Facet-Level Tracing of Evidence Uncertainty and Hallucination in RAG

Anonymous ACL submission

Abstract

Retrieval-Augmented Generation (RAG) aims to reduce hallucination by grounding answers in retrieved evidence, yet hallucinated answers remain common even when relevant documents are available. Existing evaluations focus on answer- or passage-level accuracy, offering limited insight into how evidence is used during generation. In this work, we introduce a facet-level diagnostic framework for QA that decomposes each input question into atomic reasoning facets. For each facet, we assess evidence sufficiency and grounding using a structured Facet×Chunk matrix that combines retrieval relevance with natural language inference–based faithfulness scores. To diagnose evidence usage, we analyze three controlled inference modes: Strict RAG, which enforces exclusive reliance on retrieved evidence; Soft RAG, which allows integration of retrieved evidence and parametric knowledge; and LLM-only generation without retrieval. Comparing these modes enables fine-grained analysis of retrieval–generation misalignment, defined as cases where relevant evidence is retrieved but not correctly integrated during generation. Across medical QA and HotpotQA, we evaluate three open- and closed-source LLMs (GPT, Gemini, and LLaMA), providing interpretable diagnostics that reveal recurring facet-level failure modes, including evidence absence, misalignment, and prior-driven override. Our results show that hallucinations in RAG systems are often associated with generation-side evidence usage patterns rather than retrieval accuracy alone, with facet-level analysis exposing override and misalignment behaviors that remain hidden under answer-level evaluation.

1 Introduction

RAG is widely used to improve the factual reliability of LLM by grounding generation in external evidence (Lewis et al., 2020; Guu et al., 2020). By conditioning answers on retrieved documents, RAG aims to reduce hallucinations, particularly in knowledge-intensive tasks such as question answering. Nevertheless, hallucinated or unsupported content remains common, even when seemingly relevant evidence is retrieved (Min et al., 2023; Li et al., 2023).

A prevailing assumption in RAG is that stronger retrieval leads to more faithful generation. Accordingly, prior work has focused on improving retrieval quality and evaluating systems using retrieval-centric metrics or answer-level accuracy (Asai et al., 2024). In practice, however, retrieval confidence alone often fails to guarantee faithful evidence usage, especially for complex reasoning questions (Du et al., 2024). Models may retrieve highly relevant passages yet ignore them, partially use them, or override them with parametric knowledge, behaviors that are difficult to diagnose using standard evaluation methods (Chen et al., 2022). Yet this relationship remains poorly understood (Djeddal et al., 2024; Cuconasu et al., 2024).

Does improving retrieval quality actually result in more faithful answers in RAG systems?

We evaluate our framework on multi-hop question answering using HotpotQA (Yang et al., 2018) and a medical question answering dataset (Friel et al., 2024) derived from clinical guidelines, using three LLMs: GPT-4o-mini,[1] LLaMA-3-8B-Instruct,[2] and Gemini-2.0-Flash.[3] Our contributions include: (1) a facet-level diagnostic framework that decomposes questions into atomic reasoning steps and constructs structured evidence matrices to trace retrieval-generation interactions, (2) a fine-grained evidence taxonomy that categorizes five failure modes including evidence absence, misalignment, and prior-driven overrides, and (3)

[1] GPT-4o mini is accessed via the OpenAI API.

[2] LLaMA-3-8B-Instruct is an open-weight model released by Meta.

[3] Gemini-2.0-Flash is accessed via the Google Generative AI API.

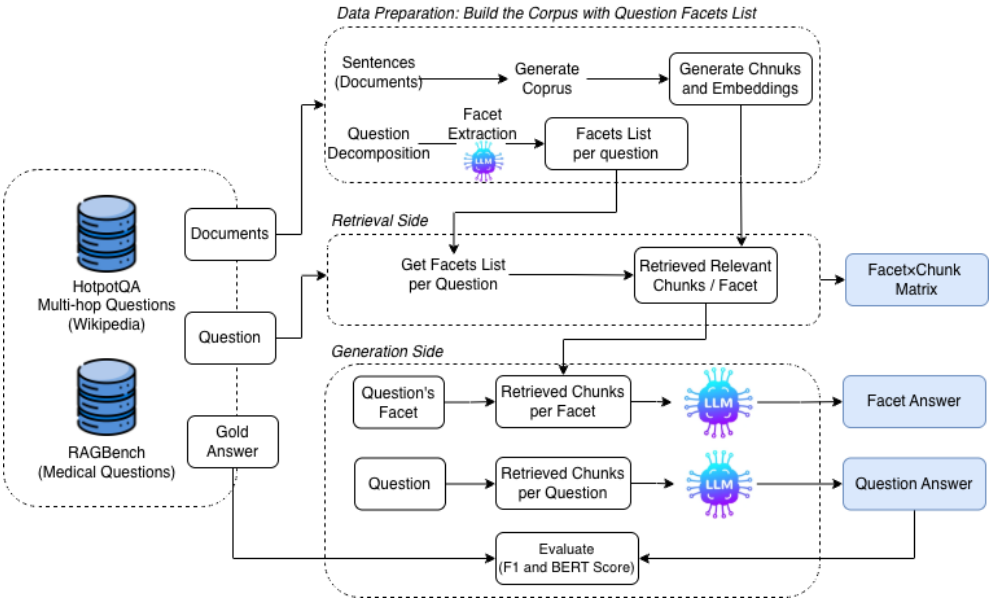

Figure 1: Facet-RAG Pipeline. The framework decomposes questions into reasoning facets, retrieves evidence at the chunk level, and evaluates facet-level entailment and generation faithfulness to diagnose hallucination and uncertainty in RAG systems.

controlled evaluation across models, revealing that hallucinations are often linked to generation-side evidence misalignment rather than retrieval failure. We further validate the robustness of our analysis through sensitivity checks and manual inspection of facet decomposition quality.

2 Related Work

RAG systems combine neural retrieval with conditional generation to ground model outputs in external knowledge (Lewis et al., 2020; Guu et al., 2020). Prior work has focused on improving retrieval quality through dense retrievers (Karpukhin et al., 2020) and iterative strategies (Jiang et al., 2023; Trivedi et al., 2023). While adaptive retrieval methods (Shi et al., 2024) dynamically adjust retrieval behavior, they do not explicitly analyze why models may fail to utilize retrieved evidence even when it is relevant and sufficient.

Hallucination detection has been studied at the answer level (Maynez et al., 2020; Li et al., 2023), with atomic fact decomposition (Min et al., 2023) and self-consistency techniques (Manakul et al., 2023) for evaluation. Multi-hop question answering work explores chain-of-thought prompting (Yao et al., 2023) and decomposition-based reasoning (Press et al., 2023) to improve accuracy.

Uncertainty estimation in RAG has focused on confidence calibration (Xiong et al., 2024), and semantic uncertainty estimation (Kuhn et al., 2023) which focus the study on semantic uncertainty through token-level entropy measures, while (Lin et al., 2023) propose uncertainty-aware decoding strategies.

3 Methodology

Our goal is to diagnose evidence uncertainty and hallucination behavior in RAG systems. Rather than proposing new retrievers or generators, our methodology: We analyze how retrieval and generation interact under controlled inference conditions to expose systematic retrieval–generation misalignment as shown in Figure 1. Given a question and its retrieved documents, we decompose the question into atomic reasoning facets, evaluate evidence support for each facet, and analyze how evidence usage changes across inference modes. A full pipeline example is provided in Appendix A.

3.1 Datasets

We evaluate our framework on two complementary question answering datasets designed to probe

Facet	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	Max
1) What is the type of place associated with Regency Road, Adelaide?	0.00	0.14	0.00	0.27	0.18	0.04	0.06	0.01	0.02	**0.47**	**0.47**
2) What is the type of place associated with Klemzig, South Australia?	0.14	0.01	0.00	0.21	0.00	0.14	0.01	**0.30**	0.14	0.03	**0.30**

Table 1: HotpotQA Example: Facet-level coverage over retrieved chunks (C1–C10). Main question: *"What type of place does Regency Road, Adelaide and Klemzig, South Australia have in common?"* Question type: *bridge*. Difficulty level: *hard*.

different aspects of evidence usage in RAG.

3.1.1 HotpotQA

HotpotQA (Yang et al., 2018) is a multi-hop question answering benchmark constructed from Wikipedia. We use a stratified subset of 3,000 questions across three difficulty levels (easy, medium, hard) to analyze evidence integration under increasing reasoning complexity.

3.1.2 RAGBench–Medical

RAGBench–Medical (Friel et al., 2024) is a medical question-answering dataset derived from clinical guidelines and authoritative texts. It contains 1,500 questions spanning fact retrieval and complex reasoning.

3.2 Facet Decomposition

We decompose each question into a set of *facets*, as shown in Table 1, where each facet represents a minimal semantic unit required for answering the question. Facets are generated as a preprocessing step and held fixed across all experiments to ensure consistent evaluation (Figure 1). For a question q, facet decomposition produces a set:

$$\mathcal{F}(q) = \{f_1, f_2, \ldots, f_m\},$$

where each facet f_i is a standalone question whose answer can be independently validated against retrieved evidence.

For HotpotQA, we use the dataset's supporting facts—gold sentences forming the reasoning chain—and convert each into an interrogative facet following prior work (Petcu et al., 2026; Fu et al., 2021). For the medical dataset, which lacks annotated supporting facts, we use GPT-4o-mini to decompose each question into facets. This yields 11,007 facets for Medical QA (1,500 questions) and 24,039 for HotpotQA (3,000 questions). Each facet is labeled with a semantic type (e.g., temporal, comparative, procedural) for analysis of evidence usage. See Appendix A and Appendix B for examples and prompts. The prompt below is used for medical question facet decomposition.

Facet Decomposition Validation. We manually reviewed 100 sampled questions for correctness, completeness, and granularity. 85–90% of decompositions are appropriate, with 1–3% showing substantive issues. Minor granularity differences do not affect taxonomy assignments. Additional annotation guidelines and validation details are provided in Appendix E.

Facet Decomposition Prompt

Task: *Given the following medical question, decompose it into a list of atomic sub-questions, where each sub-question addresses a single clinical fact or reasoning step required to answer the original question. Each sub-question should be answerable independently using medical knowledge sources.*
Question: `[INPUT_QUESTION]`
Output format: *Numbered list of sub-questions.*

3.3 Facet×Chunk Evidence Matrix

Given a facet set $\mathcal{F}(q)$ and retrieved chunks $\mathcal{C} = \{c_1, \ldots, c_K\}$, we construct a Facet×Chunk matrix that captures facet level coverage over retrieved chunks as shown in Table 1. For each facet–chunk pair (f_i, c_j), we compute the retrieval relevance score $s(f_i, c_j)$ using cosine similarity over BGE-base-en-v1.5 embeddings. To characterize retrieval uncertainty, we define the retrieval margin:

$$U_{\text{ret}}(f_i) = s(f_i, c_j^*) - s(f_i, c_j^{**}),$$

where c_j^* and c_j^{**} are the top-1 and top-2 retrieved chunks for facet f_i. Small margins indicate ambiguous or conflicting retrieval which can cause hallucination in generation side.

To assess whether a generated facet answer a_i is supported by a retrieved chunk c_j, we apply a natural language inference (NLI) RoBERTa-large-MNLI model (Liu et al., 2019) and compute:

$$\textit{faith}(a_i, c_j) = P(\text{entail}) - P(\text{contradict}),$$

yielding a signed grounding score. Positive values indicate evidential support, while negative values

Model	Strict RAG			Soft RAG			LLM-only		
	Answer	F1	BERT	Answer	F1	BERT	Answer	F1	BERT
Gemini-2.0-Flash	Bosnian	0.00	0.00	Bosnian	0.00	0.13	Bosnian	0.00	0.13
GPT-4o mini	NO_ANSWER	0.00	0.00	Bosnian	0.00	0.13	**Serb**	1.00	1.00
Llama-3-8B-Instruct	NO_ANSWER	0.00	0.00	Bosnian	0.00	0.13	The Grand Vizier...	0.00	0.11

Table 2: HotpotQA Example: Generated answers and corresponding F1 and BERTScore for each inference mode. Question: "Of what national descent was the Grand Vizier of the Ottoman Empire whose title indicated he had been castrated?" (Gold answer: *Serb*). Only GPT in the LLM-only setting produces the correct answer, while Strict and Soft RAG suppress correct parametric knowledge across models.

indicate contradiction. These retrieval and grounding scores populate the Facet×Chunk matrix and serve as the basis for all subsequent analysis.

For evidence retrieval, we use a dense retriever based on BGE-base-en-v1.5 embeddings (Xiao et al., 2024), implemented via SentenceTransformers. Source documents are segmented into chunks of 300 tokens with a 50-token overlap. For each facet query, we retrieve the top-$K = 5$ most similar chunks, balancing recall and precision, as smaller K risks missing relevant evidence while larger K introduces irrelevant or conflicting context (Yu et al., 2024). Table 1 shows the Facet×Chunk matrix for a representative question; Appendix A provides full matrix construction details, including retrieved chunk content and similarity scores.

3.4 Controlled Inference Modes

To disentangle retrieval failures from generation-level hallucinations, we evaluate three controlled inference modes for each facet:

1) Strict RAG: The model must answer using only retrieved evidence, returning `NO_ANSWER` if unsupported:

$$\text{Prompt}_{\text{strict}}(f_i) = \begin{cases} \textbf{EVIDENCE: } c_{i1} \oplus c_{i2} \oplus \cdots \oplus c_{iK} \\ \textbf{QUESTION: } f_i \\ \textbf{INSTRUCTION: } \textit{EvidenceOnly} \end{cases}$$

2) Soft RAG: The model may mix evidence with prior knowledge:

$$\text{Prompt}_{\text{soft}}(f_i) = \begin{cases} \textbf{EVIDENCE: } c_{i1} \oplus \cdots \oplus c_{iK} \\ \textbf{QUESTION: } f_i \end{cases}$$

3) LLM-only: Generation relies solely on parametric model knowledge without retrieval:

$$\text{Prompt}_{\text{lm}}(f_i) = \begin{cases} \textbf{QUESTION: } f_i \end{cases}$$

These modes distinguish retrieval insufficiency, misleading evidence, and prior-driven reasoning. Appendix A provides an example showing how their comparison enables diagnostic attribution of hallucination sources. All LLM prompts are in Appendix B.

3.5 Facet-Level Evidence Taxonomy

To understand how retrieval and generation contribute to hallucination, we develop a *Facet-Level Evidence Taxonomy* that classifies each reasoning facet based on evidence availability and usage patterns. For each model, category percentages are computed as the proportion of facets assigned to that category out of the total number of facets in the dataset.

For a facet f_i with retrieved evidence chunks $C_i = \{c_{i1}, \ldots, c_{iK}\}$ and generated answer a_i, we assess whether evidence is retrieved ($R(c_{ij}, f_i) \in \{0, 1\}$) and whether the generated answer is grounded in that evidence. For each answer-chunk pair, we compute an NLI grounding score using RoBERTa-large-MNLI (Liu et al., 2019), where positive values indicate supportive evidence and negative values indicate contradiction.

Each facet is classified by comparing the success/failure patterns across the three controlled inference modes described in Section 3.4. Using this framework, we define the following evidence taxonomy categories:

- **Evidence Failure:** Retrieval finds no relevant information for the facet. Strict RAG returns `NO_ANSWER`.
- **Evidence Misalignment:** Retrieved chunks are topically related but contradict or fail to support the generated facet answer.
- **Evidence Overridden:** Relevant evidence is retrieved, but the LLM ignores it and relies on parametric knowledge instead. Strict RAG fails, while Soft RAG or LLM-only succeeds, indicating evidence override during answer generation.
- **Evidence Helpful:** Retrieved evidence corrects LLM errors. LLM-only fails, but Strict RAG succeeds, demonstrating that retrieval reduces hallucination.
- **Evidence Robust:** All three inference modes produce correct and consistent answers. The

Models	Evidence Taxonomy				
	Failure	Overridden	Misalignment	Helpful	Robust
Gemini-2.0 Flash	20.6	12.5	22.3	9.6	35.1
GPT-4o mini	0.2	34.4	0.2	51.6	13.7
Llama-3-8B-Instruct	0.1	38.2	0.2	53.1	8.3
Overall	7.0	28.4	7.5	38.1	19.0

Table 3: Medical Dataset: Evidence Taxonomy Distribution Across Models. Percentage of medical reasoning facets (N=11,007) in each evidence category. Overall Evidence Helpful (38.1%) demonstrates that domain-specific medical retrieval frequently corrects model errors. Gemini exhibits balanced behavior across failure and override, while GPT and LLaMA demonstrate override-dominant patterns.

Question Type	Model	Strict RAG		Soft RAG		LLM-only	
		F1	BERT	F1	BERT	F1	BERT
Fact Retrieval	Gemini-2.0-Flash	0.222	0.093	**0.376**	**0.375**	**0.337**	0.351
	GPT-4o mini	0.185	0.008	0.232	0.139	0.186	0.101
	Llama-3-8B-Instruct	**0.291**	**0.231**	0.351	0.317	0.309	0.251
Complex Reasoning	Gemini-2.0-flash	0.176	0.034	0.357	0.360	0.340	**0.337**
	GPT-4o mini	0.089	0.025	0.161	0.116	0.127	0.087
	Llama-3-8B-Instruct	0.226	0.175	0.306	0.298	0.306	0.296

Table 4: Medical Dataset: Mean F1 and BERTScore by question type and inference mode. In medical dataset, Soft RAG outperforms Strict RAG across models, especially for Complex Reasoning

facet answer remains stable regardless of evidence availability.

All taxonomy assignments are performed independently at the facet level, enabling aggregation of facet-level evidence patterns to analyze question-level behavior in subsequent sections. Appendix A provides a detailed example of the *Evidence Overridden* pattern.

Threshold Sensitivity Analysis. We vary Soft-RAG faithfulness thresholds ($f_t^{low} \in [0.25, 0.35]$, $f_t^{high} \in [0.45, 0.55]$). On HotpotQA, Evidence Override shifts by only 1–3% absolute and Failure shows negligible variation ($\leq$0.12%). On Medical, GPT and LLaMA show Override within 36.0–40.1% ($\leq$4.1% drift) with Failure stable at 0.14%; Gemini's higher failure rate (20.6%) reflects retrieval characteristics independent of threshold choice. Taxonomy distributions in Tables 3 and 5 are thus robust to reasonable threshold variations.

3.6 Question-Level Aggregation

While our primary analysis operates at the facet level, users interact with final answers. We therefore aggregate facet-level outputs into a single question-level answer under each inference mode.

Question-level aggregation proceeds as follows: For each inference mode, we collect the generated facet answers $\{a_1, a_2, \ldots, a_m\}$ corresponding to facets $\{f_1, f_2, \ldots, f_m\}$. We then prompt the same LLM to synthesize these facet answers into a final question-level answer. The complete aggregation prompt is provided in Appendix B.

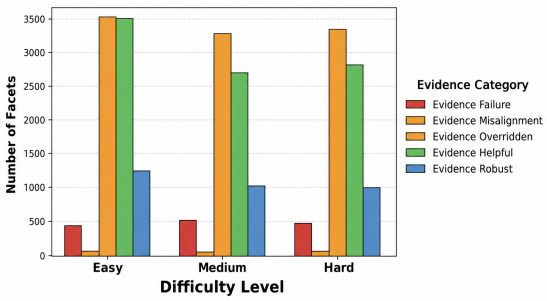

Figure 2: HotpotQA: Evidence Taxonomy Distribution by Difficulty Level. Override and Helpful are balanced on easy questions, but Helpful declines on harder questions while Override stays stable. Failure and Misalignment remain constant and minimal, while Evidence Robust decreases with difficulty, indicating complexity disrupts generation rather than retrieval.

The final answer is evaluated using token-level F1 (Rajpurkar et al., 2016) and BERTScore (Zhang et al., 2020), standard metrics for question answering (Roh et al., 2025). F1 measures token overlap between predicted and ground truth answers, balancing precision and recall. BERTScore captures semantic similarity via contextualized embeddings, addressing limitations of lexical overlap. These metrics contextualize question-level answer quality; hallucination analysis is based on

the facet-level taxonomy derived from NLI grounding scores and does not depend on BERTScore. Table 2 shows question-level answers across inference modes. Strict RAG returns NO_ANSWER, Soft RAG produces an incorrect answer contradicting retrieved evidence, and LLM-only generates the correct answer using parametric knowledge. These metrics contextualize how facet-level evidence behaviors translate into end-to-end answer quality rather than defining hallucination. All evaluations use identical inference settings across Gemini, GPT, and LLaMA (Appendix C).

4 Experimental Results

We analyze hallucination behavior in RAG across three models (Gemini, GPT, LLaMA) on two datasets Medical QA and HotpotQA. Our analysis reveals a weak and inconsistent relationship between retrieval quality and generation faithfulness: high-quality retrieval frequently co-occurs with unfaithful generation, while faithful answers arise even under imperfect retrieval. This decoupling motivates facet-level analysis of how retrieved evidence is actually used during generation.

4.1 Evidence Taxonomy Results

Table 3 presents the distribution of medical reasoning facets across five evidence categories, revealing structured failure patterns rather than random hallucination. At the aggregate level, Evidence Override is more prevalent than Evidence Failure (28.4% vs. 7.0%), suggesting that the main challenge often lies not in finding evidence but in ensuring that models use it during generation. Evidence Helpful (38.1%) further shows that domain-specific retrieval frequently corrects model errors, while Evidence Robust (19.0%) indicates stable evidence use across inference modes. Together, Helpful and Robust account for 57.1% of facets.

LLM-specific patterns diverge substantially. Gemini exhibits balanced behavior with 20.6% failure and 12.5% override, alongside highest robustness (35.1%). In contrast, GPT and LLaMA demonstrate override-dominant patterns (34.4% and 38.2% respectively) with near-zero failure rates (0.2% and 0.1%), indicating strong retrieval but systematic parametric interference that undermines evidence usage. To validate cross-domain generalization, we analyze HotpotQA (Table 5). *Evidence Override remains dominant at 42.3%*, now exceeding medical domain's 28.4%. This increase

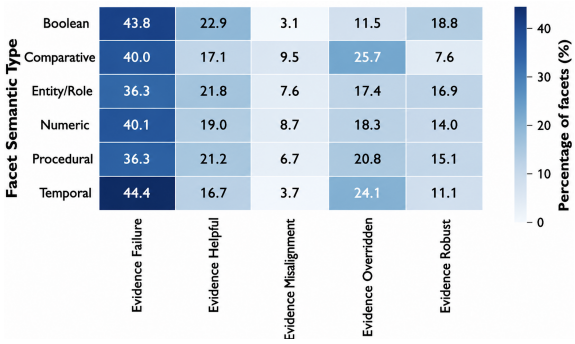

Figure 3: Medical Dataset: Facet semantic type × evidence taxonomy distribution. Boolean and Temporal facets show highest failure rates. Comparative facets are most unstable with highest misalignment and lowest robust rates.

occurs because HotpotQA's general-domain content has high coverage in LLM pretraining corpora, strengthening parametric priors that override retrieved evidence. Evidence Failure drops to 5.8% (versus 7.0% for medical), while Evidence Misalignment nearly vanishes (0.7% versus 7.5% medical), confirming Wikipedia's broad coverage provides better retrieval recall. The consistent override-to-failure gap across datasets suggests that retrieval–generation misalignment is a recurring property of current RAG behavior rather than a domain-specific artifact.

4.2 Facet-Level Evidence Patterns

Figure 2 shows that, as difficulty increases in HotpotQA, Evidence Helpful declines while Override remains stable at roughly 40–43%. In contrast, Evidence Failure stays near 6% and Misalignment remains below 2% across all levels, suggesting that increasing complexity mainly affects evidence integration during generation rather than retrieval coverage.

Figures 3 and 4 examine evidence patterns across different facet semantic types (comparison, numerical, procedural, temporal, ...), revealing domain-specific retrieval but consistent generation behavior. Failure rates vary dramatically: Boolean and Temporal facets fail most in medical QA (43.8%, 44.4%) but least in HotpotQA (6.0%, 9.9%). Override rates nearly double from medical to HotpotQA, indicating stronger parametric priors for general knowledge. However, Comparative facets remain most unstable in both datasets (highest misalignment 9%, lowest robust 7%). Critically, misalignment stays below 10% across all types in both datasets, confirming generation-side override dominates over retrieval-side contradic-

Models	Evidence Taxonomy				
	Failure	Overridden	Misalignment	Helpful	Robust
Gemini-2.0 Flash	14.7	32.1	0.7	34.5	18.0
GPT-4o mini	2.8	43.6	0.7	37.3	15.6
Llama-3-8B-Instruct	2.0	47.1	2.7	41.0	7.2
Overall	5.8	42.3	0.7	37.6	13.6

Table 5: HotpotQA: Evidence Taxonomy Distribution Across Models. Distribution of reasoning facets (N=24,039) by evidence behavior. Evidence Override (42.3%) exceeds Failure (5.8%) by 7×, confirming generation-side grounding challenges persist across domains.

Question Difficulty	Model	Strict RAG		Soft RAG		LLM-only	
		F1	BERT	F1	BERT	F1	BERT
Easy	Gemini-2.0-Flash	**0.550**	0.443	0.624	0.539	0.564	0.544
	GPT-4o mini	0.519	**0.447**	0.613	0.558	0.498	0.477
	LLaMA-3-8B-Instruct	0.196	0.026	0.315	0.174	0.230	0.108
Medium	Gemini-2.0-Flash	0.528	0.429	0.687	0.624	**0.582**	**0.623**
	GPT-4o mini	0.501	0.389	**0.699**	**0.631**	0.504	0.518
	LLaMA-3-8B-Instruct	0.196	0.033	0.351	0.236	0.182	0.053
Hard	Gemini-2.0-Flash	0.473	0.349	0.579	0.507	0.515	0.516
	GPT-4o mini	0.411	0.300	0.609	0.549	0.427	0.431
	LLaMA-3-8B-Instruct	0.149	0.131	0.296	0.171	0.195	0.059

Table 6: HotpotQA: Question-Level by Difficulty. Mean F1 and BERTScore by difficulty and inference mode. Soft RAG outperforms Strict RAG, and performance peaks at medium difficulty.

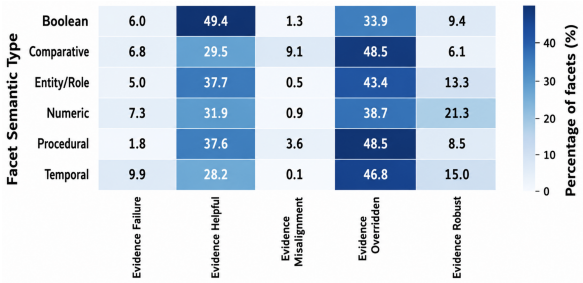

Figure 4: HotpotQA: Facet Semantic Type × Evidence Taxonomy Distribution. Boolean and Temporal facets show lowest failure rates. Override rates are consistently high across all types. Comparative facets remain most unstable in both datasets.

tions. Faithfulness Analysis: Evidence Helpful and Robust achieve high faithfulness (median > 0.7), while Override shows bimodal behavior spanning from high support to contradiction. This explains why retrieval confidence cannot predict generation quality: models unpredictably either incorporate or contradict retrieved evidence. Detailed distributions in Appendix D Figure 6.

4.3 Question-Level Performance

Table 4 presents mean F1 and BERTScore across inference modes on medical QA. We define *Complex Reasoning* as questions requiring multi-step inference across multiple facts or dependencies, in contrast to *Fact Retrieval*, which involves direct lookup of a single piece of information. Soft RAG substantially outperforms Strict RAG, with gains varying by model and question type. Complex reasoning benefits most: Gemini +103% (0.176 → 0.357 F1), GPT +81% (0.089 → 0.161 F1), and LLaMA +35% (0.226 → 0.306 F1). Fact retrieval shows more moderate gains across models.

Table 6 shows question-level results on HotpotQA stratified by difficulty. Soft RAG consistently outperforms Strict RAG across all difficulty levels: easy questions show 13–61% gains, medium questions show 30–79% gains, and hard questions show 22–99% gains. *Medium-difficulty questions achieve the highest performance across models, indicating that question decomposition and evidence integration are most effective at this level.*

Table 7 presents results by HotpotQA question meta-type. Soft RAG outperforms Strict RAG across both Bridge and Comparison questions. Bridge questions, which require connecting information across documents, show particularly strong gains for GPT (0.483 → 0.655 F1, +36%). Comparison questions exhibit more balanced performance, with LLM-only mode approaching Soft RAG performance (Gemini: 0.609 vs. 0.614 F1), suggesting that comparative reasoning may be well-captured in parametric knowledge.

Findings: LLM-only mode frequently exceeds

Question Type	Model	Strict RAG		Soft RAG		LLM-only	
		F1	BERT	F1	BERT	F1	BERT
Bridge	Gemini-2.0-Flash	**0.561**	**0.447**	0.631	0.553	0.476	0.483
	GPT-4o mini	0.483	0.350	**0.655**	**0.605**	0.375	0.374
	LLaMA-3-8B-Instruct	0.222	0.135	0.369	0.240	0.170	0.416
Comparison	Gemini-2.0-Flash	0.472	0.377	0.614	0.547	**0.609**	**0.614**
	GPT-4o mini	0.470	0.417	0.603	0.544	0.561	0.563
	LLaMA-3-8B-Instruct	0.119	0.137	0.234	0.099	0.231	0.105

Table 7: HotpotQA: Question-level answer quality by question meta-type. Mean F1 and BERTScore are reported across inference modes.

Strict RAG in both datasets (Medical: Gemini 0.340 vs. 0.176 F1 for complex reasoning; HotpotQA: Gemini 0.515 vs. 0.473 F1 for hard questions), underscoring that performance differences are driven by evidence utilization rather than knowledge availability. However, aggregate metrics mask substantial per-question variability: Appendix D shows that approximately 30% of questions degrade under Soft RAG, highlighting the lack of robust mechanisms for deciding when to rely on retrieved evidence versus parametric knowledge.

4.4 Summary

Our results demonstrate that hallucination in RAG is not solely a consequence of retrieval failure, but is frequently associated with retrieval–generation misalignment. Evidence Override emerges as the dominant pattern in the aggregate (28.4% medical, 42.3% HotpotQA), exceeding Evidence Failure by 4–7× (7.0% medical, 5.8% HotpotQA). *This indicates that the primary source of error often lies in generation-side grounding: relevant evidence is successfully retrieved but not consistently utilized during generation.* However, model-specific variations exist. In particular, for Gemini on the Medical dataset, Evidence Failure exceeds Override, highlighting differences in evidence-use behavior across model families. Question-level performance supports this interpretation: Soft RAG improves F1 by allowing evidence flexibility, while strict grounding falls below the no-retrieval baseline in 30% of cases. These patterns persist across domains, question structures, and difficulty levels, highlighting retrieval–generation misalignment as a key factor in RAG hallucination.

5 Conclusion

Our key finding overturns a foundational assumption: better retrieval does not guarantee better answers generation. Across medical QA and HotpotQA datasets, models consistently override retrieved evidence far more often than retrieval fails. This pattern holds across three LLM families (GPT, Gemini, and LLaMA), multiple difficulty levels, and distinct knowledge domains. Strict evidence grounding often underperforms generation without retrieval entirely, revealing that current RAG architectures lack mechanisms to balance external evidence against parametric knowledge. By decomposing reasoning into atomic facets and tracing evidence usage at each step, our facet-level diagnostic framework provides an interpretable view of retrieval–generation behavior.

This positions facet-level analysis as a first step toward explainable and uncertainty-aware RAG systems, enabling future models to make evidence integration decisions more transparent, adaptive, and controllable.

6 Limitations

This work is diagnostic in nature and does not propose new retrieval or generation models. While our framework enables fine-grained analysis of retrieval-generation misalignment, it relies on facet decomposition and NLI model quality, which may introduce noise or bias. Our experiments are limited to two datasets (medical QA and HotpotQA) and three LLMs; observed patterns may differ for other tasks, languages, or architectures.

Additionally, our analysis is limited to offline evaluation and does not study interactive or user-facing RAG systems. We view this as a first step toward explainable RAG systems: by making evidence usage explicit at the facet level, our framework provides transparent diagnostic signals. Future work can build on this foundation to develop interactive, uncertainty-aware RAG systems that adaptively balance external evidence with parametric knowledge.

7 Ethical considerations

Datasets and Data Usage. This work uses publicly available datasets: HotpotQA (Yang et al., 2018) and RAGBench–Medical (Friel et al., 2024). The medical dataset is derived from published clinical guidelines and authoritative medical texts, not patient data. No personally identifiable information or protected health information was used. All datasets were accessed through their official repositories and used in accordance with their licenses. We used generative AI tools to assist with language polishing and clarity of presentation. All scientific content, experimental design, analysis, and conclusions were developed and verified by the authors.

Model Access and Reproducibility. Our evaluation uses three LLMs: GPT-4o-mini (via OpenAI API), Gemini-2.0-Flash (via Google Generative AI API), and LLaMA-3-8B-Instruct (open-weight model). Access to commercial APIs requires appropriate subscriptions. We provide detailed model configurations C and prompt templates B to support reproducibility. Our diagnostic framework does not train new models and therefore does not introduce additional environmental costs beyond inference. The full pipeline code, prompts, and facet-level annotations will be released upon acceptance.

Broader Impacts. By exposing systematic retrieval-generation misalignment, this work may benefit developers of RAG systems in reducing hallucination risks. However, improved hallucination detection does not eliminate all risks of LLM deployment. Our findings that models frequently override retrieved evidence have implications for trust calibration in human-AI systems: users may over-rely on RAG systems assuming retrieval guarantees faithfulness. We encourage practitioners to implement appropriate uncertainty quantification and human-in-the-loop verification, especially in high-stakes applications.

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avi Sil, and Hannaneh Hajishirzi. 2024. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *International conference on learning representations*, volume 2024, pages 9112–9141.

Hung-Ting Chen, Michael Zhang, and Eunsol Choi. 2022. Rich knowledge sources bring complex knowledge conflicts: Recalibrating models to reflect conflicting evidence. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2292–2307.

Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. 2024. The power of noise: Redefining retrieval for rag systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 719–729.

Hanane Djeddal, Pierre Erbacher, Raouf Toukal, Laure Soulier, Karen Pinel-Sauvagnat, Sophia Katrenko, and Lynda Tamine. 2024. An evaluation framework for attributed information retrieval using large language models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 5354–5359.

Kevin Du, Vésteinn Snæbjarnarson, Niklas Stoehr, Jennifer White, Aaron Schein, and Ryan Cotterell. 2024. Context versus prior knowledge in language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13211–13235.

Robert Friel, Masha Belyi, and Atindriyo Sanyal. 2024. Ragbench: Explainable benchmark for retrieval-augmented generation systems. *arXiv preprint arXiv:2407.11005*.

Ruiliu Fu, Han Wang, Xuejun Zhang, Jun Zhou, and Yonghong Yan. 2021. Decomposing complex questions makes multi-hop qa easier and more interpretable. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 169–180.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR.

Zhengbao Jiang, Frank F Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 7969–7992.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 6769–6781.

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

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances*

in neural information processing systems, 33:9459–9474.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. Halueval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 6449–6464.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2023. Generating with confidence: Uncertainty quantification for black-box large language models. *Trans. Mach. Learn. Res.*, 2024.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. 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*, pages 9004–9017.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 1906–1919.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. 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*, pages 12076–12100.

Roxana Petcu, Kenton Murray, Daniel Khashabi, Evangelos Kanoulas, Maarten de Rijke, Dawn Lawrie, and Kevin Duh. 2026. Query decomposition for rag: Balancing exploration-exploitation. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6857–6871.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A Smith, and Mike Lewis. 2023. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5687–5711.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 2383–2392.

Keonwoo Roh, Yeong-Joon Ju, and Seong-Whan Lee. 2025. Xlqa: A benchmark for locale-aware multilingual open-domain question answering. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 28797–28809.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Richard James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2024. Replug: Retrieval-augmented black-box language models. 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 8371–8384.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 10014–10037.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, pages 641–649.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. In *International Conference on Learning Representations*, volume 2024, pages 23650–23678.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 2369–2380.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822.

Yue Yu, Wei Ping, Zihan Liu, Boxin Wang, Jiaxuan You, Chao Zhang, Mohammad Shoeybi, and Bryan Catanzaro. 2024. Rankrag: Unifying context ranking with retrieval-augmented generation in llms. *Advances in Neural Information Processing Systems*, 37:121156–121184.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

A Illustrative Facet-Level Diagnostic Example

This section illustrates our facet-level diagnostic framework using a HotpotQA example. Despite

retrieving relevant evidence with high similarity scores, the RAG system could not faithfully connect the Oklahoma player in the 1951 Sugar Bowl with the Heisman Trophy winner.

This demonstrates the "Evidence Overridden" category: Strict RAG failed completely, Soft RAG bridged the gap using parametric knowledge, and LLM-only succeeded with prior knowledge alone—revealing that retrieval quality alone does not guarantee faithful evidence utilization.

The question (HotpotQA): "Which Oklahoma player in the 1951 Sugar Bowl went on to win the Heisman Trophy?', details in Figure 5

B Prompts and Implementation Details

B.1 Facet Decomposition Prompt (HotpotQA Dataset)

For HotpotQA, we leverage the dataset's existing supporting facts annotations. We use few-shot prompting to convert supporting sentences into interrogative facets. We design question-type-specific prompts:

1) Bridge Questions Prompt:

Facet Decomposition Prompt

Task: *Convert this bridge question into reasoning steps (facets).*
Example 1:
Question: What nationality is the director of the film Masked and Anonymous?
Supporting Facts: `[["Masked and Anonymous", 0], ["Larry Charles", 0]]`
Facets:

1. Who directed the film Masked and Anonymous?
2. What is Larry Charles's nationality?

Example 2:
Question: What year was the director of Blade Runner born?
Supporting Facts: `[["Blade Runner", 1], ["Ridley Scott", 0]]`
Facets:

1. Who directed Blade Runner?
2. When was Ridley Scott born?

Now convert this:
Question: `[INPUT_QUESTION]`
Supporting Facts: `[INPUT_FACTS]`
Facets:

2) Comparison Questions Prompt:

Facet Decomposition Prompt

Task: *Convert this comparison question into reasoning steps (facets).*
Example 1:
Question: Who was born first, Arthur Conan Doyle or Artur Schnitzler?
Supporting Facts: `[["Arthur Conan Doyle", 0], ["Artur Schnitzler", 0]]`
Facets:

1. When was Arthur Conan Doyle born?
2. When was Artur Schnitzler born?

Example 2:
Question: Which has more species, genus A or genus B?
Supporting Facts: `[["Genus A", 0], ["Genus B", 0]]`
Facets:

1. How many species are in genus A?
2. How many species are in genus B?

Now convert this:
Question: `[INPUT_QUESTION]`
Supporting Facts: `[INPUT_FACTS]`
Facets:

B.2 Facet Decomposition Prompt (Medical Dataset)

For medical questions without supporting fact annotations:

Facet Decomposition Prompt

Task: *Given the following medical question, decompose it into a list of atomic sub-questions, where each sub-question addresses a single clinical fact or reasoning step required to answer the original question. Each sub-question should be answerable independently using medical knowledge sources.*
Question: `[INPUT_QUESTION]`
Output format: *Numbered list of sub-questions.*

B.3 Facet-Level Answer Generation Prompts

We generate answers for each reasoning facet under three controlled inference modes:

B.3.1 Strict RAG Prompt

For facet-level generation with strict evidence grounding:

Example: Facet-RAG Complete Pipeline
Original Question (HotpotQA): "Which Oklahoma player in the 1951 Sugar Bowl went on to win the Heisman Trophy?" **Gold Answer:** Billy Vessels
Step 1: Facet Decomposition

- Facet 1: "Which player from Oklahoma participated in the 1951 Sugar Bowl?"
- Facet 2: "Did Billy Vessels win the Heisman Trophy?"

Step 2: Chunk Retrieval (Top-5 per facet, showing top-3 as sample)

- Facet 1 most relevant: C2(0.82), C1(0.72), C3(0.72)
- Facet 2 most relevant: C8(0.77), C9(0.77), C10(0.74)

Step 3: NLI-Based Faithfulness Calculation
For each facet answer (hypothesis), run NLI vs all chunks (premises): faithfulness $= P(\text{entailment}) - P(\text{contradiction})$. Facet faithfulness = max over all chunks.
Step 4: Facet-Level Answer Generation using prompt in Appendix B:
Facet 1: "Which player from Oklahoma participated in the 1951 Sugar Bowl?"

- **Strict RAG:** "NO_ANSWER"
- **Soft RAG:** "The player from Oklahoma who participated in the 1951 Sugar Bowl was Heisman Trophy winner Billy Vessels."
- **LLM-only:** "The player from Oklahoma who participated in the 1951 Sugar Bowl was Billy Vessels."

Facet 2: "Did Billy Vessels win the Heisman Trophy?"

- **Strict RAG:** "NO_ANSWER"
- **Soft RAG:** "Yes, Billy Vessels won the Heisman Trophy in 1952."
- **LLM-only:** "Yes, Billy Vessels won the Heisman Trophy in 1952."

Step 5: NLI-Based Faithfulness Calculation
Facet 1 faithfulness_soft = 0.00 (max over chunks), and Facet 2 faithfulness_soft = 0.103 (max from C10)
Step 6: Facet-Level Evidence Taxonomy Assignment
Facet 1: Evidence Overridden – Reason: Strong retrieval (max sim: 0.82) but faithfulness = 0.00. Strict RAG fails; Soft RAG relies on parametric knowledge.
Facet 2: Evidence Overridden – Reason: Strong retrieval (max sim: 0.77) but faithfulness = 0.103. Retrieved evidence insufficient to fully support facet answer.
Step 7: Question-Level Answer Aggregation
Facet answers aggregated to full question answers using prompt in Appendix B:

- **Strict RAG:** "NO_ANSWER" (F1 = 0.00) – Both facets returned NO_ANSWER
- **Soft RAG:** "The Oklahoma player who participated in the 1951 Sugar Bowl and went on to win the Heisman Trophy is Billy Vessels. He won the award in 1952." (F1 = 0.16) – Partial match
- **LLM-only:** "Billy Vessels." (F1 = 1.00) – Exact match
- **Question Category:** Evidence Overridden (both facets classified as Evidence Overridden)

Figure 5: Illustrative facet-level diagnostic example from HotpotQA.

Strict RAG Facet Generation

System: *You are a factual question answering assistant. Give a short, direct answer in one or two sentences. State facts clearly without background explanation.*
Constraint: *Use ONLY the evidence below to answer the question. If the evidence does not clearly support an answer, respond exactly with* `NO_ANSWER`.
Evidence:

- `[CHUNK_1]`
- `[CHUNK_2]`
- ...
- `[CHUNK_K]`

Question: `[FACET]`
Output: *Short answer (1-2 sentences) or* `NO_ANSWER`

Soft RAG Facet Generation

System: *You are a factual question answering assistant. Give a short, direct answer in one or two sentences. State facts clearly without background explanation.*
Instruction: *You may use the evidence below and your general knowledge.*
Evidence:

- `[CHUNK_1]`
- `[CHUNK_2]`
- ...
- `[CHUNK_K]`

Question: `[FACET]`
Output: *Short answer (1-2 sentences)*

B.3.2 Soft RAG Prompt

For flexible facet-level generation allowing parametric knowledge integration:

B.3.3 LLM-only Prompt

For facet-level generation without retrieval:

LLM-only Facet Generation

System: *You are a factual question answering assistant. Give a short, direct answer in one or two sentences. State facts clearly without background explanation.*
Question: `[FACET]`
Output: *Short answer (1-2 sentences)*

No external evidence is provided. The model relies entirely on parametric knowledge.

B.4 Question-Level Aggregation Prompt

To aggregate facet-level answers into a final question-level answer, we use the following prompt for Soft RAG; the remaining modes follow the same procedure as in the previous section.

Question-Level Aggregation Prompt

Task: *Given the following sub-question answers, provide a concise answer to the original question.*
Original question: `[QUESTION]`
Sub-answers:

- *Facet 1:* `[FACET_QUESTION_1]` → *Answer:* `[FACET_ANSWER_1]`
- *Facet 2:* `[FACET_QUESTION_2]` → *Answer:* `[FACET_ANSWER_2]`
- ...

Output: *Final answer*

C Model Configuration

C.1 Model Configuration Details

- GPT-4o mini: Model ID `gpt-4o-mini-2024-07-18`, `temperature=0.3`, `max_tokens=160` (facet-level), `max_tokens=15` (question-level)
- Gemini-2.0-Flash: Model ID `models/gemini-2.0-flash`, `temperature=0.3`, `max_output_tokens=160` (facet-level), `max_output_tokens=15` (question-level)
- LLaMA-3-8B-Instruct: Via HuggingFace Transformers, 8-bit quantization, `temperature=0.3`, `max_new_tokens=160` (facet-level), `max_new_tokens=15` (question-level)

C.2 Retrieval Configuration

- Chunk size: 300 tokens with 50-token overlap.
- Top-K: 5 chunks per facet. We set top-k=5 to balance recall and precision: smaller k risks missing relevant evidence, while larger k introduces irrelevant or conflicting context (Yu et al., 2024). Recent work on RankRAG demonstrates that k=5 works well across diverse QA datasets.
- Similarity: Cosine similarity

D Detailed Experimental Analysis

This appendix provides additional analyses and visualizations that support the main findings reported in Section 4. We focus on distributional behavior, per-question variability, and model-specific patterns that are difficult to convey using aggregate statistics alone.

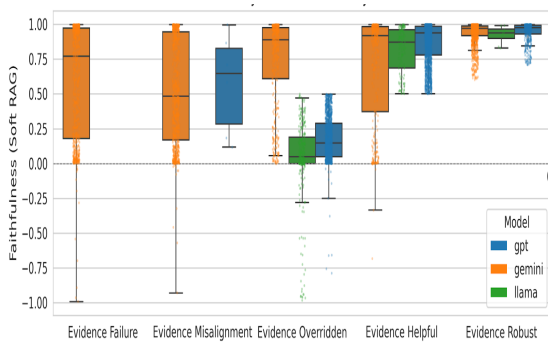

Figure 6: Medical Dataset: Facet-level faithfulness distributions under Soft RAG across evidence taxonomy categories.

D.1 Evidence Taxonomy by Question Difficulty

Figure 2 shows the distribution of evidence taxonomy categories across difficulty levels in HotpotQA. While aggregate results are reported in the main paper, this figure illustrates that Evidence Override remains consistently prevalent across easy, medium, and hard questions, whereas Evidence Robust decreases with increasing difficulty. Evidence Failure and Evidence Misalignment remain relatively rare across all levels, indicating that increased reasoning complexity primarily affects evidence integration during generation rather than retrieval coverage.

D.2 Facet-Level Faithfulness Distributions

Figure 6 presents the full distribution of facet-level faithfulness scores under Soft RAG for the medical dataset, stratified by evidence taxonomy category. Evidence Helpful and Evidence Robust facets exhibit high faithfulness with low variance across

models, while Evidence Failure and Evidence Misalignment show low faithfulness and high variance. Evidence Overridden facets display a bimodal distribution, with mass concentrated at both high entailment and strong contradiction. This distributional behavior is consistent across datasets, including HotpotQA, and highlights that overridden evidence may be either successfully incorporated or directly contradicted during generation.

D.3 Per-Question Performance Variability

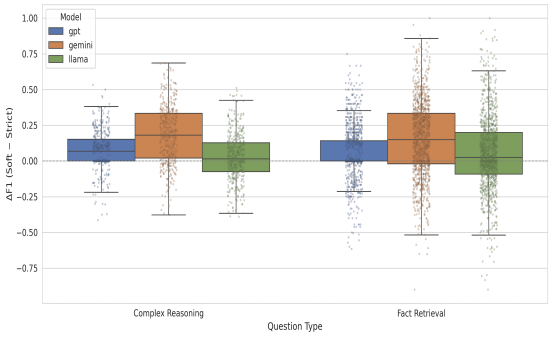

Figure 7: Medical Dataset: Per-question ΔF1 distributions (Soft − Strict) by model.

Figure 7 visualizes per-question changes in F1 induced by relaxing evidence constraints (Soft RAG vs. Strict RAG) for the medical dataset. Although Soft RAG improves average performance, the distribution reveals substantial variability across questions, with a non-trivial fraction experiencing degraded performance. This variability is more pronounced for GPT and LLaMA than for Gemini.

A similar pattern is observed for HotpotQA across difficulty levels, where medium-difficulty questions show the largest gains and hard questions exhibit increased variance. These distributions complement the mean results reported in Section 4 by illustrating the heterogeneity underlying aggregate improvements.

E Facet Decomposition Validation Details

To assess the reliability of the LLM-generated facet decomposition, we conducted a manual quality assessment on a random sample of 100 questions. The evaluation focused on three dimensions: correctness, completeness, and granularity. The validation was performed by the authors; no external participants or annotators were recruited.

Across the sampled questions, 84% of decompositions were judged correct, 88% complete, and 86% exhibited appropriate granularity. Only 1–3% showed substantive issues affecting correctness or completeness, while the remaining cases primarily reflected minor granularity differences that did not materially affect taxonomy assignments. The reported 85–90% range in the main text summarizes the overall assessment across these criteria.

Correctness assesses whether the generated facets appropriately reflect the meaning and reasoning requirements of the original question. A decomposition is considered correct if the facets correspond to the intended reasoning steps without introducing irrelevant or incorrect components.

Completeness assesses whether the decomposition covers the necessary reasoning steps required to answer the question. Missing critical reasoning components are treated as incomplete.

Granularity assesses whether facets are reasonably atomic, avoiding overly coarse decompositions that merge multiple reasoning steps or excessively fine-grained decompositions that fragment reasoning unnecessarily.