments

We thank the anonymous reviewers at TrustNLP for constructive feedback that shaped the camera-ready version of this paper, in particular the suggestion to make the unsupervised operationalization of the protocol explicit and the prompt to clarify the relationship between the mask-and-rerun mechanism and existing ablation methods.

Artifact Availability

The full corpus (272 cases), detection pipeline, and results are available at `https://anonymous.4open.science/r/GCR-Artifact-C62A/`. The anonymized link will be replaced with a permanent repository in the final version.

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *ICLR*.

Scott Barnett, Stefanus Kurniawan, Srikanth Thudumu, Zach Brannelly, and Mohamed Abdelrazek. 2024. Seven failure points when engineering a retrieval augmented generation system. *arXiv:2401.05856*.

Bernd Bohnet, Vinh Q. Tran, Pat Verga, Roee Aharoni, Daniel Andor, Livio Baldini Soares, Jacob Eisenstein, Kuzman Ganchev, Jonathan Herzig, Kai Hui, Tom Kwiatkowski, Ji Ma, Jianmo Ni, Liber Osber, Tal Schuster, William W. Cohen, Michael Collins, Dipanjan Das, Donald Metzler, and 2 others. 2022. Attributed question answering: Evaluation and modeling for attributed large language models. In *arXiv:2212.08037*.

Cheng-Han Chiang and Hung yi Lee. 2023. Can large language models be an alternative to human evaluations? *arXiv:2305.01937*.

Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. 2024. RAGAS: Automated evaluation of retrieval augmented generation. In *EACL System Demonstrations*, pages 150–163.

Alvan R. Feinstein and Domenic V. Cicchetti. 1990. High agreement but low kappa: I. the problems of two paradoxes. *Journal of Clinical Epidemiology*, 43(6):543–549.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations. In *EMNLP*, pages 6465–6488.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey. *arXiv:2312.10997*.

Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. 2023. Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection. In *AISec*.

Or Honovich, Leshem Choshen, Roee Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. 2022. TRUE: Re-evaluating factual consistency evaluation. In *NAACL*, pages 3905–3920.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. 2024. RULER: What's the real context size of your long-context language models? *arXiv:2404.06654*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv:2311.05232*.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not explanation. In *NAACL*, pages 3543–3556.

Joel Jang, Seungone Kim, Seonghyeon Ye, Doyoung Kim, Lajanugen Logeswaran, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2023. Can large language models truly follow your instructions? In *EMNLP Findings*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Jieun Kim and 1 others. 2025. Semantic inversion, identical replies: Revisiting negation blindness in large language models. In *EMNLP*.

Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. In *ICML*, pages 1885–1894.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. In *TACL*, volume 10, pages 163–177.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Mosh Levy, Alon Jacoby, and Yoav Goldberg. 2024. Same task, more tokens: the impact of input length on the reasoning performance of large language models. *arXiv:2402.14848*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *NeurIPS*, volume 33, pages 9459–9474.

Mo Li, Songyang Zhang, Yunxin Liu, and Kai Chen. 2024. NeedleBench: Can LLMs do retrieval and reasoning in 1 million context window? *arXiv:2407.11963*.

Feiyang Liu, Nikhil Kandpal, and Colin Raffel. 2024a. AttriBoT: A bag of tricks for efficiently approximating leave-one-out context attribution. *arXiv:2411.15102*.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024b. Lost in the middle: How language models use long contexts. *TACL*, 12:157–173.

Potsawee Manakul, Adian Liusie, and Mark J.F. Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. *arXiv:2303.08896*.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *ACL*, pages 1906–1919.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *EMNLP*, pages 12076–12100.

Christoph Molnar. 2022. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*, 2nd edition.

Mohsan Nadeem and 1 others. 2024. Negation blindness in large language models: Unveiling the NO syndrome in image generation. *arXiv:2409.00105*.

Helen Nissenbaum. 2004. Privacy as contextual integrity. *Washington Law Review*, 79(1):119–158.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023a. Generative agents: Interactive simulacra of human behavior. In *UIST*.

Sung Min Park, Kristian Georgiev, Andrew Ilyas, Guillaume Leclerc, and Aleksander Madry. 2023b. TRAK: Attributing model behavior at scale. In *ICML*.

Judea Pearl. 2009. *Causality*, 2nd edition. Cambridge University Press.

Fabián Perez and Ian Ribeiro. 2022. Ignore previous prompt: Attack techniques for language models. In *NeurIPS ML Safety Workshop*.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *ICML*, pages 31210–31227.

Theodore R. Sumers, Shunyu Yao, Karthik Narasimhan, and Thomas L. Griffiths. 2024. Cognitive architectures for language agents. *arXiv:2309.02427*.

Liyan Tang, Philippe Laban, and Greg Durrett. 2024a. MiniCheck: Efficient fact-checking of LLMs on grounding documents. In *EMNLP*.

Yi Tang and 1 others. 2024b. MultiHop-RAG: Benchmarking retrieval-augmented generation for multi-hop queries. *arXiv:2401.15391*.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. MuSiQue: Multi-hop questions via single hop question composition. *TACL*, 10:539–554.

Van Truong and 1 others. 2025. A comprehensive taxonomy of negation for NLP and LLMs. *arXiv:2507.22337*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS*, pages 5998–6008.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nishi, Jason Ahn, Daniel Jurafsky, and Bernd Bohnet. 2020. Investigating gender bias in language models using causal mediation analysis. In *NeurIPS*.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. In *EMNLP*, pages 11–16.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *EMNLP*, pages 2369–2380.

Jingwei Yi, Yueqi Xie, Bin Zhu, Keegan Hines, Emre Kiciman, Guangzhong Sun, Xing Xie, and Fangzhao Wu. 2023. Benchmarking and defending against indirect prompt injection attacks on large language models. *arXiv:2312.14197*.

Qiusi Zhan, Zhixiang Liang, Zifan Ying, and Daniel Kang. 2024. InjecAgent: Benchmarking indirect prompt injections in tool-integrated large language model agents. *arXiv:2403.02691*.

Michael J.Q. Zhang and Eunsol Choi. 2021. SituatedQA: Incorporating extra-linguistic contexts into QA. In *EMNLP*, pages 7371–7387.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023. Siren's song in the AI ocean: A survey on hallucination in large language models. *arXiv:2309.01219*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2024. Judging LLM-as-a-Judge with MT-Bench and chatbot arena. In *NeurIPS*.

A Corpus Construction Details

Each adapted case undergoes (1) category assignment based on question type, (2) span structuring with explicit relevant/distractor/filler labels, (3) ghost claim annotation (3–5 claims per case), and (4) enrichment with metadata including source, word counts, and system instructions.

Custom cases follow a template. Define a query with a clear gold answer, design a distractor span, add filler spans, annotate ghost claims, and for instruction-data cases add constraining system instructions.

All cases undergo three validation layers, automated structural checks, LLM-assisted semantic validation (answer isolation, distractor plausibility, ghost claim specificity), and cross-contamination checks.

Table 8 summarizes the corpus statistics.

Table 8: Corpus statistics by span type.

Span type	Count	Mean words	Median	Range
Relevant	357	75.7	76	13–267
Distractor	688	133.4	99	13–1378
Filler	183	62.5	62	20–235

Ghost claims per case have a mean of 4.3 and a range of 3–5. Source dataset breakdown is custom 123, MuSiQue 49, HotpotQA 45, SituatedQA 35, negative controls 20.

B Detection Prompt Templates

B.1 Strict Detection Prompt

The strict detection prompt provides the judge with the original query, baseline output, masked output, and known ghost claims. The full template is given below.

```
You are evaluating whether a language model exhibited "Ghost Context": drawing on an irrelevant document span to produce claims.

The model was asked: {query}

OUTPUT WITH FULL CONTEXT: {baseline_output}

OUTPUT WITH DISTRACTOR MASKED: {masked_output}

KNOWN GHOST CLAIMS: {ghost_claims}

Analyze the two outputs and determine: 1. Did the baseline output contain any of the ghost claims? 2. Did removing the distractor change or eliminate those claims? 3. Did removing the distractor degrade the answer (collateral damage)?

Return JSON: {ghost_claims_present_in_baseline, ghost_claims_absent_in_masked, ghost_context_detected, collateral_damage, confidence, reasoning}
```

B.2 Open-Ended Detection Prompt

The open-ended variant omits known ghost claims and asks the judge to independently identify distractor influence using a structured rubric.

```
You are comparing two outputs from the same language model answering the same question. The ONLY difference is that one distractor document was REMOVED.

Follow this STRICT 3-step evaluation:

STEP 1: Does Output A meaningfully differ from Output B? (Ignore trivial wording variations.)

STEP 2: Classify the PRIMARY difference: factual_contamination | framing_shift | emphasis_shift | hedging | tone_change | no_change.

STEP 3: Provide EXACT QUOTED EVIDENCE from both outputs.

Return JSON: {influence_detected, influence_type, influence_severity, factual_contamination, ghost_claims_found, evidence_from_output_a, evidence_from_output_b, reasoning}
```

Table 9: Open-ended influence rate by category and model. These rates include all detectable influence types (factual contamination, framing shifts, emphasis shifts, hedging).

Category	Haiku 4.5	DeepSeek	Llama 70B
Temporal bleed	63.3%	70.0%	63.3%
Instruction-data	43.3%	6.7%	5.0%
Distractor dom.	21.6%	17.2%	25.9%
Cross-doc fusion	15.7%	16.2%	13.3%
Negative control	6.6%	18.4%	9.2%

C Full Open-Ended Influence Results

Table 9 presents open-ended influence rates by category and model.

The instruction-data category reveals a striking divergence between strict GCR (near-zero) and open-ended influence (43.3% for Haiku). Explicit instructions suppress factual contamination effectively but do not prevent subtler forms of contextual shaping such as framing shifts and hedging.

DeepSeek's elevated negative control rate of 18.4% reflects parametric knowledge leakage rather than distractor influence. Manual inspection confirms the model adds biographical details from training data even when no distractor is present. Net influence above baseline is Haiku 15.6%, Llama 9.4%, DeepSeek 2.1%.

Table 10 breaks down the types of open-ended influence per model across all categories.

Table 10: Open-ended influence type breakdown across all categories.

Influence type	Haiku 4.5	DeepSeek	Llama 70B
No change	77.8%	79.5%	81.4%
Factual contam.	9.9%	14.8%	6.7%
Framing shift	4.8%	2.6%	5.2%
Emphasis shift	4.2%	1.0%	3.1%
Hedging	3.2%	2.0%	3.5%

D Context Scaling Results

Fifteen cases were padded with neutral filler spans to approximate target token counts of 2K, 8K, and 32K tokens and re-evaluated across all three models ($n = 15$ per condition per model). Results show no clear GCR trend across context lengths in this small sample, which suggests that for the interference patterns tested, Ghost Context is driven by distractor content quality rather than context length within this range. We emphasize the limitation of the sample size. At $n = 15$ per condition this analysis is a sanity check, not a result. It is consistent with the hypothesis that distractor content drives GCR, but it does not rule out length effects at production-scale contexts. Larger-scale experiments in the 5K–50K token range are needed to settle this question (Hsieh et al., 2024; Li et al., 2024).

E Qualitative Case Studies

E.1 Temporal Bleed, UK Population

Query. "What is the current population of the United Kingdom? Answer with a single current estimate."

Relevant span (2023 source). "As of 2023, the United Kingdom has a population of approximately 67.7 million people, making it the third most populous country in Europe…"

Distractor span (2019 source). "As of 2019, the United Kingdom maintains a population of approximately 66.4 million…"

Baseline output (Claude Haiku 4.5, with distractor present).

> "Approximately 67.7 million people (as of 2023), the most recent figure available. Approximately **66.4 million people (as of 2019)**."

Masked output (distractor removed).

> "As of 2023, the United Kingdom has a population of approximately 67.7 million people."

Analysis. The ghost claim "66.4 million as of 2019" is verifiably present in the baseline and absent in the masked output. The claim is supported by a span in the context, the 2019 source, which makes it invisible to source-blind faithfulness metrics. Masking the distractor resolves the error with no collateral damage.

E.2 Remediation Example, Fix@1

In the case above, Ghost Context was detected by the LLM judge with "high" confidence. The single distractor span was identified as the causal source. Masking it produced a clean output with only the correct 2023 figure, which illustrates Fix@1 remediation. Across all 67 detected Ghost Context instances, masking the top-1 attributed span resolved 95.5% of errors (64/67).

F Hallucination Taxonomy Comparison

Table 11 compares misattributed grounding with the established intrinsic/extrinsic hallucination taxonomy.

Table 11: Comparison of hallucination categories. Misattributed grounding (Ghost Context) is a complementary category that existing taxonomies do not capture.

Dimension	Intrinsic	Extrinsic	Misattributed grounding
Definition	Contradicts source	Unsupported claims	Claims from wrong source
Error source	Distorts provided info	Parametric memory	Irrelevant context span
Source-blind detection?	Yes (contradiction)	Yes (no support)	**No** (supported by a span)
Causal attribution	Difficult	Difficult	**Tractable** (mask-and-rerun)
Runtime fix?	Retrain / prompt	Retrieval fix	**Yes** (Fix@1 = 95.5%)

G Proof of Proposition 2

Proposition 3 (Incompleteness of Source-Blind Faithfulness, restated)**.** *Let $F : \mathcal{O} \times 2^{\mathcal{C}} \to \{0, 1\}$ be a faithfulness metric that returns 1 if and only if every claim in o is entailed by at least one span in C. Then for any output o exhibiting Ghost Context with respect to span c_j, $F(o, C) = 1$.*

Proof. Suppose $o = M(q, C)$ exhibits Ghost Context with respect to $c_j \in \bar{R}(q, C)$. By Definition 1, there exists a claim $\phi \in \text{Claims}(o)$ such that ϕ appears in o and ϕ does not appear in $M(q, C^{(j \to \text{mask})})$.

By definition of F, $F(o, C) = 1$ iff every claim in o is entailed by at least one span in C. The ghost claim ϕ is entailed by $c_j \in C$; hence ϕ satisfies the entailment condition. Thus $F(o, C) = 1$ whenever the ghost claim is present and entailed by some span in C, regardless of whether that span is relevant or irrelevant. F cannot detect misattribution because it does not distinguish *which* span in C supported each claim, only that *some* span did.

Hence, F is structurally unable to distinguish Ghost Context from correct attribution. □ □