Context Is Not Control

Source-Boundary Failures in Controlled Text-Mediated Evidence Use

R.J. Sabouhi · Symbolic Suite

Working Manuscript v0.6 — claim-tightening and replication-artifacts revision May 2026

> **Revision note (v0.6)**
>
> *This version tightens the central claim, restructures the frontier section to make the three closed objections explicit, adds a Data and Replication Artifacts section and a Claims-Supported/Not-Supported summary box, renames remaining ambiguous supported-row metrics, marks non-applicable final-token metrics as N/A, and improves wide-table formatting. It carries forward all v0.5 figures and tables without changing their underlying numbers.*

Core line

Context is not control. A model should not treat every token in the context window as equally capable of governing evidence use.

Abstract

Large language models often fail in retrieval, memory, policy, and documentation settings not only because relevant information is absent, but because the status of information that is present is not preserved. A context window may contain current evidence, stale evidence, answer choices, quoted distractors, embedded instructions, fake authority claims, and non-governing background text at the same time. The model must determine not merely what text says, but which source is admissible for answering the user question.

This paper introduces **source-boundary failure**: an observable failure to preserve the admissibility relation between a question, source role, source status, embedded control-like text, and candidate factual predicate. We evaluate this failure class in controlled synthetic memory, policy, and software-documentation tasks with supported rows, where governing evidence is present and should be used, and unsupported rows, where answer-like text may appear but no admissible governing source exists and the model should abstain.

These experiments show a source-boundary failure mode in which models treat contextually present but inadmissible text as answer-bearing evidence unless the task frame explicitly represents source admissibility. We do not claim this explains all hallucination, and we do not claim it generalizes to deployed RAG, agentic, or long-horizon tool-use systems without further testing.

Across open-weight models, raw source rendering produces substantial overrefusal, injection-paired supported-row failure, unsupported answering, and distractor capture. Sanitization often improves behavior. Explicit boundary records can further improve behavior, but are interpreted as boundary-conditioned answering tests rather than standalone proof of intrinsic boundary inference.

Across frontier models — GPT-5.1, Claude Sonnet 4.6, and Gemini 3.1 Pro Preview — the new API results close three specific objections. First, the original multiple-choice baseline produces 100% unsupported answering and 0.000 admissibility accuracy across all three models. Second, providing only an `INSUFFICIENT` channel does not recover behavior; channel-only and channel-plus-choices-not-evidence conditions remain at 0.000

admissibility accuracy. Third, a metadata-only condition that does not hand the model the no-source conclusion captures 84%–100% of the full-boundary lift. A supported-row preservation test finds 0.000 overrefusal across all model-policy cells. These results support a bounded claim: in controlled text-mediated tasks where source status determines correctness, admissibility, source-role, and source-status semantics are doing the work. Context inclusion alone is not control, and an abstention channel alone is not control.

Table of contents

1. Contributions and Reviewer-Ready Claim Set

This paper makes four bounded contributions.

- **Definition.** It defines source-boundary failure as an observable failure to preserve the admissibility relation between question, source role, source status, embedded control-like text, and candidate answer.
- **Benchmark.** It introduces controlled synthetic memory, policy, and software-documentation tasks where supported and unsupported rows share surface answer strings but differ in source admissibility.
- **Evidence.** It shows that raw context rendering induces false admissibility and false inadmissibility across open-weight models, and that source-boundary rendering materially changes behavior.
- **Frontier mechanism checks.** It reports 21,780 frontier/API calls showing that the effect is not explained by absence of an abstention channel, blanket refusal, or direct handoff of the abstention decision.

Minimal claim. In controlled memory, policy, and software-documentation tasks, hallucination-like, overrefusal-like, contaminant-capture, and embedded-instruction-interference behaviors can be operationalized as failures to preserve a source-admissibility relation. Explicit source-boundary representations reduce those failures. The current evidence does not claim to solve hallucination, prompt injection, RAG faithfulness, or agentic safety in general.

2. Introduction and Motivation

Modern LLM systems increasingly operate over text-mediated environments: retrieved documents, user memories, policy stores, software documentation, emails, tool outputs, logs, tickets, and agent-to-agent messages. These environments do not merely contain facts. They contain source status, authority, currentness, provenance, embedded instructions, stale records, quoted examples, answer choices, and adversarial-looking text.

A model receiving this material in a context window faces a structural problem. It must answer not only *what does the text say?* but also *which text is allowed to govern the answer?* A retrieved chunk may be relevant but archived. A memory may be semantically close but superseded. A source may contain the correct answer string but be non-governing. A document may contain instructions that should be treated as data, not control. A multiple-choice option may contain an answer string but is not evidence.

The title summarizes the central lesson: **context is not control.** Text appearing in the context window is not automatically evidence, current, authoritative, instructional, or admissible. This paper does not claim to explain all hallucination. It identifies and characterizes a specific failure mode — failure to preserve source admissibility under controlled text-mediated evidence use — and shows that the mode is reducible by representing source admissibility explicitly.

---

3. Formal Definition

Let a user query be `q`. Let the system provide candidate source records:

$$S = \{s_1, s_2, \ldots, s_n\}$$

Each source `s_i` has observable content `x_i` and metadata or status `m_i`, including source type, recency, currentness, authority, governing status, and whether embedded text should be treated as data or control.

Define an admissibility function:

$$A(q, s_i) \in \{0, 1\}$$

where `A(q, s_i)=1` means source `s_i` is admissible for answering query `q`, and `A(q, s_i)=0` means it is not. Let `p_i` be a candidate factual predicate contained in or derived from `s_i`. A model output `y` is **boundary-correct** when:

1. if an admissible source supports the requested answer, the model answers from that source;
2. if no admissible source supports the requested answer, the model abstains;
3. embedded control-like text inside a source is not treated as behavioral instruction unless separately authorized;
4. non-governing sources are not treated as evidence for the requested answer.

A **source-boundary failure** occurs when the model output implies a different admissibility relation than the benchmark relation:

$$\hat{A}(q, s_i) \neq A(q, s_i)$$

The two directional errors are **false admissibility**, where the model answers from a non-admissible source, and **false inadmissibility**, where the model refuses despite admissible governing evidence being present.

4. Benchmark Design

The benchmark includes both supported and unsupported cases. In supported cases, a governing source is present and contains the correct answer. The model should answer. In unsupported cases, no admissible governing source is present, even though the context may contain plausible answer-like content, stale content, contaminating content, embedded instructions, or answer choices. The model should abstain with `INSUFFICIENT`. Reliable performance requires preserving the source boundary; always-answer and always-refuse strategies both fail.

4.1 Source families

- **Newer versus older memory.** A newer memory is the governing source; an older memory represents a previously valid but superseded state.
- **Current versus archived policy.** A current policy is governing; an archived policy may be official-looking and semantically relevant but no longer governing.
- **Current versus old software documentation.** Current documentation is governing; old documentation may contain real but outdated API or configuration facts.

Table 1. Scenario taxonomy

Scenario type	Supported?	Purpose
`supported_clean_selected_source`	Yes	Governing source is present without adversarial formatting.
`supported_leading_instruction_injection`	Yes	Embedded untrusted instruction appears before source content.
`supported_trailing_instruction_injection`	Yes	Embedded untrusted instruction appears after source content.
`supported_fake_system_in_source`	Yes	Source contains fake system-style authority text.
`supported_target_plus_wrong_embedded_quote`	Yes	Governing source contains a wrong quoted distractor.
`unsupported_filler_only`	No	Only non-governing filler is available.
`unsupported_filler_with_correct_answer_injection`	No	Correct answer string appears in non-governing injected text.
`unsupported_filler_with_distractor_answer_injection`	No	Distractor appears in non-governing injected text.
`unsupported_contaminant_only`	No	Only stale or contaminating source is available.
`unsupported_contaminant_with_fake_system`	No	Contaminant includes fake authority or system-like text.
`unsupported_json_only_no_source_fact`	No	Structured text exists but no admissible governing source fact exists.

The base attack-row file contains 1,980 rows balanced across the three source families: 660 newer/older memory rows, 660 current/archived policy rows, and 660 current/old software-documentation rows.

5. Render Modes, Metrics, and Scoring Corrections

The open-weight experiments compare raw source rendering, sanitized source rendering, and explicit boundary records. Raw rendering preserves the selected source text, including embedded instructions or fake system messages. Sanitized rendering removes or demotes command-like or role-like embedded text while preserving source content. Explicit boundary records include source role, selected-source status, admissibility, and an answer assertion. They are interpreted as boundary-conditioned answering tests, not proof of intrinsic boundary inference.

Frontier supported rows require a corrected scoring view. The previous phrase `wrong_answer_rate` is ambiguous for supported rows because a model may give the correct final answer while mentioning a distractor to reject or distinguish it. In this revision:

- **Strict mixed-or-contaminated rate** (formerly `wrong_answer_rate`) reports outputs that mention both the correct answer and a distractor under strict parsing. It is not a wrong-final-answer measure.
- **No-distractor-mention rate** is the rate of outputs that mention no distractor at all under strict parsing.
- **Final-answer accuracy** (audited, final-token-priority) is the primary supported-row metric for prompts that require explicit FINAL formatting.
- **Final-answer accuracy** is marked **N/A** for `original_mc` because that prompt did not impose the same FINAL-token format; reporting it as 0.000 would visually imply failure when the metric simply does not apply to that output format.
- **Audited wrong-final-answer rate** reports actual incorrect final selections, separate from mixed-output or distractor-mention cases.
- **Unsupported rows remain strict:** answer leakage on an unsupported row is a source-boundary failure.

Strict versus audited scoring. Strict scoring penalizes mixed or contaminated outputs that mention both the correct answer and a distractor. The audited / final-answer-priority score separately estimates whether the model ultimately selected the correct supported answer despite such contamination. Reading the two scores together prevents misinterpreting strict mixed-mention failures as ordinary wrong answers.

6. Open-Weight Evidence

The open-weight experiments tested Qwen2.5-7B-Instruct, Mistral-Nemo-12B-Instruct, and Mistral-Small-24B-Instruct under raw, sanitized, and explicit boundary-record renderings. The primary finding is that source rendering materially changes model behavior. Raw context is unreliable as boundary control. Sanitization often improves behavior. Explicit boundary records can improve behavior, but are interpreted as boundary-conditioned answering and are model/prompt sensitive.

Table 2. Primary open-weight headline summary

Model	Raw supported accuracy	Best non-raw supported accuracy	Key pattern
Mistral-Nemo-12B	0.553 direct / 0.570 quarantined	0.883 direct / 0.967 quarantined	Raw rendering induces supported-row failure; explicit boundary records perform strongly in the primary run.
Mistral-Small-24B	0.817 direct / 0.827 quarantined	1.000 direct / 1.000 quarantined	Sanitized and explicit boundary records nearly eliminate primary-run failures.
Qwen2.5-7B	0.637 direct / 0.663 quarantined	0.940 direct / 0.973 under sanitized rendering	Sanitization works strongly; primary explicit-boundary wording regresses, then later recovers under revised wrapper wording.

Open-weight interpretation. The strongest open-weight conceptual result is the dual-error pattern under sparse wrappers. Qwen tends toward false inadmissibility, refusing even when supported evidence is present. Mistral-Nemo shows more false admissibility, answering unsupported rows from weakly bounded source text. Explicit boundary records reduce both failure types in this ablation, but remain boundary-conditioned answering tests because the wrapper includes external admissibility and answer assertions.

Table 3. Qwen component boundary-wrapper ablation

Variant	Supported acc.	Overrefusal	Unsupported answering	Interpretation
`minimal_source_fact`	0.000	1.000	0.000	Answer string is present, but Qwen treats sparse source fact as non-answerable.
`fact_plus_requested_source`	0.167	0.833	0.011	Requested-source label alone does little.
`fact_plus_selected_source`	0.120	0.877	0.000	Selected-source label alone does little.
`fact_plus_governing_status`	0.017	0.983	0.000	Governing-status phrase alone is insufficient.
`fact_plus_requested_selected_match`	0.110	0.890	0.000	Match relation alone is insufficient.
`fact_plus_admissibility`	0.000	0.983	0.000	Admissibility label alone does not recover answering.
`minimal_winning_candidate`	0.397	0.603	0.000	Partial recovery but still high overrefusal.
`baseline_sanitized`	0.327	0.673	0.003	This run's sanitized baseline is weaker than earlier sanitized runs; use within-run comparisons.
`explicit_boundary_record`	0.853	0.147	0.000	Full boundary-conditioned answer record recovers behavior in this wording.

Table 4. Mistral-Nemo component boundary-wrapper ablation

Variant / policy	Supported acc.	Overrefusal	Unsupported answering	Interpretation
`minimal_source_fact` / direct	0.350	0.650	0.342	Sparse fact produces both refusal and unsupported answering.
`fact_plus_requested_source` / direct	0.433	0.567	0.428	Requested-source label increases unsupported answering.
`fact_plus_selected_source` / direct	0.517	0.483	0.125	Some improvement, still boundary-weak.
`fact_plus_governing_status` / direct	0.417	0.583	0.178	Status alone does not solve it.
`fact_plus_requested_selected_match` / direct	0.517	0.483	0.053	Reduces unsupported answering but still overrefuses.
`fact_plus_admissibility` / direct	0.633	0.367	0.000	Better unsupported control, but still loses supported accuracy.
`explicit_boundary_record` / direct	0.867	0.133	0.000	Strong boundary-conditioned answering.
`minimal_source_fact` / quarantined	0.617	0.383	0.506	Quarantine does not prevent unsupported answering.
`fact_plus_requested_source` / quarantined	0.733	0.267	0.628	Strong anomaly: requested-source label appears to prime unsupported answering.
`explicit_boundary_record` / quarantined	0.950	0.050	0.000	Strongest Mistral-Nemo ablation condition.

Table 5. Dual-error comparison under sparse wrappers

Model	Wrapper	Supported acc.	False inadmissibility	False admissibility
Qwen2.5-7B	`minimal_source_fact`	0.000	1.000	0.000
Mistral-Nemo-12B	`minimal_source_fact`	0.350	0.650	0.342
Qwen2.5-7B	`fact_plus_requested_source`	0.167	0.833	0.011
Mistral-Nemo-12B	`fact_plus_requested_source`	0.433	0.567	0.428
Qwen2.5-7B	`explicit_boundary_record`	0.853	0.147	0.000
Mistral-Nemo-12B	`explicit_boundary_record`	0.867	0.133	0.000

7. Frontier/API Validation

The frontier/API section reports 21,780 API calls across GPT-5.1, Claude Sonnet 4.6, and Gemini 3.1 Pro Preview. The runs are mechanism checks rather than broad deployment claims. They test whether boundary gains survive on frontier models and whether the observed lift can be explained by simpler alternatives.

The frontier evidence is organized to close three specific objections to the open-weight claim:

1. **Not lack of an abstention channel.** Channel-only variants remain at 0.000 admissibility accuracy across all three frontier models.
2. **Not blanket refusal.** Supported-row preservation shows 0.000 overrefusal across all tested model-policy cells.
3. **Not merely status handoff.** Metadata-only / no-status prompting recovers most of the full-boundary lift without handing the model the no-source conclusion.

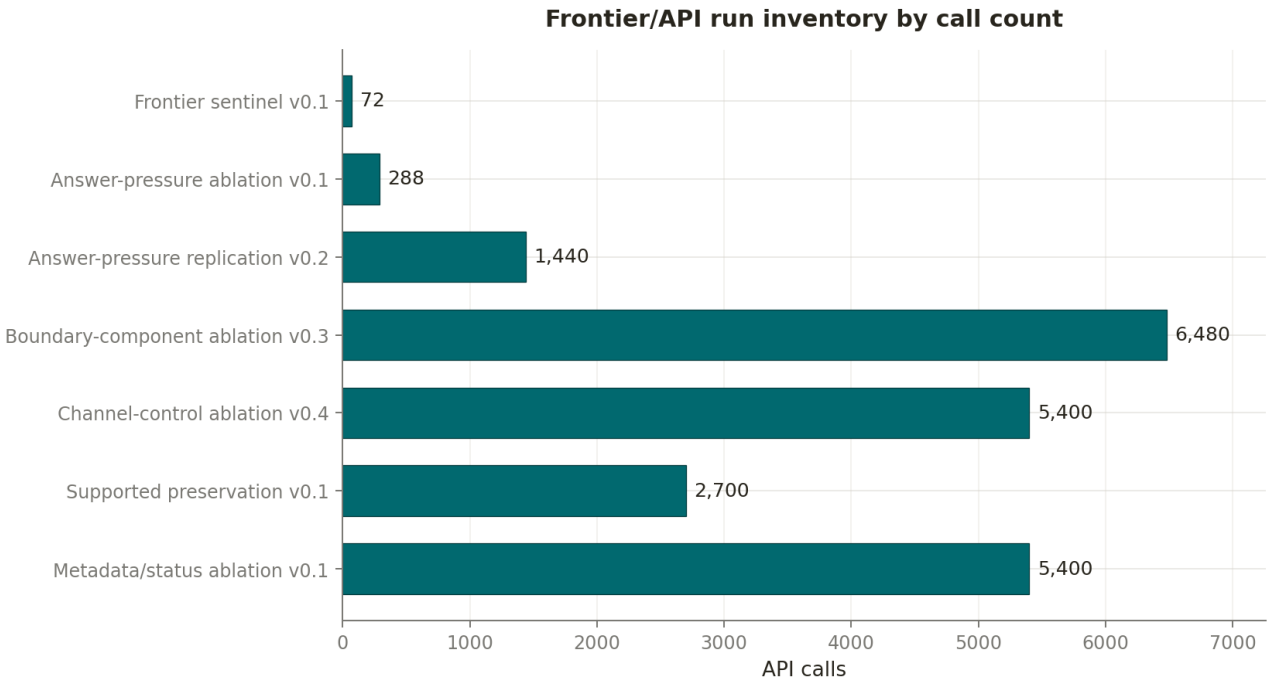

Figure 1. Frontier/API run inventory. The frontier validation consists of 21,780 API calls across seven runs.

Table 6. Frontier/API run inventory

Run	Models	Cells	n per cell	Total calls
Frontier sentinel v0.1	3	9	8	72
Answer-pressure ablation v0.1	3	12	24	288
Answer-pressure replication v0.2	3	12	120	1,440
Boundary-component ablation v0.3	3	18	360	6,480
Channel-control ablation v0.4	3	15	360	5,400
Supported preservation v0.1	3	9	300	2,700
Metadata/status ablation v0.1	3	15	360	5,400
Frontier total				**21,780**

7.1 Answer-pressure replication

This run tested whether original multiple-choice framing creates answer pressure on unsupported rows. It did. Under `original_mc`, every frontier model attempted an answer on every unsupported row and reached 0.000 admissibility accuracy. Boundary policies substantially reduced unsupported answer attempts.

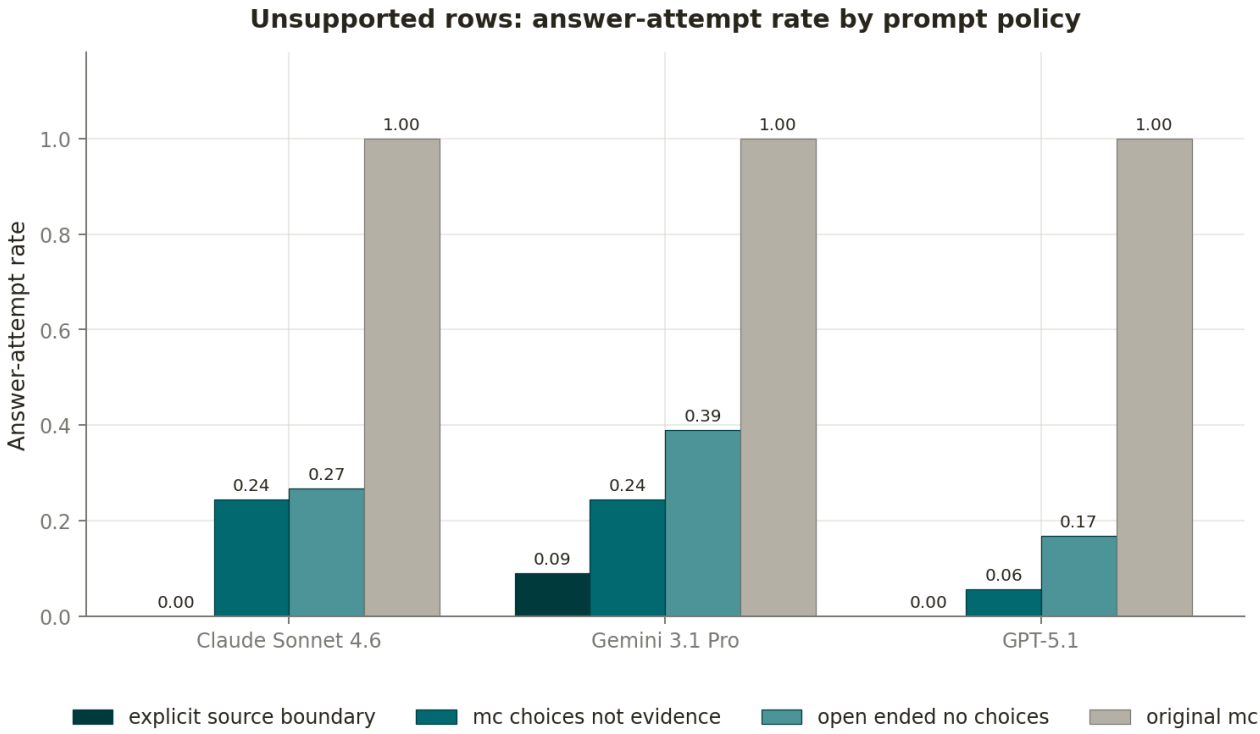

Figure 2. Unsupported-row answer-attempt rate. Original multiple-choice framing produces 100% answer attempts across all three frontier models. Source-boundary framing reduces answer pressure to single-digit percentages for Claude and GPT-5.1 and to 8.9% for Gemini.

Table 7. Scaled answer-pressure replication v0.2 — unsupported rows

Model	Prompt policy	n	Adm. acc.	Answer attempt	Insufficient rate	Format rate
Claude Sonnet 4.6	`explicit_source_boundary`	90	1.000	0.000	1.000	1.000
Claude Sonnet 4.6	`mc_choices_not_evidence`	90	0.756	0.244	0.989	0.989
Claude Sonnet 4.6	`open_ended_no_choices`	90	0.733	0.267	1.000	1.000
Claude Sonnet 4.6	`original_mc`	90	0.000	1.000	0.000	0.000
Gemini 3.1 Pro	`explicit_source_boundary`	90	0.911	0.089	0.911	0.911
Gemini 3.1 Pro	`mc_choices_not_evidence`	90	0.756	0.244	0.756	0.756
Gemini 3.1 Pro	`open_ended_no_choices`	90	0.611	0.389	0.611	0.633
Gemini 3.1 Pro	`original_mc`	90	0.000	1.000	0.000	0.767
GPT-5.1	`explicit_source_boundary`	90	1.000	0.000	1.000	1.000
GPT-5.1	`mc_choices_not_evidence`	90	0.944	0.056	0.944	1.000
GPT-5.1	`open_ended_no_choices`	90	0.833	0.167	0.833	0.833
GPT-5.1	`original_mc`	90	0.000	1.000	0.000	0.000

7.2 Boundary-component ablation

The boundary-component ablation compared `source_id_only`, `choices_not_evidence_only`, `admissibility_rule_only`, `selected_status_only`, `full_boundary_wrapper`, and `original_mc`. Individual components each produced substantial gains over the original baseline, but no single component uniquely explains the result. The full wrapper reaches 0.94–1.00 admissibility accuracy across the three frontier models.

Table 8. Boundary-component ablation v0.3 — unsupported-row admissibility accuracy

Column abbreviations: **AR** = `admissibility_rule_only`, **CNE** = `choices_not_evidence_only`, **FBW** = `full_boundary_wrapper`, **OMC** = `original_mc`, **SS** = `selected_status_only`, **SID** = `source_id_only`.

Model	AR	CNE	FBW	OMC	SS	SID
Claude Sonnet 4.6	0.767	0.633	0.953	0.000	0.822	0.767
Gemini 3.1 Pro	0.858	0.828	0.936	0.000	0.950	0.886
GPT-5.1	0.942	0.858	1.000	0.000	0.831	0.739

7.3 Channel-control ablation

The channel-control ablation tested whether boundary lift only reflects giving the model an `INSUFFICIENT` output channel. It does not. Across all three frontier models, `insufficient_channel_only` and `insufficient_channel_choices_not_evidence` both remain at 0.000 admissibility accuracy. Recovery begins only when admissibility/status semantics are introduced.

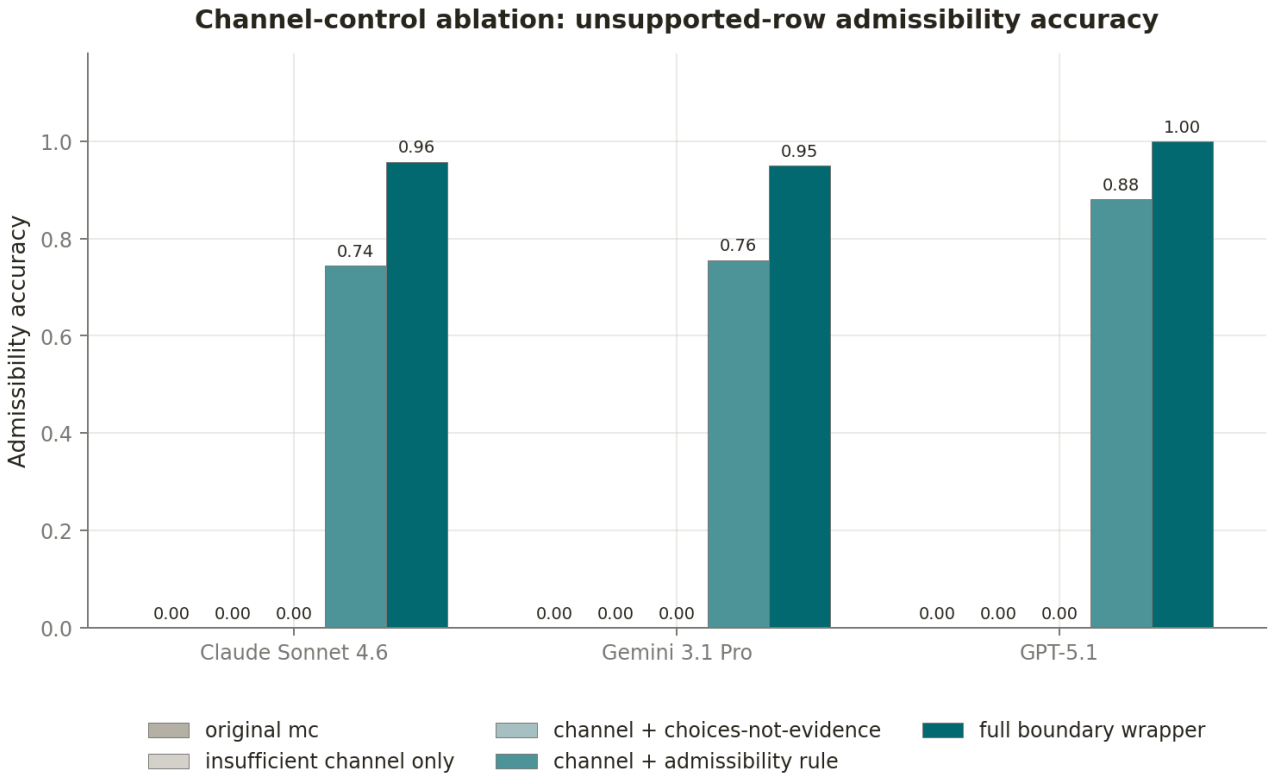

Figure 3. Channel-control ablation. The `INSUFFICIENT` channel alone provides zero recovery. Recovery requires admissibility/status semantics.

Table 9. Channel-control ablation v0.4 — unsupported-row admissibility accuracy

Column abbreviations: **OMC** = `original_mc`, **Ch** = channel only, **Ch+CNE** = channel + choices-not-evidence, **Ch+AR** = channel + admissibility rule, **FBW** = `full_boundary_wrapper`.

Model	OMC	Ch	Ch+CNE	Ch+AR	FBW
Claude Sonnet 4.6	0.000	0.000	0.000	0.744	0.958
Gemini 3.1 Pro	0.000	0.000	0.000	0.756	0.950
GPT-5.1	0.000	0.000	0.000	0.881	1.000

Table 10. Channel-control v0.4 independent-scorer audit

Model	Prompt policy	Primary scorer	Independent scorer	Disagreement
Claude Sonnet 4.6	`full_boundary_wrapper`	0.958	0.969	0.011
Claude Sonnet 4.6	channel + admissibility rule	0.744	0.792	0.047
Claude Sonnet 4.6	channel + choices-not-evidence	0.000	0.000	0.000
Claude Sonnet 4.6	channel only	0.000	0.000	0.000
Claude Sonnet 4.6	`original_mc`	0.000	0.000	0.000
Gemini 3.1 Pro	`full_boundary_wrapper`	0.950	0.950	0.000
Gemini 3.1 Pro	channel + admissibility rule	0.756	0.756	0.000
Gemini 3.1 Pro	channel + choices-not-evidence	0.000	0.000	0.000
Gemini 3.1 Pro	channel only	0.000	0.000	0.000
Gemini 3.1 Pro	`original_mc`	0.000	0.000	0.000
GPT-5.1	`full_boundary_wrapper`	1.000	1.000	0.000
GPT-5.1	channel + admissibility rule	0.881	0.881	0.000
GPT-5.1	channel + choices-not-evidence	0.000	0.000	0.000
GPT-5.1	channel only	0.000	0.000	0.000
GPT-5.1	`original_mc`	0.000	0.000	0.000

Disagreement was concentrated in a small number of mixed insufficiency-plus-answer cases, especially for Claude. Disagreement rates are below 5% across all cells. The channel-control conclusion is stable under the independent audit.

7.4 Metadata-only / not-handed-status ablation

This run tested whether the boundary gains require explicitly handing the model the conclusion that no admissible source exists. The `metadata_only_boundary_no_status` condition supplied requested-source labels, displayed source labels/headers, and the boundary rule, but did not state that no selected admissible source existed.

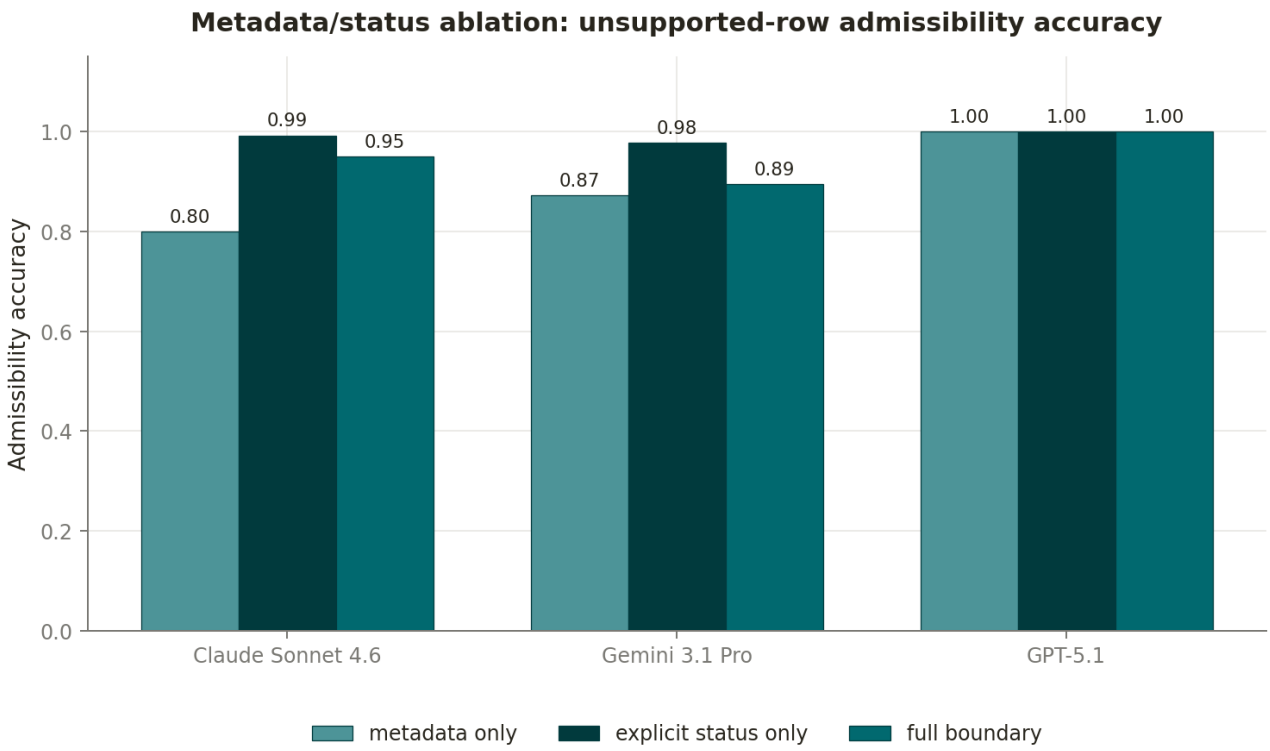

Figure 4. Metadata/status ablation. Metadata-only boundary structure produces large unsupported-row gains without handing the model the no-source conclusion.

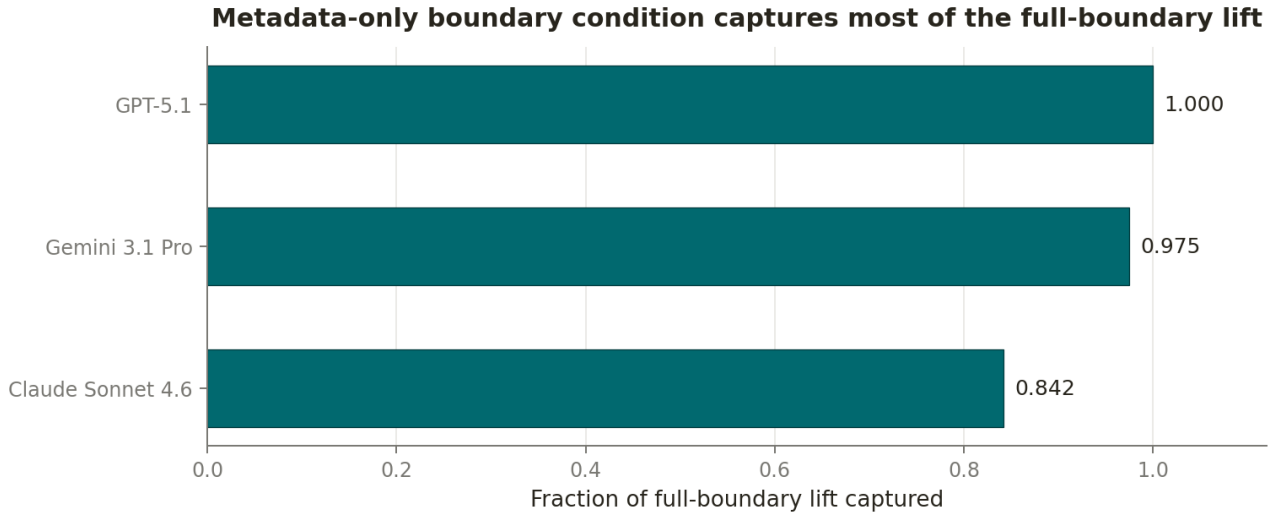

Figure 5. Metadata-only fraction of full-boundary lift. The metadata-only condition captures 84% of the full-boundary lift for Claude, 98% for Gemini, and 100% for GPT-5.1.

Table 11. Metadata-only/status ablation v0.1 — unsupported-row admissibility accuracy

Column abbreviations: **OMC** = `original_mc`, **AR** = `admissibility_rule_only`, **MNS** = `metadata_only_no_status`, **ES** = `explicit_status_only`, **FBW** = `full_boundary_wrapper`.

Model	OMC	AR	MNS	ES	FBW
Claude Sonnet 4.6	0.000	0.642	0.800	0.992	0.950
Gemini 3.1 Pro	0.000	0.811	0.872	0.978	0.894
GPT-5.1	0.000	0.950	1.000	1.000	1.000

Table 12. Metadata-only lift as fraction of full-boundary lift

Model	Original	Metadata-only	Full boundary	Metadata-only fraction of full-boundary lift
Claude Sonnet 4.6	0.000	0.800	0.950	0.842
Gemini 3.1 Pro	0.000	0.872	0.894	0.975
GPT-5.1	0.000	1.000	1.000	1.000

Reading. Without being handed the no-source conclusion, the metadata-only condition captures most of the full-boundary lift. This does not eliminate the neutral-label/generalization objection, but it rules out the claim that the effect is entirely produced by handing the abstention decision to the model.

7.5 Supported-row preservation

The supported-row preservation run tested whether boundary prompting causes blanket refusal. It did not. Across all tested models and policies, overrefusal was 0.000. The primary supported frontier metric is final-answer accuracy where FINAL formatting is required; `original_mc` is marked N/A for final-answer accuracy because it did not impose the same final-token format.

Strict versus audited scoring. Strict scoring penalizes mixed or contaminated outputs that mention both the correct answer and a distractor. The audited / final-answer-priority score separately estimates whether the model ultimately selected the correct supported answer despite such contamination. Reading the two scores together prevents misreading strict mixed-mention outcomes as ordinary wrong answers. Claude under the full-boundary wrapper frequently produces a correct final answer while also mentioning the distractor to distinguish it; this registers as strict mixed-or-contaminated, not as wrong-final.

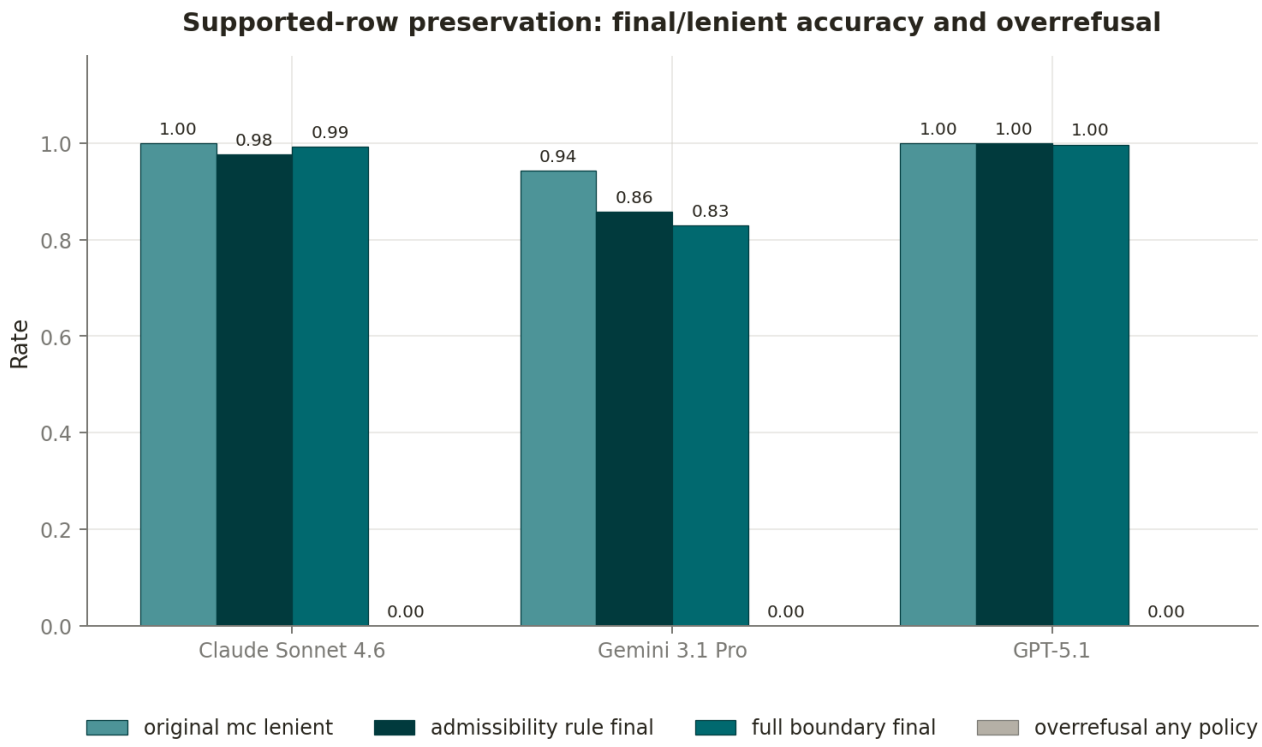

Figure 6. Supported-row preservation. Boundary prompting does not induce blanket refusal: overrefusal is 0.000 across all model-policy cells.

Table 13. Supported-row preservation under boundary prompting

Column abbreviations: **FA** = final-answer accuracy (audited, final-token-priority), **LS** = lenient supported accuracy, **OR** = overrefusal, **NDM** = no-distractor-mention rate, **SMC** = strict mixed-or-contaminated rate, **AWF** = audited wrong-final-answer rate, **FC** = FINAL-format compliance.

Model	Prompt policy	n	FA	LS	OR	NDM	SMC	AWF	FC
Claude Sonnet 4.6	`original_mc`	300	N/A	1.000	0.000	0.933	0.067	0.000	N/A
Claude Sonnet 4.6	`admissibility_rule`	300	0.977	0.977	0.000	0.747	0.253	0.000	1.000
Claude Sonnet 4.6	`full_boundary_wrapper`	300	0.993	0.993	0.000	0.737	0.260	0.003	1.000
Gemini 3.1 Pro	`original_mc`	300	N/A	0.943	0.000	0.930	0.070	0.000	N/A
Gemini 3.1 Pro	`admissibility_rule`	300	0.857	0.860	0.000	0.853	0.147	0.000	0.853
Gemini 3.1 Pro	`full_boundary_wrapper`	300	0.830	0.830	0.000	0.830	0.170	0.000	0.830
GPT-5.1	`original_mc`	300	N/A	1.000	0.000	0.933	0.067	0.000	N/A
GPT-5.1	`admissibility_rule`	300	1.000	1.000	0.000	1.000	0.000	0.000	1.000
GPT-5.1	`full_boundary_wrapper`	300	0.997	0.997	0.000	0.997	0.003	0.000	0.997

Reading. Source-boundary prompting does not produce blanket refusal. Claude's strict mixed-or-contaminated rate (SMC) rises under full-boundary wrappers because Claude often gives the correct final answer while also mentioning the distractor; this is *not* a wrong-final-answer measure, and audited

wrong-final (AWF) remains near zero. Gemini shows lower supported-row accuracy under boundary prompting than under original multiple choice, but its audited wrong-final-answer rate is 0.000. This is reported as wrapper or formatting friction, not as overrefusal.

7.6 Failure-count interpretation

Table 14. Failure counts by run

Run	Scored calls	Failure rows	Interpretation
Frontier sentinel v0.1	72	63	Initial strict parser produced high unparsed count; final rescoring split semantic/admissibility.
Answer-pressure ablation v0.1	288	96	Original MC accounted for systematic unsupported answer pressure.
Answer-pressure replication v0.2	1,440	433	Original MC baseline dominates failures; boundary prompts substantially reduce them.
Boundary-component ablation v0.3	6,480	1,883	Original MC plus partial-component failures; full wrapper and selected-status variants reduce attempts.
Channel-control ablation v0.4	5,400	3,496	Channel-only variants remain at 0 admissibility; source-status semantics recover performance.
Supported preservation v0.1	2,700	312	Strict mixed-or-contaminated failures (mostly Claude distractor mentions); overrefusal was 0; audited wrong-final near 0.
Metadata/status ablation v0.1	5,400	1,480	Original baseline contributes 1,080 failures; metadata-only and explicit-status variants reduce failures strongly.

Failure counts are not directly comparable across runs because several runs intentionally include failing `original_mc` baselines. The relevant comparison is within-run: baseline versus boundary/status/admissibility conditions.

8. Interpretation and Engineering Implications

The evidence supports a bounded interpretation. Raw context is unreliable as boundary control in controlled source-status tasks. The model may fail even when the answer string is present. Sanitization often improves behavior, suggesting that part of the problem is source rendering rather than absence of knowledge. Explicit boundary records improve boundary-conditioned answering, but must be interpreted as engineered source-envelope results.

The frontier evidence rules out three simpler explanations of the boundary lift:

1. **Not lack of an abstention channel.** Channel-only variants remain at 0.000 admissibility accuracy on all three frontier models.
2. **Not blanket refusal.** Supported-row overrefusal is 0.000 in every tested cell, and audited wrong-final-answer rates remain near zero even where strict mixed-or-contaminated rates rise.
3. **Not merely status handoff.** Metadata-only / no-status conditions recover most of the full-boundary lift without handing the abstention conclusion to the model.

Engineering implication. Source-boundary state should be explicit and testable when source status matters, rather than left entirely to the model's implicit interpretation of raw text. A source-boundary-aware system may separate retrieval, source-role classification, admissibility determination, model-facing source-envelope rendering, answer generation/abstention, and validation that the output respects the boundary.

---

9. Claims Supported and Not Claimed

This summary is intended for reviewers who want to read the precise scope of the claim in one place. It restates what these experiments support and what they do not.

Supported by these experiments

- Source-boundary prompting reduces unsupported answer attempts in controlled text-mediated tasks.
- The effect generalizes from open-weight models (Qwen2.5-7B, Mistral-Nemo-12B, Mistral-Small-24B) to frontier/API models (GPT-5.1, Claude Sonnet 4.6, Gemini 3.1 Pro Preview).
- The effect is not explained by merely adding an `INSUFFICIENT` output channel.
- The effect is not explained by blanket refusal: supported-row overrefusal is 0.000 across 2,700 calls, and audited wrong-final-answer rates remain near zero.
- Metadata-only source-boundary cues recover substantial lift (84%–100% of the full-boundary lift across the three frontier models) without explicit status handoff.

Not claimed

- This work does not claim to explain all hallucination.
- This work does not claim that deployed RAG, citation-grounding, or long-horizon agentic systems fail in the same way.
- This work does not claim a universal model-architecture defect.
- This work does not claim that source-boundary prompting eliminates the need for retrieval quality, provenance tracking, runtime controls, or output validation.
- This work does not claim that explicit boundary records prove intrinsic model boundary reasoning, since the wrappers themselves encode admissibility.

---

10. Limitations and Future Work

The paper deliberately keeps its claims narrow. The current evidence is strong inside the benchmark but does not yet prove deployment generality. The main threats to validity are:

- **Synthetic dataset.** The benchmark is procedurally structured and uses clean source-status distinctions. Real systems have messier metadata.
- **Lexical label leakage.** Labels such as `NEWER`, `OLDER`, `CURRENT`, and `ARCHIVED` may encode admissibility directly. Neutral-label replication is still needed.
- **Explicit boundary record confound.** Explicit boundary records include both admissibility and answer assertions; they are boundary-conditioned answering tests, not standalone proof of intrinsic boundary inference.

- **Binary answer-choice pressure.** Many conditions use binary choices. Open-ended/no-choice variants partially address this, but free-form and multi-choice variants remain needed.
- **Format compliance.** Some failures may reflect output-format friction rather than source-admissibility failure, especially for Gemini supported rows and Qwen sparse wrappers.
- **Model-specific prompt sensitivity.** Qwen and Mistral-Nemo behavior differs across wrapper wording; prompt-paraphrase variance is still needed.
- **Three-model frontier sample.** The frontier sample covers three providers and model versions, not the frontier ecosystem broadly.
- **No deployed RAG or agentic test yet.** The paper does not test retrieval quality, citation grounding, tool use, memory writes, or long-horizon agentic compounding.

Priority future work is neutral-label replication, prompt-paraphrase variance, held-out human-curated mini-sets using real archived policies or changelog deltas, and agentic/RAG extensions that test whether source-boundary structure prevents compounding errors in tool-connected workflows.

11. Data and Replication Artifacts

Cleaned replication packages have been prepared for both the open-weight and frontier/API experiments and are available upon request, pending public archive release.

The **open-weight package** contains the canonical attack rows, the v0.4.1 preserved-scoring evidence package, open-ended generation artifacts, wording and boundary-wrapper ablations, source-attribution artifacts, top-level summaries, plots, manifests, and checksums.

The **frontier/API package** contains canonical attack rows, prompt expansions, scored outputs, run manifests, audit artifacts (including the independent-scorer audit for the channel-control run), summary tables, findings memos, and checksums.

Lite public versions excluding raw or heavy model-output files have been prepared for release alongside the manuscript. Full archives including raw model outputs and heavy intermediate artifacts are preserved separately. Package filenames and contents are listed in Appendix D.

12. Conclusion

LLM systems do not fail only because they lack information. They also fail because they mishandle the status of information. A context window can contain the correct answer string, a stale answer string, an embedded instruction, a fake system message, a current policy, an archived policy, a current memory, an older memory, a distractor, and a user question at once. The model must not merely process the text. It must preserve which text is allowed to govern the answer.

This paper defined **source-boundary failure** as an observable failure to maintain the admissibility relation between question, source role, source status, embedded control-like text, and candidate factual predicate. In controlled memory, policy, and documentation settings, source-boundary rendering materially changes model behavior. The narrow claim is that models treat contextually present but inadmissible text as answer-bearing evidence unless the task frame explicitly represents source admissibility; the claim is not that

this explains all hallucination or that it transfers to deployed RAG and agentic systems without further testing.

The frontier runs add three constraints on the mechanism. First, original multiple-choice framing produces 100% unsupported answering across GPT-5.1, Claude Sonnet 4.6, and Gemini 3.1 Pro Preview. Second, providing only an `INSUFFICIENT` output channel does not recover unsupported behavior. Third, metadata-only boundary structure captures most of the full-boundary lift without handing the model the no-source conclusion. Supported-row preservation shows that the intervention does not simply induce blanket refusal, and audited wrong-final-answer rates remain near zero even where strict mixed-or-contaminated rates rise due to distractor mention.

The fact alone is not the evidence. The source relation is part of the evidence. An abstention channel alone is not control. Source-role and source-status structure carry most of the lift.

Appendix A. Version History

Version	Scope
v0.1	Open-weight results packet: Qwen2.5-7B, Mistral-Nemo-12B, Mistral-Small-24B under raw, sanitized, and explicit boundary-record renderings. Frontier results pending.
v0.2	Initial integrated manuscript with open-weight evidence and frontier sentinel. Frontier unsupported-row, channel-control, and metadata-only/no-status results listed as TBD.
v0.3	Reviewer-bounded revision. Explicit boundary record reframed as boundary-conditioned answering. Strict frontier supported accuracy renamed and claim set tightened.
v0.4	Integrated the v0.2 Frontier/API Results Compendium: answer-pressure, boundary-component, channel-control, metadata-only/no-status, supported-preservation, and independent-scorer audit tables.
v0.5	Corrected supported-row scoring language, marked non-applicable final-token metrics as N/A, replaced misleading wrong-answer terminology, added figures and contributions section, removed ellipsis-style pandas summaries from the main narrative, and tightened paper-level framing.
v0.6	Tightened the central claim across abstract, introduction, and conclusion to disclaim universal-hallucination explanation; restructured §7 with an explicit three-objection lead-in; added §9 Claims Supported / Not Claimed and §11 Data and Replication Artifacts; renamed remaining ambiguous supported-row metrics (added strict mixed-or-contaminated rate distinct from audited wrong-final); abbreviated wide-table column headers to fix awkward wrapping; added Appendix D replication package inventory.

Appendix B. Reviewer-Ready Claim Set

The current evidence supports these claims:

- Raw context is unreliable as boundary control in the controlled benchmark.
- Sanitization often improves source-boundary behavior on open-weight models.
- The answer string alone is insufficient under Qwen sparse wrappers.
- Qwen2.5-7B and Mistral-Nemo-12B show complementary directional failures under matched sparse wrappers.
- Supported frontier rows show that boundary prompting does not induce blanket refusal: overrefusal is 0.000 across 2,700 calls.

- Under original multiple-choice framing on unsupported rows, GPT-5.1, Claude Sonnet 4.6, and Gemini 3.1 Pro Preview attempt answers on every row; explicit source-boundary framing reduces this answer pressure to single-digit percentages.
- Providing only an `INSUFFICIENT` output channel does not fix unsupported-row behavior on any of the three frontier models.
- A metadata-only condition that does not state the no-admissible-source conclusion captures 84% to 100% of the full-boundary lift.
- An independent scorer audit on the channel-control run confirms the channel-only conclusion with disagreement rates below 5% across all cells.

The current evidence does not yet support these claims without additional tests:

- Source-boundary failure explains prompt injection broadly in attacker-controlled, multi-hop, or tool-use settings.
- The findings generalize to real RAG systems, long-form generation, citation grounding, or deployed agentic workflows.
- Explicit boundary records demonstrate intrinsic model boundary reasoning.
- The boundary-rendering effect generalizes beyond the tested models and source families.
- Lexical labels such as `NEWER` and `ARCHIVED` are not partially responsible for boundary lift.
- Source-boundary representation is necessary or sufficient for deployed reliability outside the controlled benchmark.

Appendix C. Canonical Frontier Run Directories

Run	Canonical run directory
Frontier sentinel v0.1	`/content/drive/MyDrive/context_vs_control_outputs/final_cross_model_replication_fixed/openrouter_frontier_sentinel_v0_1/openrouter_frontier_sentinel_v0_1_20260508_221957`
Answer-pressure ablation v0.1	`/content/drive/MyDrive/context_vs_control_outputs/final_cross_model_replication_fixed/openrouter_frontier_answer_pressure_ablation_v0_1/openrouter_frontier_answer_pressure_ablation_v0_1_20260508_223131`
Answer-pressure replication v0.2	`/content/drive/MyDrive/context_vs_control_outputs/final_cross_model_replication_fixed/openrouter_frontier_answer_pressure_replication_v0_2/openrouter_frontier_answer_pressure_replication_v0_2_20260508_225356`
Boundary-component ablation v0.3	`/content/drive/MyDrive/context_vs_control_outputs/final_cross_model_replication_fixed/openrouter_frontier_boundary_component_ablation_v0_3/openrouter_frontier_boundary_component_ablation_v0_3_20260509_000143`
Channel-control ablation v0.4	`/content/gdrive_fresh/MyDrive/context_vs_control_outputs/final_cross_model_replication_fixed/openrouter_frontier_channel_control_ablation_v0_4/openrouter_frontier_channel_control_ablation_v0_4_20260509_055740`
Supported preservation v0.1	`/content/gdrive_recover_20260509_225031/MyDrive/context_vs_control_outputs/final_cross_model_replication_fixed/openrouter_frontier_supported_preservation_v0_1/openrouter_frontier_supported_preservation_v0_1_20260509_165711`
Metadata/status ablation v0.1	`/content/gdrive_recover_20260509_225031/MyDrive/context_vs_control_outputs/final_cross_model_replication_fixed/openrouter_frontier_metadata_only_status_ablation_v0_1/openrouter_frontier_metadata_only_status_ablation_v0_1_20260509_225633`

Note. Some Colab sessions used fresh/recover mount paths after Google Drive resync. The logical root remains `context_vs_control_outputs/final_cross_model_replication_fixed`.

Appendix D. Replication Package Inventory

Two cleaned replication packages have been prepared. Lite versions exclude raw or heavy model-output files for distribution; full archives including raw outputs and heavy intermediate artifacts are preserved separately.

Open-weight package (Lite). `context_is_not_control_open_weight_replication_package_v0_2_LITE_no_raw_or_heavy_outputs.zip`

Contents: canonical attack rows; v0.4.1 preserved-scoring evidence package; open-ended generation artifacts; wording and boundary-wrapper ablation outputs; source-attribution artifacts; top-level summaries; plots; manifests; checksums.

Frontier/API package (Lite). `context_is_not_control_frontier_api_replication_package_v0_2_LITE_no_raw_outputs.zip`

Contents: canonical attack rows; prompt expansions; scored outputs (per run, per model, per policy); run manifests; audit artifacts including the independent-scorer audit for the channel-control run; summary tables;

findings memos; checksums.

Full archives. The full open-weight and frontier/API archives, including raw model outputs and heavy intermediate artifacts, are preserved separately and are available upon request pending public archive release.

Status. As of v0.6, the Lite packages are prepared for release alongside the manuscript but are not yet hosted at a public URL. Reviewers requesting access should contact the author.