EVIDENCE-GATED RETRIEVAL SEPARATES GROUNDING FROM ANSWERABILITY IN STATELESS LLM SYSTEMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models often conflate whether they can answer a prompt with whether that answer is actually grounded in evidence, collapsing two separable concerns into one brittle refusal decision. We introduce Human Interaction Profiles (HIP), a stateless, inference-time control layer that conditionally injects retrieved evidence and enforces predefined behavioral boundaries using a fixed, deterministic routing scheme, with no fine-tuning and no persistent memory. We evaluate whether this deterministic design behaves as specified across four complementary evaluations spanning exactly 2,630 model outputs and three model providers: a paired answerability comparison (500 no-retrieval vs. 500 evidence-gated outputs) showed 0% refusal in both conditions; a grounding-signal validation (50 vs. 50 outputs) showed a 0% vs. 100% operational grounding-signal rate; EXP-G1 (30 profiles $\times$ 41 prompts; 1,230 outputs) held 100% on all predefined deterministic evidence-signal metrics with 0% refusals; and a three-provider Phase-B suite (300 calls) showed a predefined refusal boundary enforced by the control plane remained exact across all 120 refusal-required trials. Together, these results verify the intended pipeline behavior under the evaluated configurations; the Phase-B result additionally demonstrates provider-invariant enforcement of the predefined refusal boundary. This determinism is boundary-specific, not universal, as an exact five-word formatting constraint succeeded in only 15/30 trials. Grounding here is an operational, deterministic evidence-signal measure, not a semantic-correctness judgment, and these results should be read as a systems-verification finding rather than an independent discovery about model behavior or semantic quality.

1 INTRODUCTION

Retrieval-augmented generation has emerged as a common strategy for improving factual accuracy and safety in large language models (Lewis et al., 2020). By injecting external evidence at inference time, retrieval is often assumed to simultaneously improve answer quality, reduce hallucinations (Ji et al., 2023; Huang et al., 2025), and mitigate unsafe behavior. A range of retrieval backends have been proposed to support this injection, including dense passage retrieval (Karpukhin et al., 2020), late-interaction retrieval (Khattab and Zaharia, 2020), retrieval-augmented pretraining (Guu et al., 2020), and generative retrieval-reader architectures (Izacard and Grave, 2021). In practice, retrieval-based systems conflate multiple distinct concerns, making it difficult to reason about their behavior or evaluate their effectiveness. In particular, answerability, refusal behavior, and grounding are frequently treated as a single outcome, despite arising from different system-level mechanisms.

This conflation obscures a central question: when a model fails to provide a satisfactory response, is the failure due to insufficient evidence, overly conservative refusal policies, or an inability to anchor its output to available information? Many systems resolve this ambiguity by favoring refusal, treating non-response as a proxy for safety, an approach also pursued at training time by instructing models to abstain when a question falls outside their parametric knowledge (Zhang et al., 2024). While this approach may reduce certain risks, it incurs significant downstream costs by preventing otherwise valid, well-scoped answers.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

This ambiguity is further compounded by evaluation methodologies that rely on the model itself to assess response quality (Zheng et al., 2023), as well as by sampling-based hallucination detection methods (Manakul et al., 2023) and dedicated hallucination benchmarks (Li et al., 2023), both of which target a related but distinct question of whether generated content is factually supported, rather than whether an answer was produced at all. When a large language model is used as a judge of its own outputs, refusal behavior, factual alignment, and stylistic coherence are frequently entangled. Post-hoc revision and attribution methods such as RARR (Gao et al., 2023a) attempt to address a related problem, checking and revising claims against evidence after generation, but do not by themselves separate whether an answer was produced at all (answerability) from whether it is anchored in evidence (grounding).

In this work, we argue that answerability and grounding are orthogonal concerns that must be addressed at different layers of the system. Answerability is primarily a function of prompt scoping and response policy, while grounding depends on the controlled injection and use of external evidence. This is distinct from alignment approaches that shape model behavior through training-time intervention, such as Constitutional AI (Bai et al., 2022) or targeted human judgments (Glaese et al., 2022): HIP instead operates entirely at inference time, without modifying model parameters. We introduce Human Interaction Profiles (HIP) as a concrete system design that enforces this separation.

Contributions.

1. We distinguish answerability and grounding as orthogonal system-level axes.
2. We introduce HIP, a stateless inference-time alignment layer with evidence-gated retrieval.
3. We evaluate paired answerability between a no-retrieval and an evidence-gated condition at two scales (500 and 50 outputs per condition), directly measuring operational grounding-signal outcomes at the smaller scale, and separately evaluate the evidence-gated configuration at larger scale in a generalization study (EXP-G1) spanning 30 synthetic profiles and 41 frozen prompts per profile (1,230 outputs total; not a per-condition count).
4. We report cross-provider refusal-enforcement behavior (Phase-B: 3 providers, 2 HIP presets, 5 repeats, 300 stateless calls) with behavior-specific outcomes, showing that determinism holds for the targeted refusal boundary (120/120) but does not extend uniformly to every output-format constraint (e.g., 15/30 for an exact five-word check).

2 DEFINITIONS

Answerability refers to a model's willingness to produce a response to a well-scoped prompt. Answerability is operationalized by the absence of refusal behavior and does not imply factual correctness or evidence use.

Refusal occurs when a model explicitly declines to provide an answer, typically citing uncertainty or policy constraints.

Grounding describes the extent to which a response is anchored in external evidence rather than generated solely from model priors.

Evidence Gating is a mechanism that conditionally injects external evidence at inference time based on explicit prompt intent and system configuration.

Primary Domain refers to structured contextual information that shapes how a response should be framed or interpreted but does not itself serve as external evidence.

Secondary Evidence Domain refers to an external knowledge source from which retrievable evidence is drawn to support grounding. We use "secondary-domain" as the corresponding adjective (e.g., secondary-domain anchors) throughout.

Abstraction Bias denotes a system-level behavior in which grounded responses favor paraphrased, conceptually aligned reasoning over verbatim reproduction of sources.

3 System Design and Methodology

Human Interaction Profiles define a structured alignment layer that operates entirely at inference time. A HIP specifies how primary-domain context is provided to the model and how secondary-evidence-domain evidence may be conditionally injected. The system does not modify model parameters, maintain interaction history, or encode behavioral preferences across sessions, and is designed to remain compatible with heterogeneous model families, including large general-purpose models such as GPT-4 (OpenAI, 2023).

This inference-time framing distinguishes HIP from agentic approaches that interleave reasoning and tool calls within a single decoding trajectory, such as Toolformer (Schick et al., 2023) and ReAct (Yao et al., 2023). Rather than expanding the model's action space with callable tools, HIP constrains a single inference call through conditional evidence injection, which keeps the system stateless and avoids the additional latency and failure surface introduced by multi-step agentic control loops.

HIP's retrieval-gating decision is also related to a line of work on adaptive retrieval, which asks when a system should retrieve at all rather than retrieving unconditionally for every query. Self-RAG (Asai et al., 2024) trains a model to emit special reflection tokens that judge whether retrieval is needed and whether generated content is supported by retrieved passages; FLARE (Jiang et al., 2023) instead triggers retrieval reactively, mid-generation, based on the model's own token-level confidence; and Adaptive-RAG (Jeong et al., 2024) routes queries to no-retrieval, single-step, or multi-step retrieval pipelines using a separately trained query-complexity classifier. HIP differs from all three along two axes relevant to this paper's contribution. First, HIP's routing decision is produced by a fixed, deterministic validator operating on pre-established per-prompt criteria (evidence dependence and stability) rather than by a component trained end-to-end on model-generated reflection tokens or confidence signals, which is what allows the tiered comparison in this paper (Section 4.3) to be exactly reproducible given the same configuration. Second, and more specific to this paper's framing, HIP explicitly separates the retrieval-gating decision from refusal-boundary enforcement (Section 4.5): none of the three adaptive-retrieval systems above are evaluated, to our knowledge, on whether a targeted refusal boundary can be made provider-invariant independently of the retrieval decision itself. We see HIP's contribution here as complementary to, rather than a replacement for, learned adaptive-retrieval routing – it is a lighter-weight, fully deterministic alternative suited to settings where reproducibility and auditability of the routing decision itself are a priority.

At a high level, the HIP payload construction pipeline operates in three stages. First, a routing step determines whether a given prompt requires grounded evidence and, if so, which evidence namespace to draw from. This routing is performed by a deterministic validator that classifies each prompt along two axes – evidence dependence and stability – using a fixed, pre-established classification for each prompt rather than a general-purpose semantic analysis of the prompt text at inference time; given the same validator and configuration, routing decisions are fully reproducible, and no prompt text is modified during this step. Throughout this paper, *prompt intent* refers specifically to this deterministic routing/classification signal, and we do not claim it reflects a general-purpose semantic intent classifier. Second, if the routing step signals a grounding requirement, a retrieval step returns a bounded set of evidence passages from the selected secondary evidence domain. Third, an injection step deterministically compiles the validated HIP payload – together with any retrieved evidence passages, for Tier B – into a bounded governance context; this governance context is then combined with the original user request to form the final input passed to the underlying model for inference. This compilation step is deterministic and bounded: for a given payload (and, where applicable, retrieved evidence), it produces a fixed governance context rather than a stochastically varying one, which is what allows the system's behavior to be reproduced exactly given the same inputs. Evidence sufficiency is likewise checked deterministically: evidence is injected only when the prompt is routed as evidence-requiring, retrieval is enabled for the active condition, and the relevant secondary evidence domain is available; when these conditions are not jointly met, the system either proceeds without evidence or refuses when evidence is strictly required, rather than answering ungrounded.

The two conditions evaluated in this paper differ only in which of these stages are active. Tier A disables retrieval and supplies no secondary-domain evidence, so the model receives only the prompt and primary-domain context. Tier B runs all three stages as described above, so evidence is injected only when the routing step signals a grounding requirement. Grounding success (Table 3; see also

EXP-G1, Section 4.2) is measured via a deterministic, pattern-based detector: model outputs are scanned for citation-like retrieval markers together with primary-domain and secondary-domain lexical/concept anchors, and a per-output anchor count is computed directly from these pattern matches, rather than via a separately trained/learned classifier. This detector is not a semantic-correctness judgment.

3.1 Tiered Configuration Design

Tier A (Baseline: No Retrieval). Retrieval is disabled. The model receives the prompt and structured primary-domain context, but no external evidence from the secondary evidence domain.

Tier B (HIP with Evidence-Gated Retrieval). Evidence-gated retrieval is enabled. The model receives secondary-evidence-domain evidence only when prompt intent explicitly signals a grounding requirement.

4 Experiments and Results

4.1 Experimental Setup

Paired answerability. We evaluate paired outputs across the no-retrieval (Tier A) and evidence-gated (Tier B) conditions at full scale: 500 outputs per condition, exactly paired across 50 synthetic users $\times$ 10 prompts (Section 4.3).

Grounding-signal validation. A separate, smaller evaluation directly measures the operational grounding signal for the same two conditions: 50 no-retrieval outputs versus 50 evidence-gated outputs (Section 4.3).

EXP-G1 (generalization study). We evaluate the evidence-gated configuration across 30 synthetic profiles and 41 frozen prompts per profile, yielding 1,230 model outputs. This figure is a single total for the evidence-gated configuration, reported once rather than per condition. This is a scale/generalization study distinct from, and does not replace, the paired-answerability and grounding-signal evaluations above.

Phase-B (cross-provider enforcement suite). Phase-B evaluates ten prompts across three providers, two HIP configurations, and five repetitions, yielding 300 stateless calls. The prompt set contains four refusal-required cases (B1–B4), two JSON-output cases (B5–B6), one exact five-word case (B7), one exact-token case (B8), and two answerable-control cases (B9–B10). All calls are executed independently with no shared state, memory, or caching across runs or providers.

Together, these four evaluations span exactly 2,630 model outputs (500 + 500 for paired answerability, 50 + 50 for grounding-signal validation, 1,230 for EXP-G1, and 300 for Phase-B).

4.2 EXP-G1 Output Count

Table 1: EXP-G1 generalization study: output count for the evidence-gated configuration.

Study	Configuration	Profiles	Prompts per profile
EXP-G1	Evidence-gated	30	41

Beyond the raw output count, the following deterministic evidence-signal metrics are reported for the full EXP-G1 set (30 profiles $\times$ 41 prompts = 1,230 outputs): 100% non-empty evidence previews, 100% citation-like grounding signal, 100% primary-domain anchor signal, 100% secondary-domain anchor signal, 0% refusals, and 0% verbatim quote blocks. As with the tier comparison, these are deterministic evidence-signal measures rather than factual-correctness scores.

4.3 Paired Answerability and Grounding-Signal Validation

We evaluate paired answerability and grounding-signal outcomes for the no-retrieval (Tier A) and HIP evidence-gated retrieval (Tier B) conditions at two scales. This is distinct from EXP-G1 (Section 4.2),

which evaluates only the evidence-gated condition at larger scale; EXP-G1 expands the evidence-gated condition but does not replace the paired comparisons below.

Paired answerability ($n = 500$ per condition). Table 2 reports refusal outcomes over 500 outputs per condition, exactly paired across 50 synthetic users $\times$ 10 prompts.

Table 2: Paired answerability across the no-retrieval (Tier A) and HIP evidence-gated retrieval (Tier B) conditions ($n = 500$ per condition, exactly paired across 50 synthetic users $\times$ 10 prompts).

Metric	Tier A (no retr.)	Tier B (HIP gated)
Refusal rate (%)	0	0
Refusal 95% CI (Wilson)	[0.0, 0.8]	[0.0, 0.8]
n (paired outputs)	500	500

Grounding-signal validation ($n = 50$ per condition). A separate, smaller evaluation directly measures the operational grounding signal for the same two conditions, defined as the Frozen-41 prompt set (41 items) combined with the Evidence-Required-9 subset (9 items), for 50 aligned items per condition.

Table 3: Grounding-signal validation across the no-retrieval (Tier A) and HIP evidence-gated retrieval (Tier B) conditions, combining the Frozen-41 prompt set (41 items) and the Evidence-Required-9 subset (9 items) for $n = 50$ aligned items per condition. Each row is a deterministic evidence-signal component (presence of a citation-like retrieval marker or a domain anchor), not a human-rated semantic-correctness score.

Metric	Tier A (no retr.)	Tier B (HIP gated)
Refusal rate (%)	0	0
Refusal 95% CI (Wilson)	[0.0, 7.1]	[0.0, 7.1]
Retrieved secondary-domain evidence	No	Yes (gated)
Citation-like marker present	0/50	50/50
Secondary-domain evidence anchor present	0/50	50/50
Primary-domain anchor present	0/50	48/50
n (aligned items)	50	50

Component breakdown of the grounding validation. Within the Frozen-41 prompt set (41 items), the no-retrieval condition showed 0/41 citation-like markers, 0/41 primary-domain anchoring, and 0/41 secondary-domain anchoring; the evidence-gated condition showed 41/41 on all three measures. Within the nine-prompt Evidence-Required-9 subset, the no-retrieval condition showed 0/9 on citation-like and secondary-domain markers, and the evidence-gated condition showed 9/9 on citation-like and secondary-domain markers but 7/9 on primary-domain anchoring. Combined across both subsets ($n = 50$), this gives 0/50 versus 50/50 for citation-like markers and for secondary-domain anchoring, and 0/50 versus 48/50 for primary-domain anchoring, as reported in Table 3. Where this paper reports a headline "0% versus 100%" operational grounding signal, that headline refers specifically to the citation-like and secondary-domain-anchor components, which are the components a gated secondary-domain retrieval step is designed to produce; it does not claim 100% on primary-domain anchoring, which is a distinct component at 48/50 in the evidence-gated condition. We report the grounding metric as a deterministic evidence-signal measure specifically to avoid overstating it as a semantic-correctness judgment, which is a distinct question not evaluated in this submission.

4.4 Latency Analysis

Tier A vs. Tier B matched-condition comparison. We report end-to-end inference latency for the no-retrieval (Tier A) and HIP evidence-gated retrieval (Tier B) conditions under matched settings ($n = 40$ runs per condition). Table 4 shows that evidence-gated retrieval did not introduce a measurable latency overhead relative to the no-retrieval baseline under this configuration; the nominal 1.2 ms difference between conditions is well within run-to-run variance and is not interpreted as a speed improvement, only as an absence of measurable overhead.

Table 4: End-to-end inference latency for Tier A (no retrieval) vs. Tier B (HIP evidence-gated retrieval), $n = 40$ runs per condition. All values in milliseconds.

Condition	Mean	Std	p50	p95
Tier A (no retrieval)	242.8	10.1	241.0	246.6
Tier B (HIP evidence-gated)	241.6	3.6	241.1	249.4

The below table reports a separate measurement, *Phase-B* per-provider, per-preset latency, which must not be conflated with the matched-condition overhead comparison above.

Table 5: Phase-B provider and configuration latency (percentiles, $n = 50$ trials per row unless noted).

Provider	HIP preset	HTTP 200	p50 (ms)	p90 (ms)
OpenAI	HIP_01	50/50	771	1,434
OpenAI	HIP_03	50/50	793	1,575
Google	HIP_01	50/50	6,392	16,545
Google	HIP_03	50/50	4,654	15,931
Anthropic	HIP_01	50/50	737	1,061
Anthropic	HIP_03	49/50	751	979

Refusal enforcement was provider-invariant for the B1–B4 refusal-required prompts (Section 4.5), while end-to-end latency remained strongly provider-dependent. These values should not be collapsed into a single global "HIP latency" claim. We have not investigated why Google's latency is substantially higher than OpenAI's or Anthropic's in Table 5; this could reflect provider-side infrastructure or routing differences unrelated to HIP itself, and we do not speculate on the cause here.

4.5 Cross-Provider Refusal Enforcement (Phase-B)

To test whether strict refusal behavior can be enforced robustly across providers, we evaluate a subset of refusal-required prompts (B1–B4) under the Phase-B suite. These prompts require a strict, machine-checkable refusal output when evidence is missing or incomplete.

For B1–B4, the evaluation harness supplied a fixed refusal-required tag alongside each prompt. When enforcement was enabled, HIP deterministically mapped this tag to the canonical refusal response and suppressed numeric output, independent of the underlying model's own judgment about the prompt. We describe this transparently as cross-provider control-plane enforcement of a predefined boundary, rather than the models independently detecting that boundary, or the models' own calibrated self-knowledge of what they do or do not know (Kadavath et al., 2022; Kuhn et al., 2023). This mirrors the design of external input/output safeguard classifiers in prior work, which likewise enforce a boundary from outside the generating model rather than relying on the model's internal judgment. Because the refusal-required cases are tagged in advance and the mapping itself is deterministic by construction, the 120/120 result is best understood as a verification that this fixed policy-to-output mapping holds reliably across three heterogeneous providers and two HIP presets, rather than as a novel behavioral finding about model capability: strict refusal compliance is achieved deterministically at the output level across all tested providers and both HIP presets, with the exact refusal string produced on every refusal-required trial, zero refusal drift, and zero numeric guessing under uncertainty. This result complements the tiered grounding experiments by demonstrating that inference-time enforcement can provide hard, testable guarantees across heterogeneous providers – i.e., that the system contract remained provider-invariant across the three providers – while avoiding prompt-surface fragility such as refusal rephrasing or partial answering; it should not be read as evidence that the underlying models discovered or reasoned about the refusal boundary themselves.

Note on precision. The per-provider/preset breakdown of the refusal-required result (20/20 for each of six provider/preset combinations) has a wider individual-cell 95% CI of [83.9, 100.0]%; the pooled [96.9, 100.0]% reported for the aggregate result in Table 6 should not be read as applying uniformly to every individual cell. The five-word result (B7) has the widest interval of any reported behavior

Table 6: Phase-B outcomes by behavior group.

Behavior group	Att.	Eval.	Result	95% CI
Required refusal (B1–B4)	120	120	120/120 exact; 0 drift; 0 risky guesses	[96.9, 100.0]%
JSON dictionaries (B5–B6)	60	60	59/60 (98.3%)	[91.1, 99.7]%
Five words (B7)	30	30	15/30 (50.0%)	[33.2, 66.8]%
Exact token (B8)	30	30	30/30 (100%)	[88.6, 100.0]%
Answerable controls (B9–B10)	60	59	59/59 non-refusal; 1 provider timeout	[93.9, 100.0]%

([33.2, 66.8]%), reflecting both its low success rate and its small sample ($n = 30$); it should be read as a clear limitation rather than a borderline pass.

Interpretation. These findings support a boundary-specific view of determinism: inference-time enforcement made the targeted, high-risk refusal boundary provider-invariant without converting the system into a blanket-refusal mechanism. This guarantee did not automatically generalize to every surface-form constraint (Table 6, B7). We do not claim that all output constraints were made deterministic by this mechanism; that broader claim is contradicted by the five-word result.

Semantic quality (not evaluated). The 59 successfully completed answerable-control responses (30 operations, 29 security, per the predefined task split) establish response preservation rather than semantic correctness. Factual correctness, completeness, unsupported-claim rate, operational usefulness, and safety appropriateness for these responses were not evaluated in this submission, and no correctness or unsupported-claim rate is reported here as a result.

5 Discussion and Conclusion

This paper separates two distinct findings that should not be collapsed into one narrative. First, evidence-gated retrieval changed grounding behavior without causing refusal on aligned, evidence-dependent prompts: Tier B achieved a 100% deterministic grounding-signal rate at 0% refusal, whereas the no-retrieval condition showed a 0% grounding-signal rate at the same 0% refusal rate. Second, and separately, a targeted, high-risk refusal boundary (B1–B4) was enforced deterministically and provider-invariantly through a predefined control-plane policy, not through the retrieval mechanism itself. Grounding and answerability thus arise from distinct system-level mechanisms: evidence-gated retrieval separates grounding from answerability rather than one causing or suppressing the other, while the specific refusal boundary evaluated in Phase-B is governed by explicit policy enforcement rather than by retrieval or by the model's own judgment.

On the recurrence of 100%/0% outcomes. Several of the headline numbers in this paper (Tier B's 100% grounding-signal rate, EXP-G1's 100% evidence-signal metrics, Phase-B's 120/120 refusal enforcement) are, to varying degrees, expected consequences of how the system is constructed rather than independent discoveries about model behavior. The grounding-signal detector is deterministic and pattern-based, and is scanning for markers (citation-like text, domain anchors) that a deterministic injection mechanism is designed to produce whenever it fires; a high match rate under these conditions demonstrates that the intended pipeline behavior occurs reliably, not that the resulting text is semantically better grounded in a human-judged sense. Similarly, Phase-B's refusal-enforcement result follows from a fixed tag-to-output mapping rather than from the underlying models discovering the boundary themselves (Section 4.5). We report these figures because verifying that a deterministic pipeline behaves deterministically is itself a non-trivial systems claim worth establishing empirically across three heterogeneous providers, but readers should not read the repeated appearance of "100%" across sections as independent, compounding evidence of semantic quality or model capability. No component of this evaluation measures semantic quality; that would require a blinded, rubric-based review, which is not included in this submission.

Limitations. Grounding in the paired comparisons (Table 2, Table 3) is reported as a deterministic evidence-signal measure (citation-like markers and primary-/secondary-domain anchoring), not as a judgment of semantic correctness. The routing validator underlying both conditions operates on a fixed, pre-established per-prompt classification rather than a general-purpose intent classifier evaluated at inference time; the reported determinism therefore reflects reproducibility of a fixed

policy given known prompts, not generalization to novel, unseen prompt distributions. Comparison against unstructured, ungated context injection is left for future work and is not evaluated in this paper. Likewise, this paper does not include an empirical comparison against learned adaptive-retrieval baselines such as Self-RAG (Asai et al., 2024), FLARE (Jiang et al., 2023), or Adaptive-RAG (Jeong et al., 2024); it establishes the behavior of HIP's deterministic mechanism on its own terms rather than benchmarking it head-to-head against these learned alternatives. Within Phase-B, refusal-boundary enforcement (B1–B4) was deterministic and provider-invariant across 120 trials, but this determinism did not extend to every surface-form constraint: exact five-word compliance succeeded in only 15/30 trials and was provider-dependent. The answerable-control results (B9–B10) establish response preservation, i.e., that evidence gating did not cause over-refusal on prompts that should be answered, rather than semantic correctness of those answers; one Anthropic HIP_03 call timed out after approximately 60 seconds and was excluded from the answerable-control count (59/59). Factual correctness, completeness, unsupported-claim rate, usefulness, and safety appropriateness of the 59 successfully completed answerable-control responses were not evaluated in this submission. All reported figures reflect a single frozen evaluation run under stateless conditions; behavior under significantly larger trial counts, multi-turn interaction, or adversarial context is not evaluated here. The broader control-plane research lineage that this system is part of (covering canonical semantic state, recovery/hold behavior, paraphrase normalization, and cross-provider consistency at a system-architecture level) is deliberately out of scope for this manuscript and is reserved for a separate, follow-on paper. Taken together, the contribution here is a narrow but empirically verified claim: that answerability and grounding can be decoupled at inference time through deterministic, auditable mechanisms, without extending that determinism – or its underlying safety guarantees – further than the data support.

6 Broader Impact

This work addresses a system-design question: how to decouple answerability from grounding at inference time, rather than introducing a new model capability or a new pretrained model. We see the primary potential positive impact as reducing over-reliance on blanket refusal as a proxy for safety: by making a narrow, high-risk refusal boundary deterministic and provider-invariant (Section 4.5) without suppressing answerable requests, systems built on this pattern may be able to preserve legitimate utility while still enforcing specific, well-scoped safety constraints.

We highlight two categories of risk that follow directly from our own results rather than from speculation. First, the mechanism's determinism is boundary-specific, not universal: our own data show that an exact five-word formatting constraint held in only 15/30 trials (Section 4.5). A system operator who mistakenly assumes that evidence-gating enforcement generalizes uniformly to all output constraints could deploy an insufficiently tested safety boundary. Second, the same conditional-injection architecture that enables beneficial, scoped grounding could in principle be reconfigured to selectively suppress or inject evidence in ways that mislead rather than ground a user; auditability of the gating configuration itself, not just of model outputs, is therefore an important deployment consideration that this paper does not evaluate.

Since semantic correctness of the answerable-control responses was not evaluated in this submission (Section 5), practitioners considering HIP-style enforcement in production settings should not cite the refusal-determinism result in isolation as evidence of general-purpose safety or production readiness: it demonstrates enforcement of one specific, narrow boundary under a single frozen evaluation run, not a general safety guarantee across models, providers, or deployment conditions.

AI Use Statement

The authors used ChatGPT/Codex (OpenAI) as an assistive tool at several stages of this project, including research framing, feedback on experimental and prompt design, code generation and debugging for the HIP system and evaluation harness, and drafting and editing portions of the text. All experimental design choices, the interpretation of results, and the paper's scientific claims remain the responsibility of the human authors, who reviewed and verified all AI-assisted content.

Reproducibility Statement

We report exact sample sizes and prompt-set construction used for the paired answerability and grounding-signal comparisons (Section 4.3), the EXP-G1 profile/prompt design (Section 4.2), and the Phase-B provider/preset/repeat configuration (Section 4). Wilson score confidence intervals and the method for computing them are reported alongside all proportions (Tables 2, 3, and 6). Archived run identifiers and artifact names use internal naming conventions that are not renamed in the underlying archived artifacts, only in this manuscript's reviewer-facing prose (e.g. the evidence sets referred to here as Frozen-41 and Evidence-Required-9). The exact HIP implementation code, the specific routing-validator implementation, and the verbatim synthetic prompt sets are not being released alongside this submission, as this system is part of ongoing development work outside the scope of this submission. Appendix A instead provides a higher-level, sanitized description of the HIP payload construction and evaluation harness (Section 3) sufficient for a reader to understand the design and construct an independent implementation of the described mechanism, without exposing the underlying proprietary code or prompt text.

References

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

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, et al. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Luyu Gao, Chenyan Xiong, Jianfeng Gao, et al. RARR: Researching and revising what language models say, using language models. *ACL*, 2023a.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2023b. doi: 10.18653/v1/2023.emnlp-main.398.

Amelia Glaese, Nat McAleese, Maja Trębacz, et al. Improving alignment of dialogue agents via targeted human judgments. *arXiv preprint arXiv:2209.14375*, 2022.

Kelvin Guu, Kenton Lee, Zora Tung, et al. REALM: Retrieval augmented language model pre-training. In *ICML*, 2020.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. TRUE: Re-evaluating factual consistency evaluation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.naacl-main.287.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, 43 (2), 2025. doi: 10.1145/3703155.

Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering. In *ACL*, 2021.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong C. Park. Adaptive-RAG: Learning to adapt retrieval-augmented large language models through question complexity. In *NAACL*, 2024.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12), 2023. doi: 10.1145/3571730.

Zhengbao Jiang, Frank F. Xu, Luyu Gao, et al. Active retrieval augmented generation. In *EMNLP*, 2023.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, et al. Dense passage retrieval for open-domain question answering. In *EMNLP*, 2020.

Omar Khattab and Matei Zaharia. ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. In *SIGIR*, 2020.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations*, 2023.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *NeurIPS*, 2020.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. HaluEval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.397.

Potsawee Manakul, Adian Liusie, and Mark Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.557.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.741.

OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, et al. Toolformer: Language models can teach themselves to use tools. *NeurIPS*, 2023.

Shunyu Yao, Jeffrey Zhao, Chuanyang Yu, et al. ReAct: Synergizing reasoning and acting in language models. In *ICLR*, 2023.

Hanning Zhang, Shizhe Diao, Yong Lin, Yi Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji, and Tong Zhang. R-tuning: Instructing large language models to say 'I don't know'. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7113–7139. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.naacl-long.394.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, et al. Judging LLM-as-a-judge with MT-bench and chatbot arena. *NeurIPS Datasets and Benchmarks Track*, 2023.

A Technical appendices and supplementary material

A.1 Sanitized description of the HIP payload construction and evaluation harness

This section summarizes, at a sanitized level of detail, the mechanism described conceptually in Section 3, sufficient for a reader to understand the design and construct an independent implementation of the approach without access to the underlying proprietary code or prompt text.

Routing/intent-detection. Each prompt is passed through a deterministic validator that classifies it along two axes – evidence dependence and stability – using a fixed, pre-established classification for each prompt. We do not claim this reflects a general-purpose semantic analysis of prompt content at

inference time; the specific validator implementation is not disclosed in this submission. What the reported evidence directly supports is that routing decisions are deterministic and reproducible given the same validator and configuration, and that prompt text itself is never modified by this step. This classification step is what determines both (a) whether a prompt requires grounded evidence at all, and (b) which secondary evidence domain a retrieval step should query if so.

Retrieval. When the routing step signals a grounding requirement, the selected evidence namespace returns a bounded set of evidence passages, which are deterministically compiled with the HIP payload before inference. No retrieval is performed for prompts the routing step does not flag as evidence-requiring, or when Tier A conditions are active (Section 3).

Injection/compilation. Retrieved evidence (if any) is combined with the validated HIP payload through a deterministic, bounded compilation step that produces a fixed governance context; this context is combined with the original user request to form the model input. The same payload and retrieved evidence always compile to the same governance context, which is what makes the tiered comparison in this paper exactly reproducible given the same inputs and configuration.

Evidence sufficiency and refusal. Evidence is injected only when three conditions jointly hold: the prompt is routed as evidence-requiring, retrieval is enabled for the active condition, and the relevant secondary evidence domain is available. When these conditions are not jointly met, the system either proceeds without evidence or refuses, depending on whether evidence is strictly required for the prompt in question; this refusal behavior is deterministic given the same inputs and configuration. Separately, for the Phase-B B1–B4 refusal-required prompts specifically, the evaluation harness supplies a fixed refusal-required tag directly, and enforcement deterministically maps this tag to the canonical refusal output (Section 4.5); this is a control-plane enforcement mechanism operating on a predefined tag, not a model-side judgment of the prompt.

Grounding measurement. Grounding success (Table 3; see also EXP-G1, Section 4.2) is scored by a deterministic, pattern-based detector that scans model outputs for citation-like retrieval markers and primary-/secondary-domain conceptual anchoring, and computes a per-output anchor count from these pattern matches. This is a rule-based measure rather than a separately trained/learned classifier, which is why it is reported as a deterministic evidence-signal measure rather than a semantic-correctness judgment (Section 4.5).

Raw logs, verbatim prompt sets, and the underlying implementation code are not being released with this submission, as the system is part of ongoing development work outside the scope of this submission.