How You Ask Matters! Adaptive RAG Robustness to Query Variations

Anonymous ACL submission

Abstract

Adaptive Retrieval-Augmented Generation (RAG) promises accuracy and efficiency by dynamically triggering retrieval only when needed and is widely used in practice. However, real-world queries vary in surface form even with the same intent, and their impact on Adaptive RAG remains under-explored. We introduce the first large-scale benchmark of diverse yet semantically identical query variations, combining human-written and model-generated rewrites. Our benchmark enables systematic evaluation of Adaptive RAG robustness across answer, computational cost, and retrieval decisions. We discover a critical robustness gap, where small surface-level changes in queries dramatically alter retrieval behavior and accuracy. Although larger models show better performance, robustness does not improve accordingly. These findings reveal that Adaptive RAG methods are highly vulnerable to query variations that preserve identical semantics, exposing a critical robustness challenge.

1 Introduction

Retrieval-Augmented Generation (RAG) reduces hallucinations in large language models (LLMs) by grounding outputs in retrieved evidence, thereby improving factual accuracy in downstream tasks (Khandelwal et al., 2019; Lewis et al., 2020). While effective in reducing factual errors, conventional RAG often retrieves irrelevant or noisy documents, which can mislead reasoning and degrade answer quality (Li et al., 2023; Ding et al., 2025).

To address these limitations, Adaptive RAG conditions retrieval decisions on the input query and its evolving context (Jeong et al., 2024). Unlike conventional RAG with a fixed retrieval pattern, it adaptively decides *when* and *what* to retrieve as generation proceeds. For simple factoid questions, a small number of retrievals is sufficient, whereas multi-hop questions typically benefit from iterative subqueries that aggregate evidence across hops (Ho et al., 2020; Trivedi et al., 2022, 2023). As a result, prior adaptive approaches can improve both accuracy and efficiency by allocating retrieval and LLM calls based on query difficulty and intermediate states (Su et al., 2024; Jiang et al., 2023; Ding et al., 2025). Due to their strong performance, these frameworks are increasingly deployed in real-world applications (Singh et al., 2025).

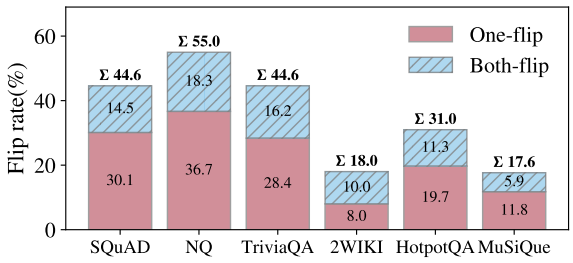


Figure 1: **Retrieval-decision flip rates** of the Qwen-32B model, measuring how often the model changes its retrieval judgment under meaning-preserving human query rewrites. Higher rates signal instability. Flips are categorized as *One-flip* (only one rewrite flips) and *Both-flip* (both rewrites flip).

In real-world settings, users express the same intent in diverse forms, including paraphrases, stylistic variations, typos or other noise (Fu et al., 2025; Zhang et al., 2025; Cao et al., 2025). To examine the model's decision stability under query variations, we measure how often its retrieval decision (retrieve or not) flips between the original query and a human rewrite. As shown in Figure 1, human rewrites trigger decision flips in up to 55% of cases, revealing a severe lack of robustness in the system's retrieval decisions under human queries. Despite these risks, robustness to real-world query variation has not been systematically evaluated, which is particularly concerning in high-stakes settings where factual correctness and user trust are critical (Sharma et al., 2024; Oche et al., 2025).

This paper presents the first empirical study of Adaptive RAG robustness under realistic query

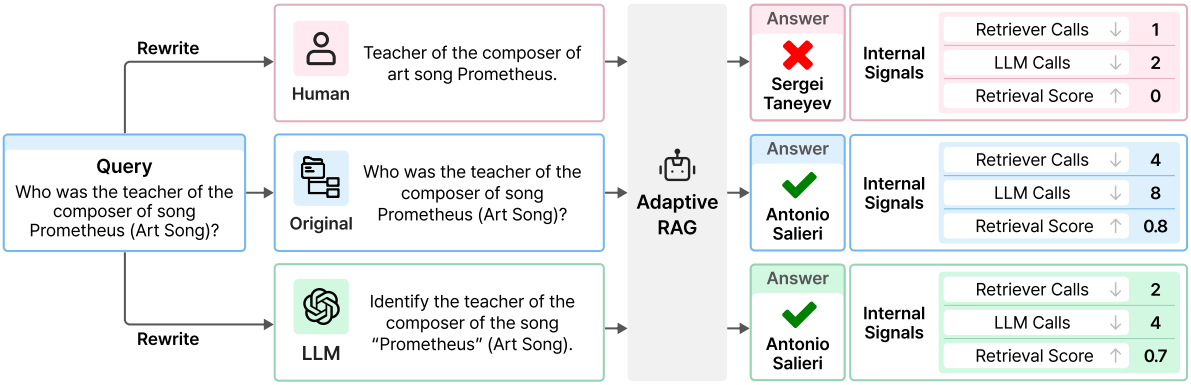


Figure 2: **Example of Adaptive RAG responses** under human, original, and LLM-generated query rewrites, highlighting differences in answer correctness, computation overhead, and retrieval score.

variations. To facilitate this, we introduce a benchmark spanning controlled linguistic paraphrases and naturally written human queries. We evaluate on six widely used question-answering (QA) datasets covering both single-hop and multi-hop questions and analyze how surface-level differences affect both intermediate retrieval trajectories and final answers (Figure 2). For each question, we include six model-generated variants and two human rewrites, yielding 27K QA pairs spanning diverse levels of complexity. Across these query variations, we comprehensively assess model behavior along three dimensions: answer quality, computational cost, and retrieval decisions.

This work presents the following key findings:

- Spelling errors cause the largest accuracy and robustness drop at the cost of semantic distortion, while human rewrites cause the second-largest drop but preserve the original meaning.
- Improved performance does not imply greater robustness; even strong baselines (e.g., the generation-based method, QwQ-32B) show substantially greater inconsistency in answer semantics and computational cost.
- Even minor changes to the initial query can alter the intermediate reasoning path and the subqueries generation, leading to retrieval failure and ultimately lower-quality final answers.

2 Related Works

2.1 Adaptive RAG

Adaptive RAG implements dynamic, model-driven decisions about when to retrieve and what to retrieve. We broadly categorize prior methods into two main streams: logit-based and generation-based methods. Logit-based methods monitor model confidence and trigger retrieval when the model appears uncertain, for example, using token-level probabilities as in Jiang et al. (2023) or uncertainty estimates based on token probabilities and attention weights as in Su et al. (2024). Generation-based methods, on the other hand, let the model decide on the fly when it needs retrieval by generating search queries during the generation process (Li et al., 2025). Our work focuses on approaches that use a single model with a retriever and do not rely on external classifiers, such as Jeong et al. (2024).

2.2 Query Robustness

A growing line of work studies robustness of RAG systems. Several studies show that linguistic perturbations such as query-entry errors, grammatical errors, typos, and stylistic variations affect QA performance (Cao et al., 2025; Perçin et al., 2025; Zeng et al., 2025). These findings demonstrate that even minor modifications to a query can substantially alter retrieved documents and downstream answers (Zhang et al., 2025; Fu et al., 2025). However, prior work has two key limitations: (1) it focuses primarily on conventional RAG, leaving the additional complexity introduced by adaptive RAG largely unexplored; (2) it does not examine the effects of real human query reformulations across different RAG frameworks.

3 Benchmark Construction

3.1 Dataset

We evaluate single-hop and multi-hop QA, using Natural Questions (NQ) (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017), and SQuAD v1.1

Original Query: Who is the child of the director of film Mukhyamantri (1996 Film)?		
Category	**Variation**	**Sample Rewrites**
Style	Formality (Form.) ↓	Yo, quick one — who's the kid of the director behind the 1996 flick Mukhyamantri?
	Readability (Read.) ↓	By whom is the progeny of the director of Mukhyamantri (1996 Film) constituted?
Type	Declarative (Decl.)	I want to know who is the child of the director of the film Mukhyamantri (1996 Film).
	Imperative (Impr.)	Name the child of the director of film Mukhyamantri (1996 Film).
Error	Spelling (Spell.) ↓	Who is ght childer ar the directer of filme Mukhyamantri (1996 Film)?
	Grammar (Gram.) ↓	Who are the childs of the director from film Mukhyamantri 1996?

Table 1: **Examples of query rewrites** grouped by category, illustrating stylistic, structural, and error-based variations. ↓ denotes reduction in corresponding writing variation.

(Rajpurkar et al., 2016) for single-hop, and 2WikiMultiHopQA (2WIKI) (Ho et al., 2020), MuSiQue (Trivedi et al., 2022), and HotpotQA (Yang et al., 2018) for multi-hop evaluation.

3.2 Query Variation Generation

Building upon prior works (Moskvoretskii et al., 2025; Jeong et al., 2024; Trivedi et al., 2023), we evaluate the systems on the same 500 samples per dataset. To generate query variations, we employ GPT-5-mini and produce six distinct linguistic variants grouped into three categories: sentence style, sentence type, and error, as shown in Table 1. While earlier studies examined style and error (Zhang et al., 2025; Perçin et al., 2025; Cao et al., 2025), we extend this paradigm by incorporating sentence type variations, covering 24k queries.

Sentence Style. Human queries exhibit diverse linguistic styles, often characterised by less formality, simplified language, or omission of polite expressions. Following previous work (Cao et al., 2025), we introduce a set of natural queries with reduced formality and readability. Each variation is generated by LLM with prompts detailed in the Appendix B. For validation, we follow the process outlined in Cao et al. (2025) (see Appendix C.1). We formally define these variations as follows:

- **Formality**: We generate informal versions using simpler vocabulary, contractions, and more casual expressions.
- **Readability**: We reduce the readability of the queries, making them more difficult to understand. Although this may sacrifice clarity, it more accurately reflects how users often communicate in less structured contexts.

Sentence Type. Sentence type variations capture the grammatical function of the sentence, reflecting how the query is structured. Most current datasets cover interrogative questions, particularly in the form of wh-questions (e.g., who, what, when) (Rogers et al., 2023; Kwong and Yorke-Smith, 2012). Therefore, we challenge the model by generating two additional sentence types: declarative and imperative. We instruct GPT-5-mini to generate queries that fit each of these sentence types, and the used prompts are given in Appendix B.

- **Declarative sentence**: These queries make statements or assertions. They are often used to express knowledge, desire, or intention.
- **Imperative sentence**: These queries are commands or requests. They can suggest actions or instructions, typically used when the user wants the system to perform a task.

Error. We focus on the most frequent error types observed in human-generated queries: grammatical and spelling errors (Hagiwara and Mita, 2019). These kinds of errors commonly occur in user input due to hurried typing, typographical mistakes, and limited mastery of grammatical rules.

- **Spelling errors**: These errors occur when a user mistakenly types a word with incorrect letters or letter combinations, resulting in a non-standard word form. We use the python library *nlpaug* tool[1] to inject the spelling errors to the original query (Zhang et al., 2025).
- **Grammatical errors**: We implement these errors by prompting GPT-5-mini to inject grammatical mistakes. The prompts are provided in

[1] pypi.org/project/nlpaug/0.0.5

the Appendix B. After that, we use LanguageTool to check whether the query variations include at least one grammar error.

3.3 Human Rewrite Annotation

Human queries exhibit substantial variation in form and do not follow a single, consistent pattern; instead, they are a combination of multiple categories (Penha et al., 2021; Li et al., 2024). However, existing robustness evaluations for RAG systems largely overlook this dimension, as they do not incorporate human written queries (Cao et al., 2025; Perçin et al., 2025; Zhang et al., 2025). In this work, we collected two human rewrites for 600 queries (100 per dataset across six QA benchmarks), yielding 1,200 rewrites, as well as additional query reformulations from 12 annotators. Annotators were given the original query and its corresponding answer and were instructed to generate alternative formulations as if posing the question themselves. Detailed guidelines are provided in the Appendix C.2.

4 Experimental Setup

4.1 Methods

We consider two Adaptive RAG methods that opened new paradigms: logit-based control (LB) and generation-based control (GB). For the LB approach, we use DRAGIN (Su et al., 2024), which determines when and what to retrieve using token-level signals (e.g., probabilities and attention). For the GB approach, we use Search-o1 (Li et al., 2025), where the model uses its internal reasoning to construct a retrieval plan, including whether to initiate search, how to formulate queries, and how to derive the final answer. Unlike the threshold-style control of LB, Search-o1 performs retrieval control end-to-end without external triggers such as hand-tuned thresholds or auxiliary classifiers.

4.2 Metrics

Answer Robustness. We evaluate answer correctness using InAccuracy, following Baek et al. (2023), to check whether the gold answer is included in the model answer. In addition, we measure the semantic consistency of generated answers across variations using similarity and diversity.

1. **InAccuracy:** For each instance i and perturbation type v, let $\hat{a}_{i,v}$ be the generated answer and g_i the gold answer. InAccuracy is reported as the dataset-level mean for each perturbation type v.

$$\mathrm{InAcc}_v = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}[g_i \subseteq \hat{a}_{i,v}]$$

2. **Similarity:** Let $\phi(\cdot)$ denote an embedding function (SBERT). We denote the answers to the original and reformulated queries as $\hat{a}i, \mathrm{orig}$ and $\hat{a}i, v$, respectively, and define $e_{i,v} = \phi(\hat{a}i, v)$. We compute the cosine similarity between $\hat{a}i, v$ and $\hat{a}_{i,\mathrm{orig}}$.

$$\mathrm{Sim}_v = \frac{1}{N}\sum_{i=1}^{N} \cos(e_{i,v}, e_{i,orig})$$

3. **Diversity (Div.):** Let $\mathcal{V}$ be the set of variants (including the original) for instance i, and let $K = |\mathcal{V}|$. We quantify answer spread as:

$$\mathrm{Div}_i = \frac{2}{K(K-1)}\sum_{v<w}\big(1 - \cos(e_{i,v}, e_{i,w})\big)$$

$$\mathrm{Div} = \frac{1}{N}\sum_{i=1}^{N}\mathrm{Div}_i$$

Computation Robustness. We measure call robustness for the LLM and the retriever using the average number of calls made for a single question. To assess robustness, we introduce an evaluation protocol that measures (i) the change in computational cost relative to the original queries and (ii) the consistency of cost across query variations. Let $c_{i,v}$ denote the call count for instance i under variant v, with $v = 0$ denoting the original query.

1. **Relative Error (RE):** For each instance i and variant v, we compute the deviation from the original-query call count:

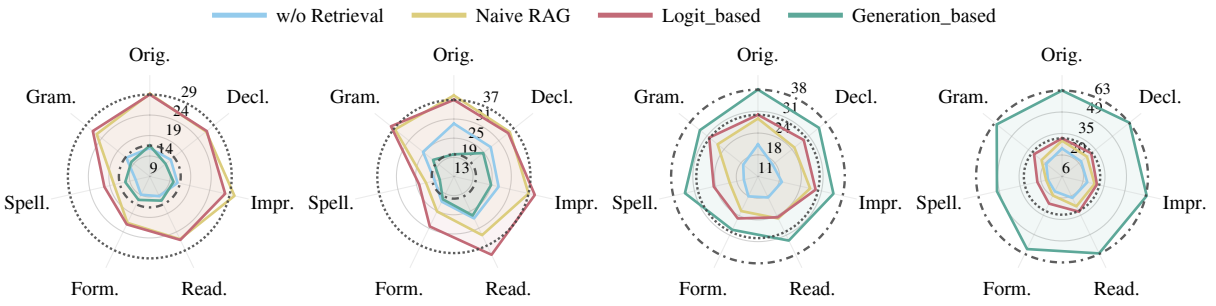
$$\mathrm{RE}_{i,v} = \frac{|c_{i,v} - c_{i,0}|}{\max(1, c_{i,0})}$$

We then report the dataset-level mean for each perturbation type v:

$$\mathrm{RE}_v = \frac{1}{N}\sum_{i=1}^{N}\mathrm{RE}_{i,v}$$

2. **Coefficient-of-Variation Robustness (CVR):** Let $\mathcal{V}$ be the set of variants (including the original) and $r_i = (c_{i,v})_{v\in\mathcal{V}}$. We compute instance-wise variability across variants:

$$\mathrm{CV}_i = \frac{\mathrm{std}(r_i)}{\mathrm{mean}(r_i)}, \qquad \mathrm{CVR}_i = \frac{1}{1+\mathrm{CV}_i}$$



(a) SQuAD / Llama-3.1 (b) 2WIKI / Llama-3.1 (c) SQuAD / QwQ (d) 2WIKI / QwQ

Figure 3: **InAccuracy results** across six perturbation types for a given dataset / model pair.

The final score is averaged over instances:

$$\mathrm{CVR} = \frac{1}{N}\sum_{i=1}^{N}\mathrm{CVR}_i$$

We report RE for each perturbation type and CVR aggregated across all perturbations, for both **retriever calls** and **LLM calls**.

Retrieval Decision Robustness. We evaluate whether the decision to retrieve is correct. The variable *act* indicates whether retrieval is triggered ($act = 1$ if triggered; $act = 0$ otherwise). The variable *need* is defined based on a no-retrieval baseline: if the model answers the question correctly without retrieval, then $need = 0$; otherwise, $need = 1$.

1. **Overconfidence**: The rate of cases where the model fails to retrieve despite retrieval being needed (i.e., $need = 1$, $act = 0$), following Moskvoretskii et al. (2025).

2. **Underconfidence**: The rate of cases where the model retrieves despite retrieval not being needed (i.e., $need = 0$, $act = 1$), following Moskvoretskii et al. (2025).

4.3 Implementation Details

We use two LLMs, QwQ-32B (Team, 2025) and Llama 3.1-8B-Instruct (Grattafiori et al., 2024). For retrieval, we employ the sparse retriever BM25, applied to a document pool constructed from Wikipedia documents (Moskvoretskii et al., 2025). See Appendix A for further implementation details.

5 Results

We present results on two representative datasets – SQuAD (single-hop) and 2WIKI (multi-hop) – and report others in the Appendix.

Dataset	Mtd.	Similarity ↑						Div. ↓
		Form.	Read.	Decl.	Impr.	Spell.	Gram.	
Llama-3.1								
SQuAD	LB	0.551	0.539	0.593	**0.634**	0.552	0.599	0.469
	GB	0.426	0.408	**0.433**	0.430	0.411	0.414	0.585
2WIKI	LB	0.560	0.628	0.649	**0.655**	0.537	0.627	0.414
	GB	0.423	**0.458**	0.300	0.403	0.313	0.385	0.667
QwQ								
SQuAD	LB	0.573	0.579	0.615	**0.631**	0.544	0.584	0.440
	GB	0.552	0.558	0.579	**0.589**	0.552	0.579	0.459
2WIKI	LB	0.500	0.545	0.551	**0.565**	0.462	0.519	0.511
	GB	0.520	0.557	0.554	**0.559**	0.499	0.548	0.465

Table 2: **Answer similarity and diversity** across query variations. Darker shades indicate more desirable outcomes: higher similarity and lower diversity.

5.1 Answer Robustness

Model performance drastically drops on query variations. We first analyze how query variations affect the answer accuracy. Figure 3 shows spelling error perturbation yields the largest accuracy drop across dataset/model pairs, indicating strong sensitivity to surface-form noise. Generation-based methods benefit from larger models and harder datasets. For example, on 2WIKI with QwQ, accuracy still degrades with changes in query phrasing (63.4% for the original query vs. 50.6% for spelling errors), with drops of up to 12.8% across variations. Results across all 6 datasets are shown in the Appendix D.1.

Generation-based method shows inconsistent outputs. Table 2 confirms the same trend: imperative rewrites maintain the highest answer similarity to the original, whereas spelling error consistently produces the lowest similarity. Moreover, generation-based methods exhibit higher diversity (lower is better), implying less stable outputs across perturbation.

	Target	Mtd.	SQuAD RE ↓ Form.	Read.	Decl.	Impr.	Spell.	Gram.	SQuAD CVR ↑	2WIKI RE ↓ Form.	Read.	Decl.	Impr.	Spell.	Gram.	2WIKI CVR ↑
Llama-3.1	RTR	LB	0.472	0.592	0.416	**0.338**	0.431	0.439	0.734	0.853	0.859	**0.455**	0.846	0.928	0.941	0.710
		GB	0.092	0.094	**0.086**	0.090	0.100	0.092	0.937	0.468	0.416	**0.402**	0.542	0.496	0.435	0.742
	LLM	LB	0.441	0.592	0.396	**0.313**	0.416	0.411	0.767	1.001	1.001	**0.427**	1.007	1.093	1.130	0.742
		GB	0.056	0.059	**0.048**	0.055	0.061	0.051	0.963	0.447	0.404	**0.401**	0.520	0.477	0.418	0.745
QwQ	RTR	LB	0.494	0.450	**0.405**	0.415	0.495	0.441	0.740	0.560	0.502	0.446	**0.443**	0.565	0.511	0.719
		GB	0.360	0.322	**0.298**	0.315	0.361	0.330	0.762	0.788	0.808	**0.780**	0.825	0.787	0.884	0.689
	LLM	LB	0.477	0.466	**0.389**	0.399	0.491	0.424	0.760	0.547	0.481	**0.426**	0.430	0.565	0.493	0.747
		GB	0.265	0.230	**0.225**	0.238	0.250	0.243	0.858	1.078	1.142	1.092	1.203	**1.049**	1.219	0.716

Table 3: **Computation robustness** across different model-generated query variations. Darker shading indicates more desirable outcomes: lower RE and higher CVR. RTR = Retriever. Bold denotes maximum across the variations.

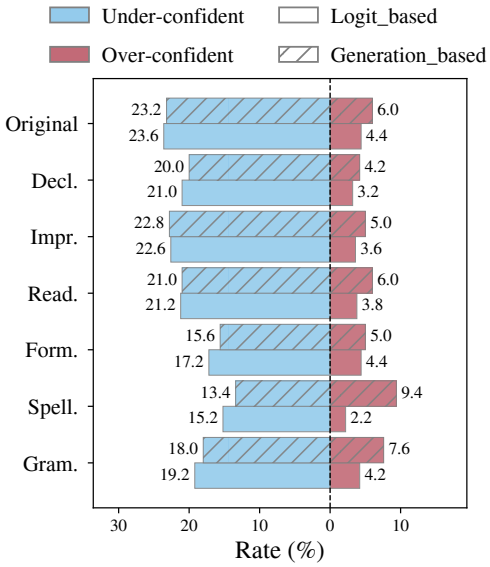
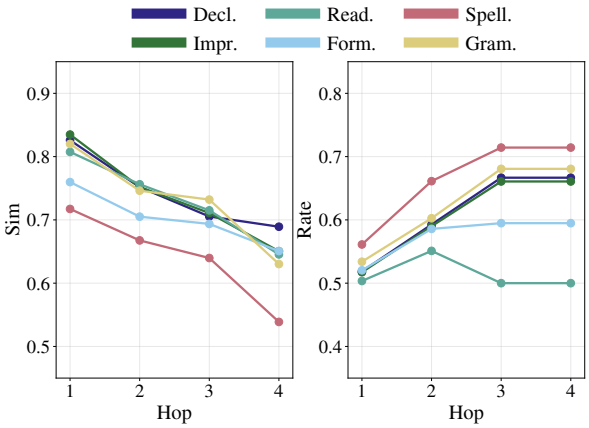


Figure 4: **Under- and over-confident rates** on model query variations.



(a) Subquery Drift (b) Retrieval Failure Rate

Figure 5: **Subquery drift effect on retrieval quality.** 2WIKI Dragin results (a) Subquery similarity to the original decreases across turns. (b) Retrieval failure rate (no gold document in top-k) across turns.

5.2 Computation Robustness

Computation robustness destabilizes on harder questions. We evaluate computational robustness using RE and CVR to assess whether the model triggers the retriever and the LLM consistently across query variations. Table 3 shows that retrieval-call counts are most sensitive on 2WIKI, exhibiting larger RE across most perturbations. Spelling and grammatical errors produce the highest RE, indicating that surface-level noise most strongly shifts the retrieval-call budget.

LLM-call robustness follows a similar trend. On SQuAD, RE remains moderate and CVR stays high for Llama (up to 0.963 with the generation-based method), suggesting stable call counts under rewrites. In contrast, on 2WIKI, RE increases and CVR decreases, with the largest degradation under spelling errors.

Computation robustness does not imply higher accuracy. In Figure 3, generation-based method achieves strong accuracy on harder datasets when paired with larger models. However, this trend reverses for computational robustness in Table 3. For 2WIKI with QwQ model, computational robustness degrades substantially. This contrast indicates that high accuracy does not necessarily correspond to strong computational robustness.

5.3 Retrieval Decision Robustness

Query errors invoke over-confidence. We investigate the retrieval decision correctness using under- and over-confidence metrics. Under-confident instances incur additional costs, while over-confidence harms accuracy by skipping necessary retrievals. Figure 4 shows spelling-error appear on high over-confident cases, reaching up to 9.4%. This tendency persists particularly in larger models.

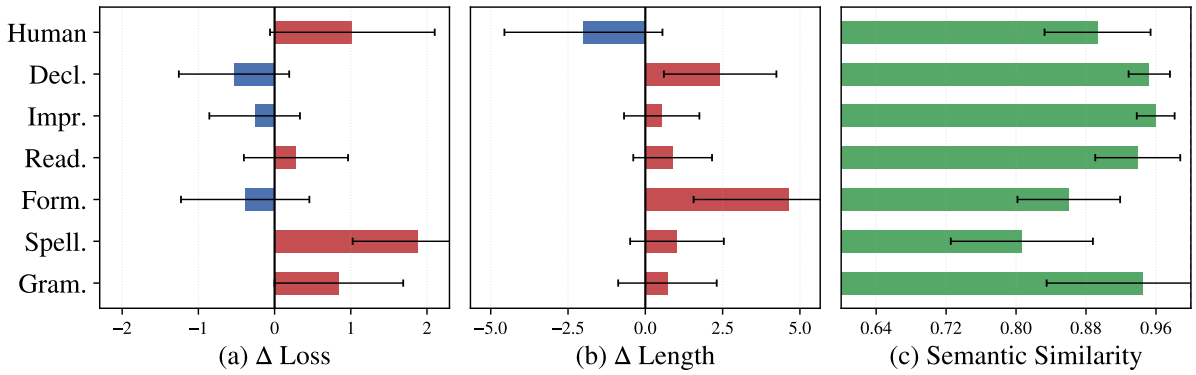


Figure 6: **Query analysis.** From left to right: (a) change in LM loss relative to the original query (ΔLoss; nats/token), (b) change in query length (ΔLength; words), and (c) semantic similarity to the original query (cosine similarity).

Mtd.	Similarity ↑							Div. ↓
	Human	Form.	Read.	Decl.	Impr.	Spell.	Gram.	
LB	0.518	0.528	0.570	0.573	0.612	0.522	0.572	0.477
GB	0.529	0.542	0.534	0.568	0.598	0.520	0.558	0.460

Table 4: **Answer Robustness with human queries.** Similarity to the original query and diversity on 2WIKI (QwQ-32B) including human rewrites.

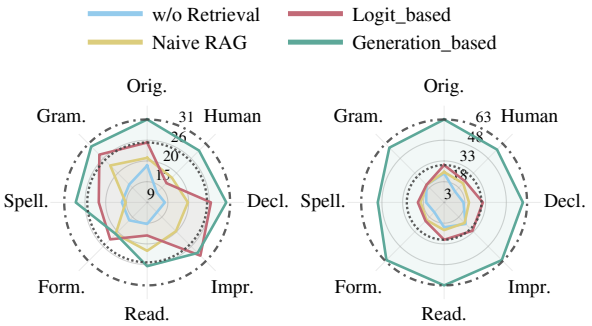


Figure 7: **InAccuracy results** with human rewrites

Drifted subqueries miss gold documents. We investigate retrieval-decision robustness in a multi-hop Adaptive RAG setting. At hop t, the model generates a subquery q_t and retrieves a top-k document set D_t. We analyze how query variations perturb the multi-hop retrieval process by tracking (i) semantic drift in intermediate subqueries and (ii) hop-level retrieval failures rates. To quantify drift, we compute the mean subquery similarity (MSS), defined as the average semantic similarity between each variant subquery and its corresponding original subquery across hops. Retrieval failures are quantified using the retrieval failure rate (RFR), defined as the proportion of hops in which the top-k retrieved documents contain no gold document, i.e., $\text{recall@}k(D_t^{\text{var}}, D_t^{\text{Gold}}) = 0$.

As subqueries deviate from the original trajectory (lower MSS), the retrieved top-k sets increasingly fail to cover gold evidence, resulting in a higher RFR. Figure 5 illustrates this trend: subquery similarity decreases while RFR increases across hops, indicating that subquery drift directly undermines retrieval quality.

6 Human Query Analysis

Human rewrites are semantically faithful but harder for the model. Figures 6(a) and (c) show that human rewrites preserve semantic intent of the original queries, yet they induce substantially higher token-level loss. They are also shorter, frequently compressing or omitting surface cues as in Figure 6(b).

In Figure 7, human-authored rewrites lead to accuracy degradation compared to original query, in some cases comparable to that caused by spelling-errors. Table 4 further shows reduced answer stability relative to the original queries. This degradation is similar to that observed under simple surface-level perturbations, indicating that robustness failures are not limited to synthetic noise. To identify the source of errors, we trace a causal chain from properties of human rewrites to subquery drift, retrieval failures, and ultimately reduced end-to-end answer accuracy.

Human query length and uncertainty are associated with subquery drift. Shorter and more uncertain human rewrites are reflected in subquery drift. For Search-o1 on 2WIKI, query loss is negatively correlated with mean subquery similarity (MSS; Spearman's ρ=-0.15), while query length is

t	Original Subqueries	Human Subqueries	Sim	Hit(O)	Hit(H)
	Which film has the director born earlier, The Adventures Of Priscilla, Queen Of The Desert or Harvest: 3,000 Years?	Who was born first director of The Adventures Of Priscilla, Queen Of The Desert or Harvest: 3,000 Yearst			
1	director of the adventures of priscilla, queen of the desert	director of the adventures of priscilla queen of the desert	0.99	✓	✓
2	director of harvest 3000 years	director of harvest 3000 years	1.00	✓	✓
3	“harvest: 3,000 years” director birth year	harvest 3000 years movie director	0.85	✓	×
4	giuliano carnimeo birth year	harvest 3000 years korean movie director	0.13	✓	×

Table 5: **Per-hop comparison of original and human subqueries.** Sim is the semantic similarity between the two subqueries at hop t. Hit(O) indicates whether the retrieved top-k set contains any ground-truth supporting evidence for the original(O) and human(H) query.

positively correlated with MSS (ρ=0.29). These results suggest that higher-loss, more compressed human phrasing destabilizes intermediate subqueries, leading to lower MSS.

Semantic drift leads to retrieval trajectory collapse and evidence failures. Once intermediate subqueries drift, retrieval becomes unstable. For Search-o1 on human rewrites, MSS is strongly negatively correlated with trajectory collapse (ρ=-0.75), indicating that retrieved documents increasingly diverge from those for the original queries. MSS is also negatively correlated with RFR (ρ=-0.35), suggesting that gold documents are more often missing from the retrieved set. These correlations imply that human rewrites leads not only to different retrieval trajectories but also to failures in retrieving necessary evidence.

Figure 8 visualizes this pattern: the gold-document inclusion rate decreases as multi-hop reasoning progresses, ultimately dropping well below the rates for the original and model queries. Table 5 illustrates the underlying mechanism with an example from QwQ outputs. In the initial hops, the model-generated subqueries remain highly similar to the original query (similarity ≈ 0.99–1.00), but they begin to drift starting at hop 3. This drift leads to retrieval failures, as reflected by the hit metric.

Retrieval errors propagate. Failures in intermediate retrieval steps translate into end-to-end accuracy degradation. Retrieval failure is negatively correlated with the performance gap Y (Spearman's $\rho = -0.20$ for Search-o1). When the system fails to retrieve gold documents, it lacks the evidence required to answer the query, and downstream generation rarely recovers.

Overall, our findings indicate that human rewrites preserve semantic meaning yet increase the difficulty for Adaptive RAG systems. Higher query loss and more compressed phrasing amplify subquery drift, which in turn destabilizes the retrieval trajectory and increases retrieval failure, ultimately degrading final answer accuracy.

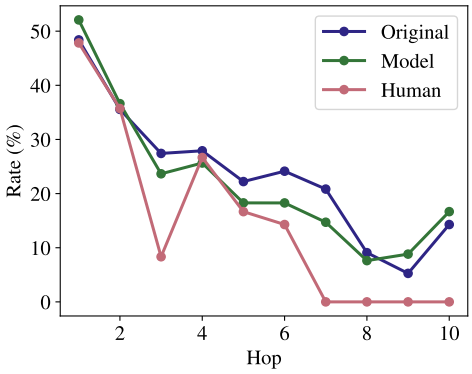


Figure 8: **Gold document inclusion rate across hops.** For each hop, we measure how often the top-k retrieved results contain at least one gold document.

7 Conclusion

In this work, we introduced a large-scale benchmark of query variations to evaluate the robustness of Adaptive RAG systems. We comprehensively assessed three dimensions—answer robustness, computation robustness, and retrieval-decision robustness—covering both final answers and intermediate reasoning steps. We showed that both logit-based and generation-based adaptive methods were sensitive to query perturbations, which degraded accuracy, efficiency, and retrieval-decision stability. Human rewrites were particularly detrimental: they induced drift in intermediate subqueries and triggered retrieval-trajectory collapse, leading to substantial end-to-end performance drops. Finally, we found that human rewrites differed systematically from model-generated variations, highlighting a realism gap in current robustness evaluations.

Limitations

Although our study offers a systematic view of Adaptive RAG responses under realistic query variation, several aspects remain outside our current scope. First, our human rewrites are designed to be meaning-preserving paraphrases of existing questions. This setting captures a common and well-controlled form of user rephrasing, but it does not include broader interactive behaviors such as multi-turn follow-up questions, conversational reformulations, or intentionally adversarial inputs. Second, we evaluate a specific set of retrievers and Adaptive RAG controllers. Other retrieval architectures or decision mechanisms may exhibit different robustness–efficiency profiles, and extending coverage to a wider range of systems is a straightforward next step. Finally, although our human rewrites are carefully collected and anonymized, they represent only a small portion of the full diversity of real-world query phrasing. Scaling human-written evaluations across languages, domains, and user populations would further improve coverage.

References

Jinheon Baek, Soyeong Jeong, Minki Kang, Jong C. Park, and Sung Ju Hwang. 2023. Knowledge-augmented language model verification. *Preprint*, arXiv:2310.12836.

Tianyu Cao, Neel Bhandari, Akhila Yerukola, Akari Asai, and Maarten Sap. 2025. Out of style: Rag's fragility to linguistic variation. *Preprint*, arXiv:2504.08231.

Hanxing Ding, Liang Pang, Zihao Wei, Huawei Shen, and Xueqi Cheng. 2025. Rowen: Adaptive retrieval-augmented generation for hallucination mitigation in llms. *Preprint*, arXiv:2402.10612.

Daocheng Fu, Jianbiao Mei, Licheng Wen, Xuemeng Yang, Cheng Yang, Rong Wu, Tao Hu, Siqi Li, Yufan Shen, Xinyu Cai, Pinlong Cai, Botian Shi, Yong Liu, and Yu Qiao. 2025. Re-searcher: Robust agentic search with goal-oriented planning and self-reflection. *Preprint*, arXiv:2509.26048.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Masato Hagiwara and Masato Mita. 2019. Github typo corpus: A large-scale multilingual dataset of misspellings and grammatical errors. *CoRR*, abs/1911.12893.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6609–6625, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong C. Park. 2024. Adaptive-rag: Learning to adapt retrieval-augmented large language models through question complexity. *Preprint*, arXiv:2403.14403.

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. *Preprint*, arXiv:2305.06983.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2019. Generalization through memorization: Nearest neighbor language models. *CoRR*, abs/1911.00172.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Helen Kwong and Neil Yorke-Smith. 2012. Detection of imperative and declarative question–answer pairs in email conversations. *AI Communications*, 25(4):271–283.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. *CoRR*, abs/2005.11401.

Daliang Li, Ankit Singh Rawat, Manzil Zaheer, Xin Wang, Michal Lukasik, Andreas Veit, Felix Yu, and Sanjiv Kumar. 2023. Large language models with controllable working memory. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1774–1793, Toronto, Canada. Association for Computational Linguistics.

Xiaoxi Li, Guanting Dong, Jiajie Jin, Yuyao Zhang, Yujia Zhou, Yutao Zhu, Peitian Zhang, and Zhicheng Dou. 2025. Search-o1: Agentic search-enhanced large reasoning models. *Preprint*, arXiv:2501.05366.

Zhicong Li, Jiahao Wang, Zhishu Jiang, Hangyu Mao, Zhongxia Chen, Jiazhen Du, Yuanxing Zhang, Fuzheng Zhang, Di Zhang, and Yong Liu. 2024. Dmqr-rag: Diverse multi-query rewriting for rag. *Preprint*, arXiv:2411.13154.

Viktor Moskvoretskii, Maria Lysyuk, Mikhail Salnikov, Nikolay Ivanov, Sergey Pletenev, Daria Galimzianova, Nikita Krayko, Vasily Konovalov, Irina Nikishina, and Alexander Panchenko. 2025. Adaptive retrieval without self-knowledge? bringing uncertainty back home. *Preprint*, arXiv:2501.12835.

Agada Joseph Oche, Ademola Glory Folashade, Tirthankar Ghosal, and Arpan Biswas. 2025. A systematic review of key retrieval-augmented generation (rag) systems: Progress, gaps, and future directions. *Preprint*, arXiv:2507.18910.

Gustavo Penha, Arthur Câmara, and Claudia Hauff. 2021. Evaluating the robustness of retrieval pipelines with query variation generators. *CoRR*, abs/2111.13057.

Sezen Perçin, Xin Su, Qutub Sha Syed, Phillip Howard, Aleksei Kuvshinov, Leo Schwinn, and Kay-Ulrich Scholl. 2025. Investigating the robustness of retrieval-augmented generation at the query level. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²)*, pages 439–457, Vienna, Austria and virtual meeting. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *Preprint*, arXiv:1606.05250.

Anna Rogers, Matt Gardner, and Isabelle Augenstein. 2023. Qa dataset explosion: A taxonomy of nlp resources for question answering and reading comprehension. *ACM Computing Surveys*, 55(10):1–45.

Sanat Sharma, David Seunghyun Yoon, Franck Dernoncourt, Dewang Sultania, Karishma Bagga, Mengjiao Zhang, Trung Bui, and Varun Kotte. 2024. Retrieval augmented generation for domain-specific question answering. *Preprint*, arXiv:2404.14760.

Aditi Singh, Abul Ehtesham, Saket Kumar, and Tala Talaei Khoei. 2025. Agentic retrieval-augmented generation: A survey on agentic rag. *Preprint*, arXiv:2501.09136.

Weihang Su, Yichen Tang, Qingyao Ai, Zhijing Wu, and Yiqun Liu. 2024. Dragin: Dynamic retrieval augmented generation based on the information needs of large language models. *Preprint*, arXiv:2403.10081.

Qwen Team. 2025. Qwq-32b: Embracing the power of reinforcement learning.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. Musique: Multi-hop questions via single-hop question composition. *Preprint*, arXiv:2108.00573.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. *Preprint*, arXiv:2212.10509.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *Preprint*, arXiv:1809.09600.

Yixiao Zeng, Tianyu Cao, Danqing Wang, Xinran Zhao, Zimeng Qiu, Morteza Ziyadi, Tongshuang Wu, and Lei Li. 2025. Rare: Retrieval-aware robustness evaluation for retrieval-augmented generation systems. *Preprint*, arXiv:2506.00789.

Kepu Zhang, Zhongxiang Sun, Weijie Yu, Xiaoxue Zang, Kai Zheng, Yang Song, Han Li, and Jun Xu. 2025. Qe-rag: A robust retrieval-augmented generation benchmark for query entry errors. *Preprint*, arXiv:2504.04062.

A Implementation Details

We conduct our experiments with two LLMs: Llama 3.1-8B-Instruct and QwQ-32B. Llama-3.1-8B-Instruct is a lightweight model from the Llama-3.1 family, optimized for instruction-following tasks. QwQ-32B is part of the Qwen model family and is specifically optimized for reasoning, supporting long-context and multi-step inference with strong performance on mathematical, programming, and logic-intensive tasks. We generate responses using nucleus sampling with temperature $\mathcal{T} = 0.7$ and top-$p = 0.95$, and report all results from a single run. We run inference locally using 2 × NVIDIA A100 GPUs with Intel(R) Xeon(R) Silver 4210R CPU @ 2.40GHz.

For the DRAGIN method, following prior work (Su et al., 2024), we tune the retrieval threshold by evaluating several high-performing candidate values on the Natural Questions (NQ) dataset and select 0.6. We then fix this setting and apply the same configuration to all remaining datasets to ensure a consistent evaluation of computational cost across benchmarks. For the Search-o1 method, maximum search limit hyperparameter is set 15 for the multi-hop QA and 5 for the single-hop QA. Rest of the hyperparamer follows the (Li et al., 2025) setting. All of the embedding similarity is calculated by the `sentence-transformers/all-mpnet-base-v2`, which is suitable for clustering and semantic search task.

B Prompts

We report the prompts for model-generated query variations.

Formality Variation. We adopt the lower-formality rewriting prompt proposed in prior work (Cao et al., 2025) to rewrite each query in a more casual style while preserving its original meaning.

Formality Variation Prompt Template

You are an AI assistant skilled at transforming formal queries into casual, everyday language.
Rewrite the following query so that it sounds very informal. Experiment with different colloquial openings, varied sentence constructions, and a mix of slang, idioms, and casual expressions throughout the sentence. Avoid using the same phrase repeatedly (e.g., 'hey, so like') and ensure the meaning remains unchanged.

Readability Variation. We utilize the prompt from Cao et al. (2025) shown in Table 6 to reduce the readability of the original query.

Sentence Type Variation. This prompt rewrites the given query into a specified sentence type (declarative or imperative) while preserving meaning. We implement this prompt for both declarative and imperative variations.

Sentence Type Variation Prompt Template

Task: Rewrite Query as {*sentence_type*} Sentence
Take the given query and rewrite it as a {*sentence_type*} sentence, changing only the sentence type while keeping the original meaning intact.
Do not add extra details or change the intent.

Grammar Error Variation. This prompt injects at least one grammar error to the original query.

Grammar Error Variation Prompt Template

You are an AI assistant skilled at generating realistic grammatical errors in queries.
Rewrite the given query by introducing grammatical errors.
Constraints:
- Preserve the original meaning
- Introduce at least three grammatical errors per query
- Make it realistic and natural by varying the types of errors

C Benchmark Construction

C.1 Model-generated Queries

Style and sentence-type variations are generated using GPT-5-mini[2] with the prompts provided in Appendix B. For the error categories, we use external tools to inject or verify errors. Spelling errors are injected using the `SpellingAug` function in `nlpaug`. For the grammar error category, we use LanguageTool[3] to verify that each query contains at least one grammatical error. If a query is classified as having no grammar errors, we regenerate it using GPT-5-mini to inject a grammatical error.

For all query variations, the co-authors manually inspected sampled queries. The annotation process was conducted by two co-authors, both fluent in English. Annotators were instructed to focus on the quality of the rewritten queries, whether the intended perturbation was correctly applied while preserving the original meaning. We randomly sampled 100 examples across query variations, and the two annotators marked 96% and 98% as pass-

[2] gpt-5-mini-2025-08-07
[3] https://languagetool.org

Readability Variation Prompt Template

1. Task Definition:
You are rewriting a query to make it significantly less readable while preserving the original semantic meaning as closely as possible.
2. Constraints & Goals:

- Flesch Reading Ease Score: The rewritten text must have a Flesch score below 60 (preferably below 50).
- Semantic Similarity: The rewritten text must have SBERT similarity > 0.7 compared with the original query.
- Length: The rewritten text must remain approximately the same length as the original query (±10%).
- Preserve Domain Terminology: Do not remove or drastically change domain-specific words, abbreviations, or technical terms (e.g., "IRS," "distance," etc.).
- Abbreviation: Do not expand abbreviations unless the original query already used the expanded form.
- No New Information: You must not add additional details beyond what the original query states.
- Question Format: Retain the form of a question if the original is posed as a question.
3. How to Increase Complexity:

- Lexical Changes: Use advanced or academic synonyms only for common words. For domain or key terms (e.g., "distance," "IRS," "tax"), keep the original term or use a very close synonym if necessary to maintain meaning.
- Syntactic Complexity: Introduce passive voice, nominalizations, embedded clauses, and parenthetical or subordinate phrases. Ensure the sentence flow is more formal and convoluted without changing the core meaning.
- Redundancy & Formality: Employ circumlocution and excessively formal expressions (e.g., "due to the fact that" instead of "because") while avoiding any semantic drift.
- Dense, Indirect Construction: Favor longer phrases, indirect references, and wordiness. Avoid direct or simple phrasing.

Table 6: Prompt for lower-readability query generation.

ing, indicating that the rewrites are largely meaning-preserving and adhere well to the intended query categories.

C.2 Human Annotation Details

We recruited 12 volunteer annotators to produce human query rewrites. Annotation guidelines are provided in Table 7. All volunteers were fluent in English and between 20 and 40 years old; with 5 female and 7 male participants. All annotators were graduate students or postdoctoral researchers with research experience in NLP or closely related fields. They reported high familiarity with AI tools (e.g., LLM-based assistants) and used them frequently in everyday tasks. Participation was unpaid and voluntary. Prior to the annotation process, all annotators provided informed consent for their rewrites to be collected and used for research purposes.

D Robustness Results

In this section, we report results for all six datasets and model variants in our benchmark.

D.1 Answer Robustness

InAccuracy. Figure 9 and 10 show consistent degradation under spelling errors. In contrast, sentence-type variations (imperative and declarative) achieve performance comparable to the original, suggesting that these rewrites have little effect on accuracy. For human rewrites, Figures 11 and 12 show consistently lower performance across most datasets.

Answer Consistency. Tables 8 and 9 report model-generated answer consistency across all six datasets. Full results including human query variations are shown in Tables 10 and 11. Overall, human rewrites tend to yield answers that are less similar to those produced for the original queries.

D.2 Computation Robustness

We measure computation robustness in terms of the computational cost incurred by the LLM and the retriever. We report the actual number of calls for each method–dataset pair. Figures 13 and 14 visualize the number of LLM calls across query variations. Figures 15 and 16 report the corresponding retriever calls for each model. As dataset difficulty increases, the generation_based method adaptively increases both LLM and retriever calls.

Robustness across model-generated query variations is reported in Tables 12 and 13. Results including human queries are shown in Tables 14 and 15. Overall, increased query complexity and higher call counts are associated with lower computational robustness.

E Self-Knowledge Results

We visualize the under- and over-confidence results for model-generated queries in Figures 17 and 18. As dataset difficulty increases, the system exhibits more overconfident cases. Figures 19 and 20 report the results when including human rewrites as well.

F Human Query Analysis

In this section, we visualize human query features in terms of loss, length, and semantic similarity. As shown in Figure 22, human queries are consistently shorter than other query variations. Consistent with their shorter length, Figure 21 shows that human, readability, spelling, and grammar-error variations tend to have higher loss. Lastly, semantic similarity to the original query shows a similar trend across datasets, indicating that human queries preserve the original query's meaning adequately.

G Usage of GenAI

We used AI-assisted coding tools to support data analysis and visualization (e.g., drafting and debugging scripts for metric aggregation and figure/table generation). All code and reported results were reviewed and validated by the authors.

Human Annotation Instruction

Task Overview

1. You will be given a query (question) and its corresponding answer.
2. Your task is to rewrite the given query — imagine how you would ask the same question if you were speaking to a large language model (like ChatGPT).

What You Should Do

1. Understand the original query.
   - Identify what the user is trying to find out.
   - Grasp the intent and main topic clearly.
2. Rewrite the query in a new way while keeping the same meaning.
   - You may change the tone, phrasing, or structure.
   - Information loss, addition, or simplification is acceptable.
   - The goal is to make the rewritten query similar in meaning but different in expression.
   - Make sure that your rewritten query can still be answered appropriately with the same answer as the original query.
3. Incorporate your personal style.
   - You can make the question sound more casual, detailed, concise, or clear.
   - Different writing styles and tones are encouraged.
4. Avoid copying the original query.
   - Do not simply rephrase it with only minor surface changes.
   - The rewritten query should feel genuinely reworded and natural.
5. Use the answer only for context.
   - The answer helps you understand what the original query intended to ask.
   - You should not change your rewrite based on the answer's content.
   - Your task is to rewrite the question, not the answer.

Table 7: Human Annotation Instruction: Query Rewriting Task.

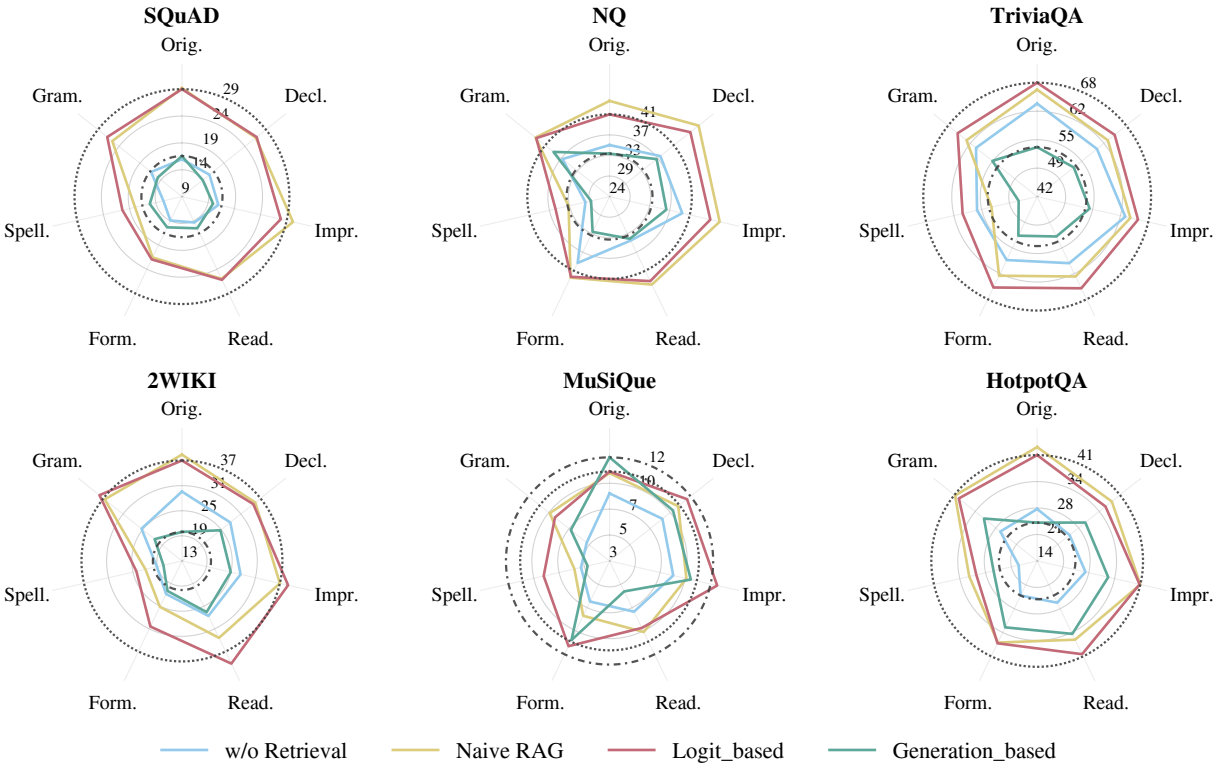


Figure 9: **InAccuracy (Llama-3.1-8B).** Performance across datasets for four Adaptive RAG methods. Each panel reports InAccuracy on seven query variations: the original query and six model-generated perturbations.

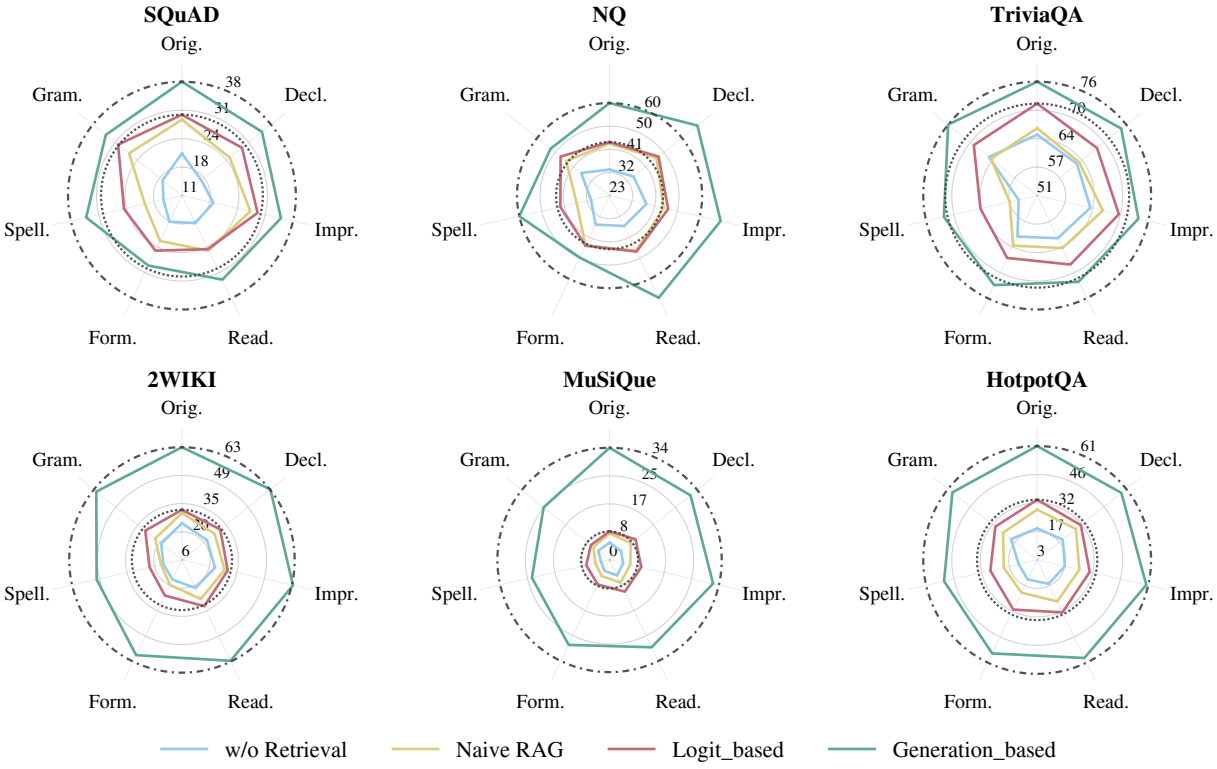


Figure 10: **InAccuracy (QwQ-32B).** Performance across datasets for four Adaptive RAG methods. Each panel reports InAccuracy on seven query variations: the original query and six model-generated perturbations).

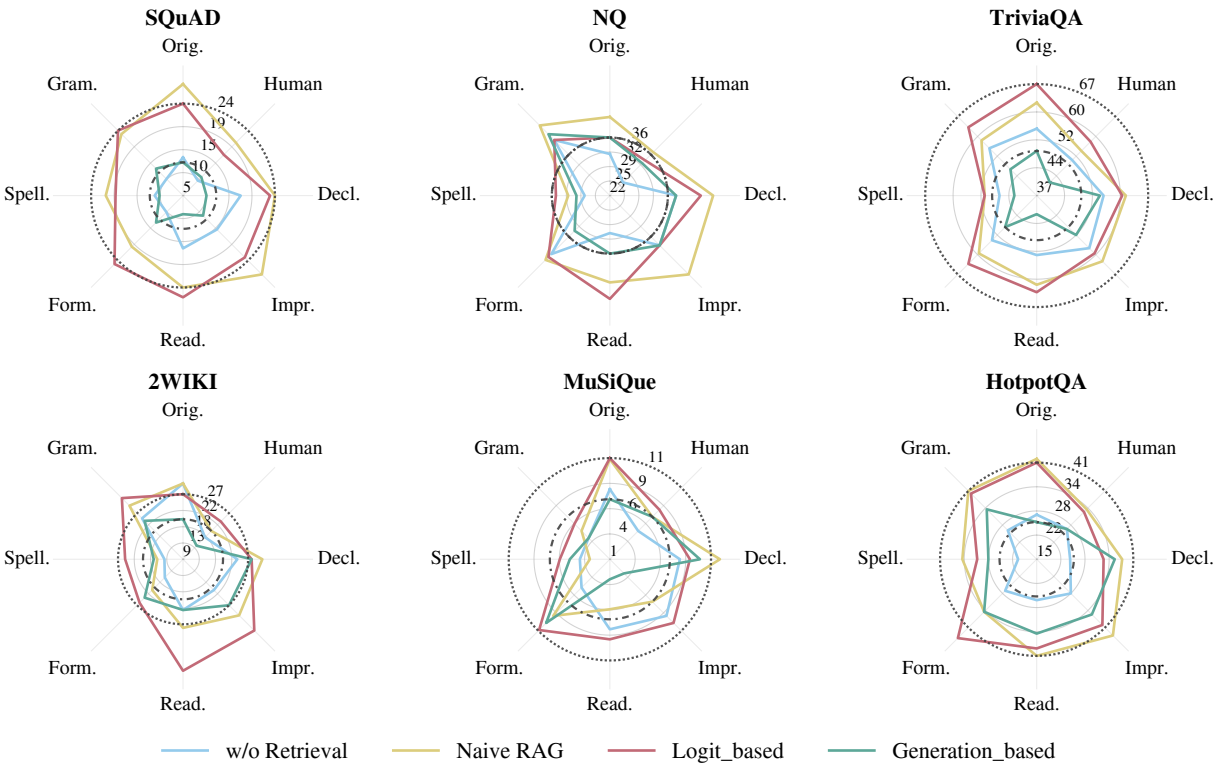


Figure 11: **InAccuracy (Llama-3.1-8B) including human rewrites.** Performance across datasets for four Adaptive RAG methods, evaluated on a matched subset of instances that have human rewrites. For each panel, InAccuracy is reported for eight query variations (original, human rewrite, and six model-generated perturbations), all computed on the same human-rewrite subset.

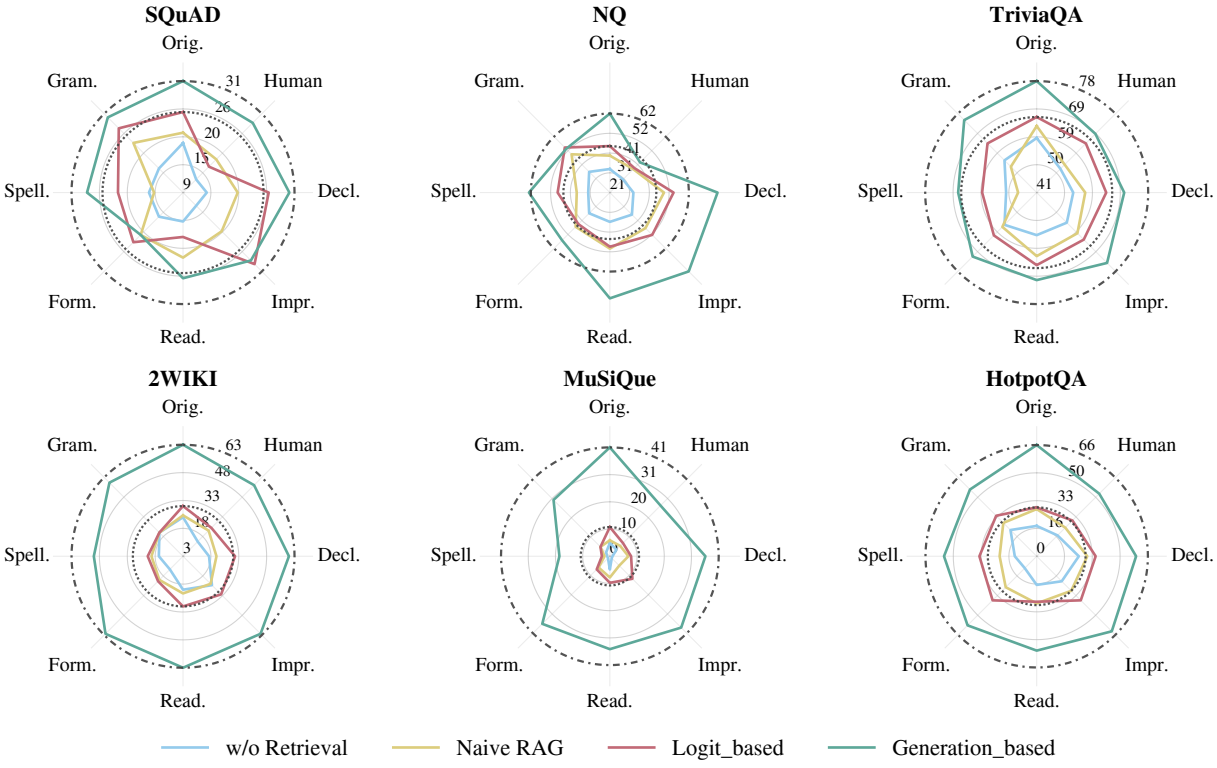


Figure 12: **InAccuracy (QwQ-32B) including human rewrites.** Performance across datasets for four Adaptive RAG methods, evaluated on a matched subset of instances that have human rewrites. For each panel, InAccuracy is reported for eight query variations (original, human rewrite, and six model-generated perturbations), all computed on the same human-rewrite subset.

Dataset	Method	Similarity ↑						Diversity ↓
		Form.	Read.	Decl.	Impr.	Spell.	Gram.	
Llama-3.1-8B								
SQuAD	WR	0.568	0.580	0.598	**0.635**	0.538	0.584	**0.459**
	NR	0.509	0.552	0.606	**0.631**	0.466	0.578	**0.499**
	LB	0.551	0.539	0.593	**0.634**	0.552	0.599	**0.469**
	GB	0.426	0.408	**0.433**	0.430	0.411	0.414	**0.585**
NQ	WR	0.672	0.670	0.652	**0.712**	0.610	0.675	**0.373**
	NR	0.311	0.483	**0.499**	0.494	0.425	0.321	**0.621**
	LB	0.706	0.686	0.708	**0.729**	0.626	0.724	**0.346**
	GB	0.382	**1.000**	**1.000**	**1.000**	**1.000**	0.435	**0.347**
TriviaQA	WR	0.767	0.773	0.762	**0.787**	0.734	0.763	**0.268**
	NR	0.700	0.716	0.725	**0.758**	0.648	0.738	**0.327**
	LB	0.781	**0.799**	0.769	0.798	0.727	0.785	**0.259**
	GB	0.528	0.537	0.543	0.547	0.504	**0.549**	**0.467**
2WIKI	WR	0.561	0.595	0.630	**0.660**	0.520	0.593	**0.442**
	NR	0.568	0.628	0.644	**0.651**	0.530	0.641	**0.432**
	LB	0.560	0.628	0.649	**0.655**	0.537	0.627	**0.414**
	GB	0.423	**0.458**	0.300	0.403	0.313	0.385	**0.667**
HotpotQA	WR	0.536	0.566	0.547	**0.580**	0.509	0.557	**0.484**
	NR	0.532	0.542	0.546	0.543	0.500	**0.564**	**0.500**
	LB	0.594	**0.628**	0.591	0.625	0.557	0.617	**0.415**
	GB	0.479	**0.502**	0.481	0.483	0.456	0.489	**0.523**
MuSiQue	WR	0.469	0.489	0.491	**0.512**	0.423	0.480	**0.560**
	NR	0.427	0.454	**0.471**	0.459	0.385	0.452	**0.588**
	LB	0.513	0.480	0.502	**0.528**	0.428	0.481	**0.537**
	GB	0.424	0.426	0.419	**0.438**	0.389	0.392	**0.576**

Table 8: **Answer similarity to the original query and diversity across perturbation types (excluding human rewrites).**We compare w/o retrieval (WR), naive RAG (NR), logit-based (LB), and generation-based (GB). Darker shading indicates more desirable outcomes: higher Similarity and lower Diversity.

Dataset	Method	Similarity ↑						Diversity ↓
		Form.	Read.	Decl.	Impr.	Spell.	Gram.	
				QwQ-32B				
SQuAD	WR	0.566	0.576	0.602	**0.634**	0.549	0.593	**0.438**
	NR	0.511	0.570	0.625	**0.649**	0.488	0.599	**0.475**
	LB	0.573	0.579	0.615	**0.631**	0.544	0.584	**0.440**
	GB	0.552	0.558	0.579	**0.589**	0.552	0.579	**0.459**
NQ	WR	0.660	0.634	0.671	**0.678**	0.583	0.673	**0.373**
	NR	0.367	0.518	0.375	**0.520**	0.483	0.389	**0.566**
	LB	0.344	0.304	0.321	0.309	0.305	**0.353**	**0.662**
	GB	0.438	0.769	0.783	**0.797**	0.739	0.442	**0.428**
TriviaQA	WR	0.743	0.735	**0.759**	0.759	0.699	0.752	**0.288**
	NR	0.698	0.738	0.735	**0.759**	0.671	0.747	**0.309**
	LB	0.770	0.775	0.785	0.779	0.730	**0.788**	**0.255**
	GB	0.608	0.588	0.603	**0.626**	0.580	0.604	**0.395**
2WIKI	WR	0.528	0.554	0.589	**0.591**	0.481	0.483	**0.531**
	NR	0.443	0.500	0.520	**0.531**	0.422	0.514	**0.547**
	LB	0.500	0.545	0.551	**0.565**	0.462	0.519	**0.511**
	GB	0.520	0.557	0.554	**0.559**	0.499	0.548	**0.465**
HotpotQA	WR	0.359	**0.402**	0.385	0.401	0.333	0.394	**0.639**
	NR	0.377	**0.414**	0.408	0.401	0.357	0.397	**0.624**
	LB	0.444	0.452	**0.468**	0.458	0.406	0.444	**0.568**
	GB	0.582	0.585	0.581	**0.608**	0.563	0.580	**0.429**
MuSiQue	WR	0.394	0.390	**0.408**	0.405	0.371	0.408	**0.629**
	NR	0.387	0.377	0.396	**0.405**	0.374	0.383	**0.634**
	LB	0.395	0.400	0.414	**0.422**	0.380	0.404	**0.610**
	GB	0.593	**0.607**	0.606	0.607	0.589	0.582	**0.422**

Table 9: **Answer similarity to the original query and diversity across perturbation types (excluding human rewrites).** We compare w/o retrieval (WR), naive RAG (NR), logit-based (LB), and generation-based (GB). Darker shading indicates more desirable outcomes: higher Similarity and lower Diversity.

Dataset	Method	Similarity ↑							Diversity ↓
		Human	Form.	Read.	Decl.	Impr.	Spell.	Gram.	
					Llama-3.1-8B				
SQuAD	WR	0.542	0.579	0.610	0.644	**0.653**	0.542	0.600	**0.446**
	NR	0.505	0.580	0.544	0.613	**0.622**	0.479	0.607	**0.503**
	LB	0.527	0.557	0.557	0.599	**0.626**	0.566	0.584	**0.469**
	GB	0.447	0.416	0.421	0.423	**0.449**	0.427	0.449	**0.583**
NQ	WR	0.624	0.698	0.646	0.683	**0.748**	0.586	0.692	**0.385**
	NR	0.292	0.300	0.470	**0.507**	0.500	0.385	0.319	**0.616**
	LB	0.642	0.674	0.662	0.661	0.686	0.577	**0.691**	**0.370**
	GB	0.356	0.382	**1.000**	**1.000**	**1.000**	**1.000**	0.435	**0.416**
TriviaQA	WR	0.660	0.734	0.753	0.741	**0.755**	0.672	0.700	**0.331**
	NR	0.634	0.673	0.691	0.690	**0.743**	0.635	0.725	**0.370**
	LB	0.697	0.735	**0.776**	0.718	0.729	0.708	0.735	**0.314**
	GB	0.475	0.499	0.458	0.541	**0.555**	0.460	0.538	**0.517**
2WIKI	WR	0.540	0.516	0.567	0.628	**0.628**	0.511	0.607	**0.480**
	NR	0.488	0.516	0.585	**0.641**	0.592	0.453	0.615	**0.473**
	LB	0.550	0.541	**0.617**	0.616	0.610	0.502	0.590	**0.451**
	GB	0.279	0.423	**0.458**	0.300	0.403	0.313	0.385	**0.647**
HotpotQA	WR	0.514	0.544	0.545	0.555	**0.592**	0.511	0.565	**0.482**
	NR	0.524	0.568	0.578	0.574	0.583	0.520	**0.616**	**0.465**
	LB	0.543	0.630	0.617	0.586	0.634	0.588	**0.652**	**0.401**
	GB	0.462	0.477	0.484	0.452	0.465	0.489	**0.515**	**0.507**
MuSiQue	WR	0.459	0.417	0.511	0.471	**0.525**	0.403	0.454	**0.564**
	NR	0.407	0.415	**0.476**	0.433	0.452	0.382	0.400	**0.600**
	LB	0.504	0.528	**0.578**	0.486	0.576	0.436	0.487	**0.551**
	GB	0.360	0.423	**0.442**	0.380	0.415	0.372	0.370	**0.590**

Table 10: **Answer similarity to the original query and diversity across perturbation types (including human rewrites).** We compare w/o retrieval (WR), naive RAG (NR), logit-based (LB), and generation-based (GB). Darker shading indicates more desirable outcomes: higher Similarity and lower Diversity.

Dataset	Method	Similarity ↑							Diversity ↓
		Human	Form.	Read.	Decl.	Impr.	Spell.	Gram.	
					QwQ-32B				
SQuAD	WR	0.585	0.567	0.608	0.630	**0.642**	0.583	0.605	**0.448**
	NR	0.556	0.514	0.619	0.655	**0.664**	0.518	0.584	**0.459**
	LB	0.565	0.608	0.590	0.608	**0.635**	0.577	0.599	**0.435**
	GB	0.544	0.549	0.560	0.596	**0.601**	0.575	0.588	**0.461**
NQ	WR	0.594	0.631	0.612	0.649	0.635	0.573	**0.660**	**0.374**
	NR	0.370	0.358	**0.517**	0.374	0.510	0.461	0.374	**0.551**
	LB	0.330	0.348	0.294	0.319	0.346	0.300	**0.387**	**0.628**
	GB	0.438	0.438	0.769	0.783	**0.797**	0.739	0.442	**0.454**
TriviaQA	WR	0.684	**0.718**	0.678	0.696	0.705	0.661	0.687	**0.336**
	NR	0.646	0.649	0.683	0.687	**0.720**	0.597	0.664	**0.363**
	LB	0.697	0.730	0.712	**0.745**	0.693	0.665	0.717	**0.307**
	GB	0.548	0.551	0.531	0.539	**0.582**	0.546	0.558	**0.440**
2WIKI	WR	0.560	0.542	0.556	0.623	**0.634**	0.519	0.570	**0.496**
	NR	0.511	0.492	0.545	**0.579**	0.569	0.482	0.551	**0.516**
	LB	0.518	0.528	0.570	0.573	**0.612**	0.522	0.572	**0.486**
	GB	0.529	0.542	0.534	0.568	**0.598**	0.520	0.558	**0.463**
HotpotQA	WR	0.356	0.394	**0.455**	0.412	0.416	0.368	0.421	**0.636**
	NR	0.379	0.366	0.405	0.398	**0.412**	0.364	0.381	**0.621**
	LB	0.420	0.467	0.417	**0.486**	0.476	0.428	0.468	**0.560**
	GB	0.616	0.595	0.596	0.598	**0.629**	0.573	0.612	**0.426**
MuSiQue	WR	0.395	0.384	0.388	**0.411**	0.402	0.372	0.395	**0.633**
	NR	0.363	**0.427**	0.396	0.425	0.404	0.354	0.382	**0.627**
	LB	0.397	**0.416**	0.371	0.403	0.408	0.389	0.390	**0.619**
	GB	0.584	0.582	0.604	**0.627**	0.603	0.595	0.573	**0.424**

Table 11: **Answer similarity to the original query and diversity across perturbation types (including human rewrites).** We compare w/o retrieval (WR), naive RAG (NR), logit-based (LB), and generation-based (GB). Darker shading indicates more desirable outcomes: higher Similarity and lower Diversity.

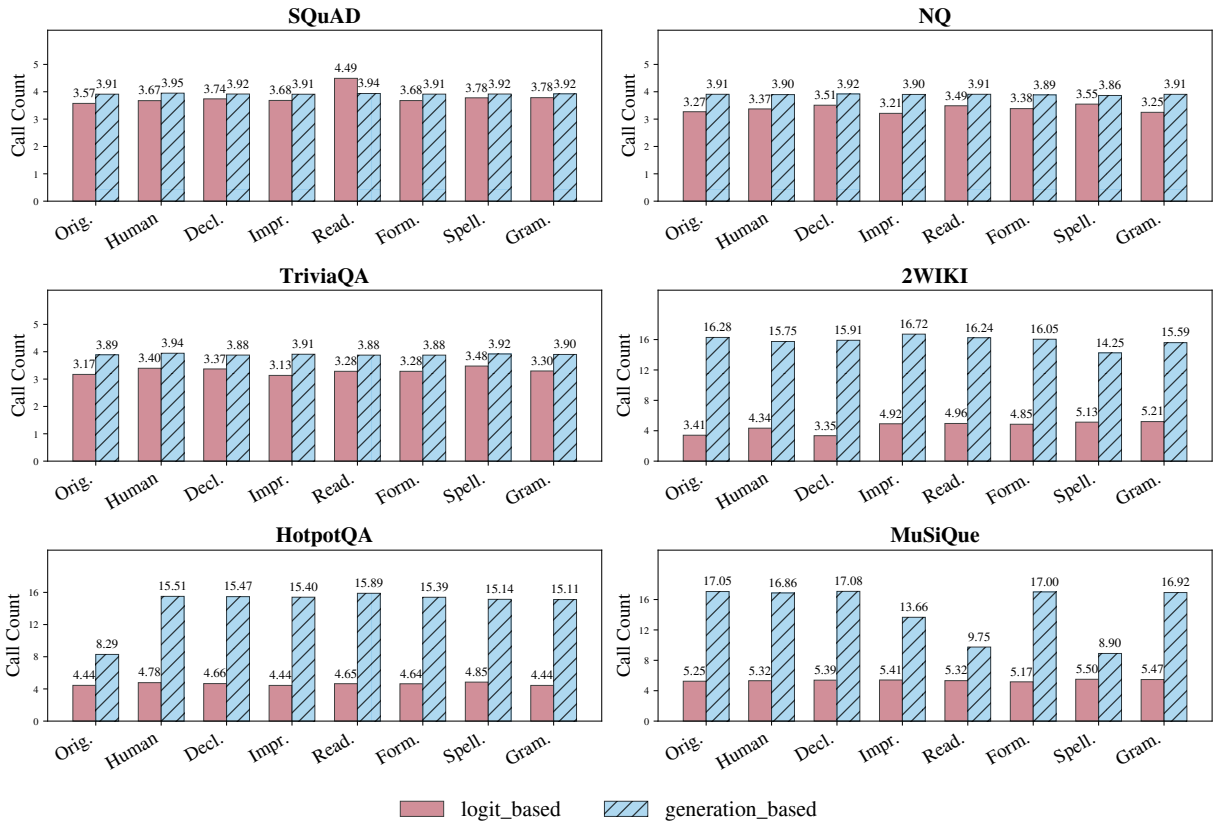


Figure 13: **LLM Call** on Llama-3.1-8B on both model-generated and human queries.

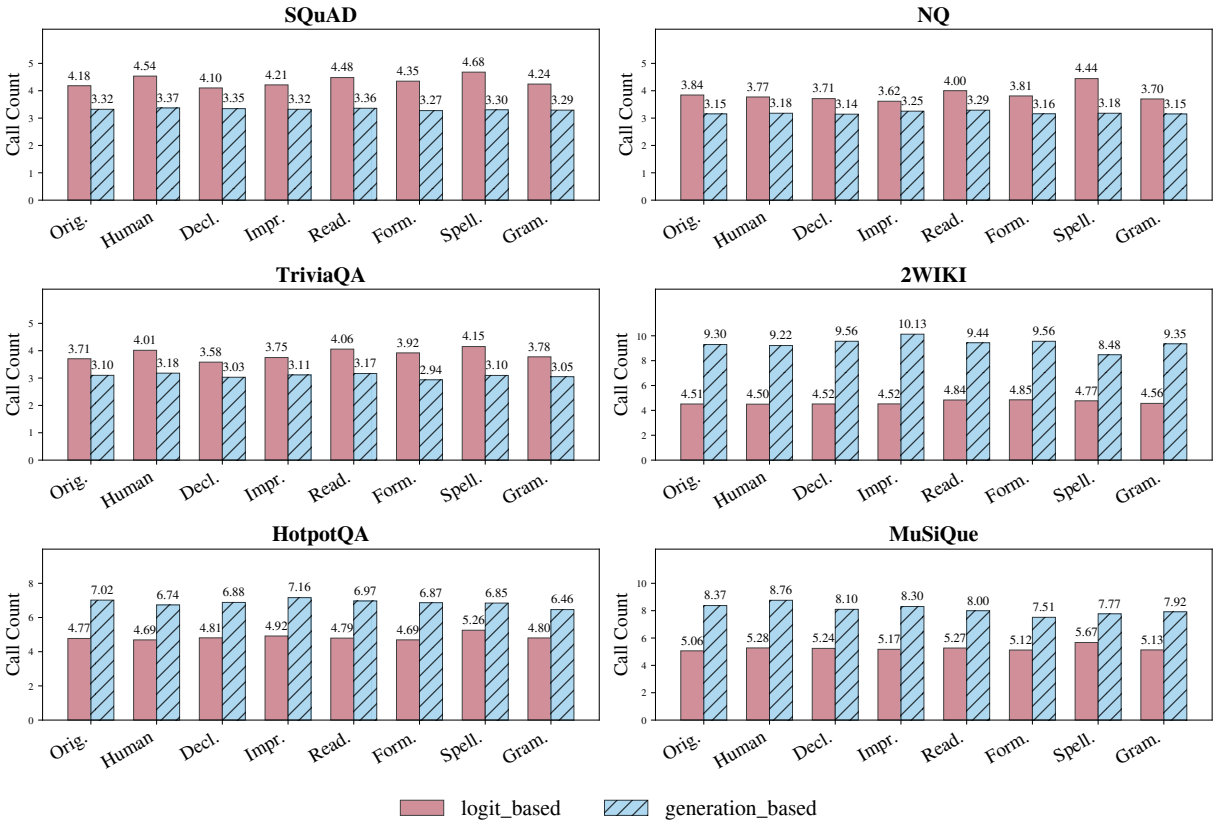


Figure 14: **LLM Call** on QwQ-32B on both model-generated and human queries.

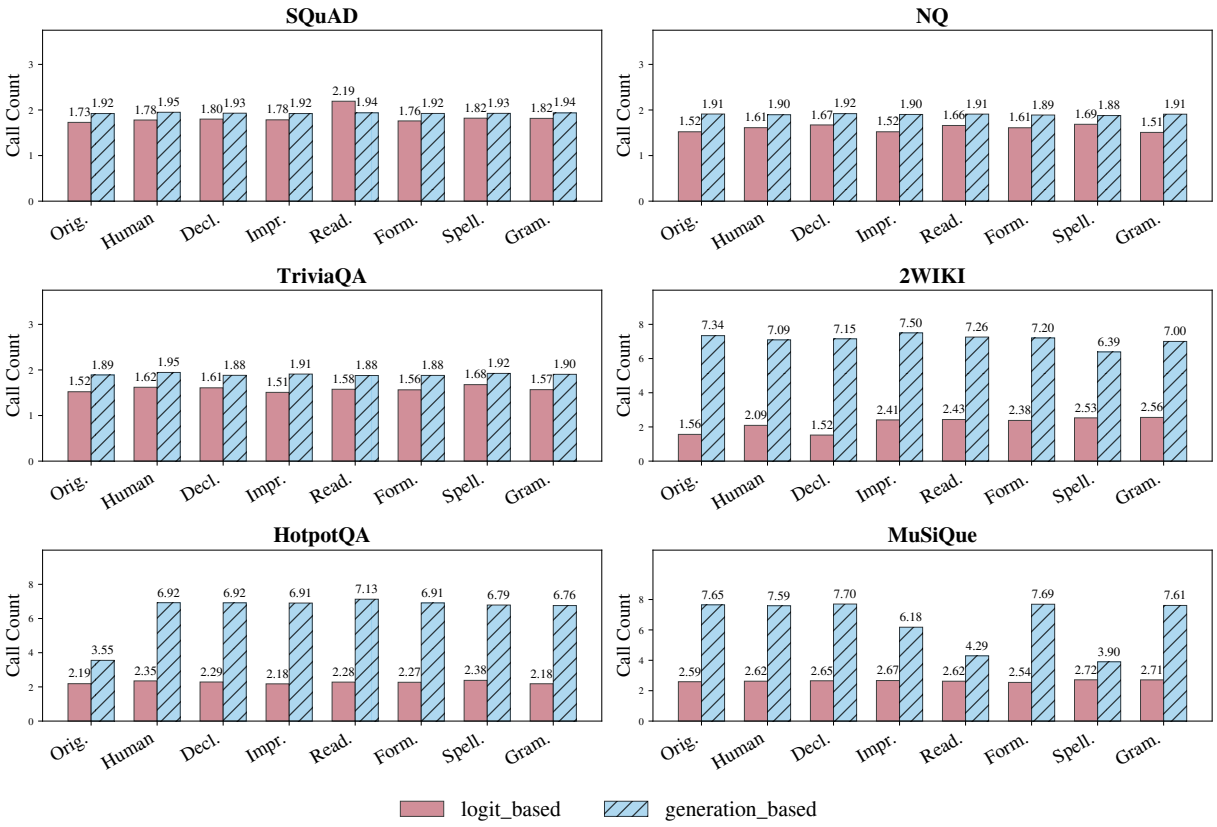


Figure 15: **Retriever Call** on Llama-3.1-8B on both model-generated and human queries.

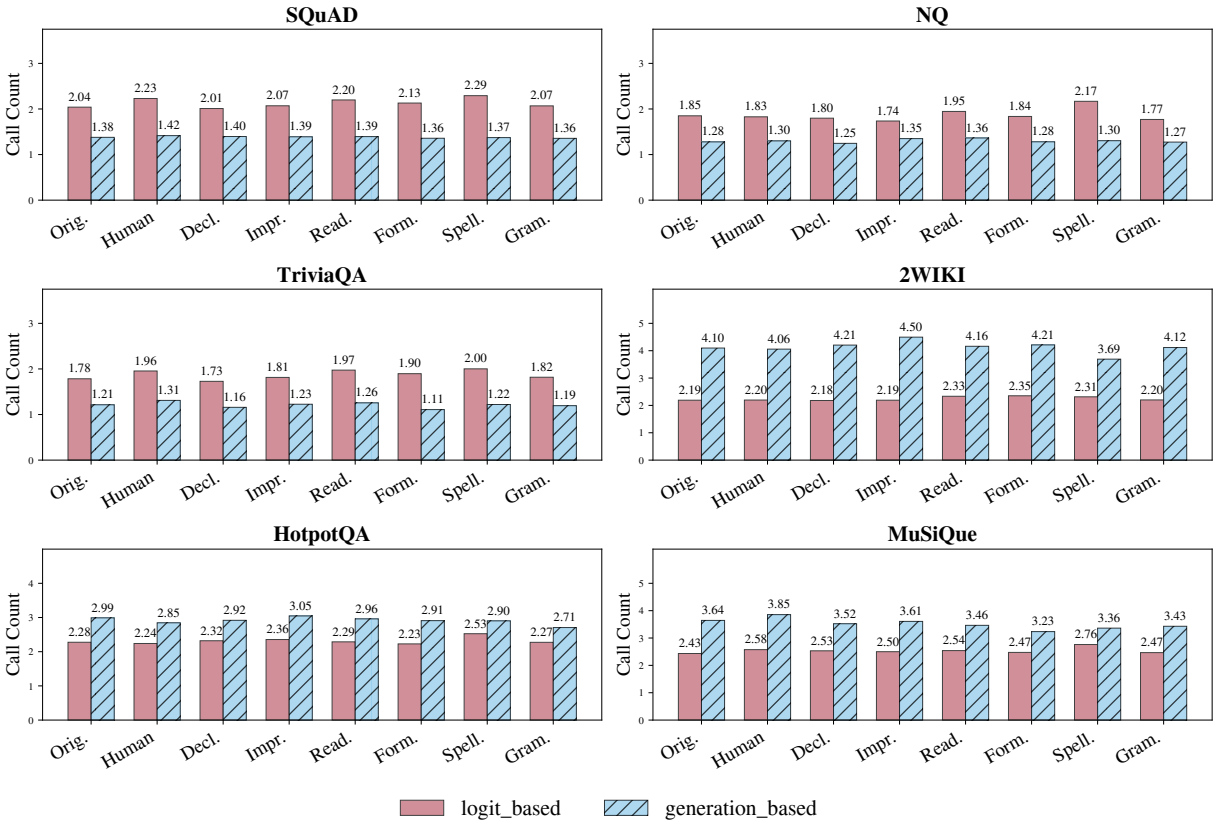


Figure 16: **Retriever Call** on QwQ-32B on both model-generated and human queries.

Dataset	Target	Method	RE ↓						CVR ↑
			Form.	Read.	Decl.	Impr.	Spell.	Gram.	
Llama-3.1-8B									
SQuAD	**Retriever**	LB	0.472	0.592	0.416	0.338	0.431	0.439	0.734
		GB	**0.092**	**0.094**	**0.086**	**0.090**	**0.100**	**0.092**	0.937
	LLM	LB	0.441	0.592	0.396	0.313	0.416	0.411	0.767
		GB	**0.056**	**0.059**	**0.048**	**0.055**	**0.061**	**0.051**	0.963
NQ	**Retriever**	LB	0.520	0.485	0.576	0.465	0.578	0.456	0.702
		GB	**0.096**	**0.000**	**0.000**	**0.000**	**0.000**	**0.096**	0.949
	LLM	LB	0.499	0.462	0.541	0.428	0.545	0.440	0.750
		GB	**0.038**	**0.000**	**0.000**	**0.000**	**0.000**	**0.042**	0.973
TriviaQA	**Retriever**	LB	0.452	0.442	0.493	0.436	0.536	0.471	0.712
		GB	**0.137**	**0.120**	**0.121**	**0.122**	**0.124**	**0.122**	0.917
	LLM	LB	0.463	0.420	0.472	0.410	0.527	0.466	0.747
		GB	**0.054**	**0.047**	**0.048**	**0.046**	**0.047**	**0.047**	0.956
2WIKI	**Retriever**	LB	0.853	0.859	0.455	0.846	0.928	0.941	0.710
		GB	**0.468**	**0.416**	**0.402**	**0.542**	**0.496**	**0.435**	0.742
	LLM	LB	1.001	1.001	0.427	1.007	1.093	1.130	0.742
		GB	**0.447**	**0.404**	**0.401**	**0.520**	**0.477**	**0.418**	0.745
HotpotQA	**Retriever**	LB	**0.486**	**0.461**	**0.428**	**0.440**	**0.487**	**0.428**	0.747
		GB	1.914	1.992	1.850	1.848	1.924	1.825	0.687
	LLM	LB	**0.490**	**0.439**	**0.421**	**0.427**	**0.502**	**0.418**	0.761
		GB	1.560	1.626	1.525	1.517	1.565	1.495	0.694
MuSiQue	**Retriever**	LB	**0.404**	**0.417**	**0.382**	**0.345**	**0.399**	**0.340**	0.753
		GB	0.521	0.525	0.533	0.446	0.574	0.510	0.691
	LLM	LB	**0.365**	**0.398**	**0.353**	**0.323**	**0.388**	**0.318**	0.763
		GB	0.502	0.502	0.519	0.429	0.550	0.484	0.698

Table 12: **Computation robustness across query variations (Llama-3.1-8B; no human rewrites).** Darker shading indicates more desirable outcomes: lower RE and higher CVR.

Dataset	Target	Method	RE ↓						CVR ↑
			Form.	Read.	Decl.	Impr.	Spell.	Gram.	
QwQ-32B									
SQuAD	**Retriever**	LB	0.494	0.450	0.405	0.415	0.495	0.441	0.740
		GB	**0.360**	**0.322**	**0.298**	**0.315**	**0.361**	**0.330**	0.762
	LLM	LB	0.477	0.466	0.389	0.399	0.491	0.424	0.760
		GB	**0.265**	**0.230**	**0.225**	**0.238**	**0.250**	**0.243**	0.858
NQ	**Retriever**	LB	0.551	0.586	0.525	0.513	0.673	0.520	0.689
		GB	**0.411**	**0.402**	**0.374**	**0.416**	**0.439**	**0.452**	0.721
	LLM	LB	0.546	0.598	0.511	0.504	0.679	0.516	0.724
		GB	**0.379**	**0.375**	**0.358**	**0.372**	**0.411**	**0.424**	0.820
TriviaQA	**Retriever**	LB	0.581	0.563	0.495	0.504	0.587	0.546	0.697
		GB	**0.362**	**0.377**	**0.347**	**0.374**	**0.349**	**0.387**	0.721
	LLM	LB	0.564	0.565	0.470	0.509	0.601	0.530	0.727
		GB	**0.329**	**0.373**	**0.331**	**0.348**	**0.324**	**0.348**	0.818
2WIKI	**Retriever**	LB	**0.560**	**0.502**	**0.446**	**0.443**	**0.565**	**0.511**	0.719
		GB	0.788	0.808	0.780	0.825	0.787	0.884	0.689
	LLM	LB	**0.547**	**0.481**	**0.426**	**0.430**	**0.565**	**0.493**	0.747
		GB	1.078	1.142	1.092	1.203	1.049	1.219	0.716
HotpotQA	**Retriever**	LB	**0.589**	**0.620**	**0.580**	**0.590**	**0.671**	**0.613**	0.707
		GB	0.806	0.774	0.788	0.818	0.894	0.663	0.654
	LLM	LB	**0.516**	**0.520**	**0.487**	**0.504**	**0.605**	**0.505**	0.743
		GB	0.968	0.920	0.902	1.000	1.078	0.778	0.695
MuSiQue	**Retriever**	LB	**0.632**	**0.660**	**0.598**	**0.607**	**0.702**	**0.603**	0.711
		GB	0.811	0.890	0.767	0.883	0.845	0.881	0.644
	LLM	LB	**0.538**	**0.572**	**0.537**	**0.515**	**0.607**	**0.523**	0.740
		GB	1.032	1.168	0.984	1.203	1.100	1.117	0.679

Table 13: **Computation robustness across query variations (QwQ-32B; no human rewrites).** Darker shading indicates more desirable outcomes: lower RE and higher CVR.

Dataset	Target	Method	RE ↓							CVR ↑
			Human	Form.	Read.	Decl.	Impr.	Spell.	Gram.	
Llama-3.1-8B										
SQuAD	**Retriever**	LB	0.415	0.452	0.577	0.423	0.344	0.460	0.450	0.748
		GB	**0.105**	**0.090**	**0.100**	**0.105**	**0.090**	**0.120**	**0.120**	0.946
	LLM	LB	0.407	0.418	0.587	0.398	0.306	0.414	0.414	0.778
		GB	**0.061**	**0.055**	**0.060**	**0.065**	**0.048**	**0.068**	**0.067**	0.969
NQ	**Retriever**	LB	0.549	0.495	0.487	0.562	0.481	0.551	0.425	0.696
		GB	**0.077**	**0.096**	**0.000**	**0.000**	**0.000**	**0.000**	**0.096**	0.942
	LLM	LB	0.542	0.485	0.475	0.546	0.410	0.498	0.427	0.739
		GB	**0.029**	**0.038**	**0.000**	**0.000**	**0.000**	**0.000**	**0.042**	0.969
TriviaQA	**Retriever**	LB	0.492	0.508	0.457	0.500	0.397	0.535	0.472	0.709
		GB	**0.065**	**0.095**	**0.105**	**0.090**	**0.080**	**0.090**	**0.095**	0.924
	LLM	LB	0.496	0.563	0.466	0.537	0.378	0.532	0.468	0.747
		GB	**0.024**	**0.039**	**0.044**	**0.041**	**0.030**	**0.037**	**0.041**	0.959
2WIKI	**Retriever**	LB	0.745	0.862	0.870	0.510	1.000	0.890	0.960	0.717
		GB	**0.506**	**0.468**	**0.416**	**0.402**	**0.542**	**0.496**	**0.435**	0.740
	LLM	LB	0.827	1.065	1.041	0.526	1.259	1.126	1.195	0.747
		GB	**0.480**	**0.447**	**0.404**	**0.401**	**0.520**	**0.477**	**0.418**	0.743
HotpotQA	**Retriever**	LB	**0.426**	**0.495**	**0.429**	**0.350**	**0.413**	**0.496**	**0.415**	0.747
		GB	1.962	1.895	1.928	1.627	1.697	1.805	1.832	0.684
	LLM	LB	**0.426**	**0.479**	**0.396**	**0.328**	**0.381**	**0.498**	**0.401**	0.757
		GB	1.668	1.524	1.556	1.326	1.389	1.481	1.527	0.693
MuSiQue	**Retriever**	LB	**0.284**	**0.275**	**0.377**	**0.448**	**0.343**	**0.296**	**0.219**	0.751
		GB	0.597	0.568	0.544	0.554	0.438	0.586	0.585	0.704
	LLM	LB	**0.268**	**0.262**	**0.331**	**0.410**	**0.307**	**0.274**	**0.221**	0.761
		GB	0.528	0.510	0.504	0.491	0.415	0.542	0.505	0.710

Table 14: **Computation robustness across query variations (Llama-3.1-8B; including human rewrites).** Darker shading indicates more desirable outcomes: lower RE and higher CVR.

Dataset	Target	Method	RE ↓							CVR ↑
			Human	Form.	Read.	Decl.	Impr.	Spell.	Gram.	
QwQ-32B										
SQuAD	**Retriever**	LB	0.439	0.500	0.568	0.446	0.404	0.451	0.547	0.741
		GB	**0.370**	**0.295**	**0.280**	**0.310**	**0.240**	**0.335**	**0.290**	0.766
	LLM	LB	0.461	0.468	0.597	0.426	0.389	0.417	0.494	0.757
		GB	**0.285**	**0.252**	**0.233**	**0.242**	**0.210**	**0.273**	**0.242**	0.859
NQ	**Retriever**	LB	0.565	0.485	0.462	0.533	0.558	0.662	0.537	0.692
		GB	**0.404**	**0.411**	**0.402**	**0.374**	**0.416**	**0.439**	**0.452**	0.715
	LLM	LB	0.554	0.456	0.492	0.510	0.555	0.669	0.533	0.725
		GB	**0.370**	**0.379**	**0.375**	**0.358**	**0.372**	**0.411**	**0.424**	0.816
TriviaQA	**Retriever**	LB	0.488	0.510	0.438	0.538	0.514	0.600	0.557	0.700
		GB	**0.455**	**0.405**	**0.420**	**0.430**	**0.395**	**0.440**	**0.440**	0.716
	LLM	LB	0.473	0.459	**0.438**	0.483	0.530	0.590	0.486	0.727
		GB	**0.472**	**0.417**	0.499	**0.481**	**0.407**	**0.448**	**0.448**	0.811
2WIKI	**Retriever**	LB	**0.407**	**0.531**	**0.486**	**0.461**	**0.437**	**0.546**	**0.510**	0.720
		GB	0.746	0.686	0.643	0.760	0.757	0.751	0.794	0.679
	LLM	LB	**0.437**	**0.506**	**0.480**	**0.459**	**0.425**	**0.548**	**0.502**	0.749
		GB	0.911	0.939	0.942	1.083	1.130	0.996	1.135	0.708
HotpotQA	**Retriever**	LB	**0.500**	**0.497**	**0.552**	**0.508**	**0.451**	**0.570**	**0.465**	0.706
		GB	1.023	0.928	0.712	0.910	0.901	0.877	0.716	0.632
	LLM	LB	**0.446**	**0.462**	**0.448**	**0.440**	**0.411**	**0.575**	**0.392**	0.746
		GB	1.146	1.092	0.776	0.976	1.105	1.083	0.874	0.678
MuSiQue	**Retriever**	LB	**0.652**	**0.629**	**0.657**	0.602	0.689	**0.655**	**0.602**	0.719
		GB	0.770	0.797	0.767	**0.548**	**0.642**	0.713	0.709	0.645
	LLM	LB	**0.557**	**0.501**	**0.537**	**0.512**	**0.579**	**0.570**	**0.510**	0.745
		GB	0.888	0.908	0.916	0.584	0.718	0.767	0.719	0.679

Table 15: **Computation robustness across query variations (QwQ-32B; including human rewrites).** Darker shading indicates more desirable outcomes: lower RE and higher CVR.

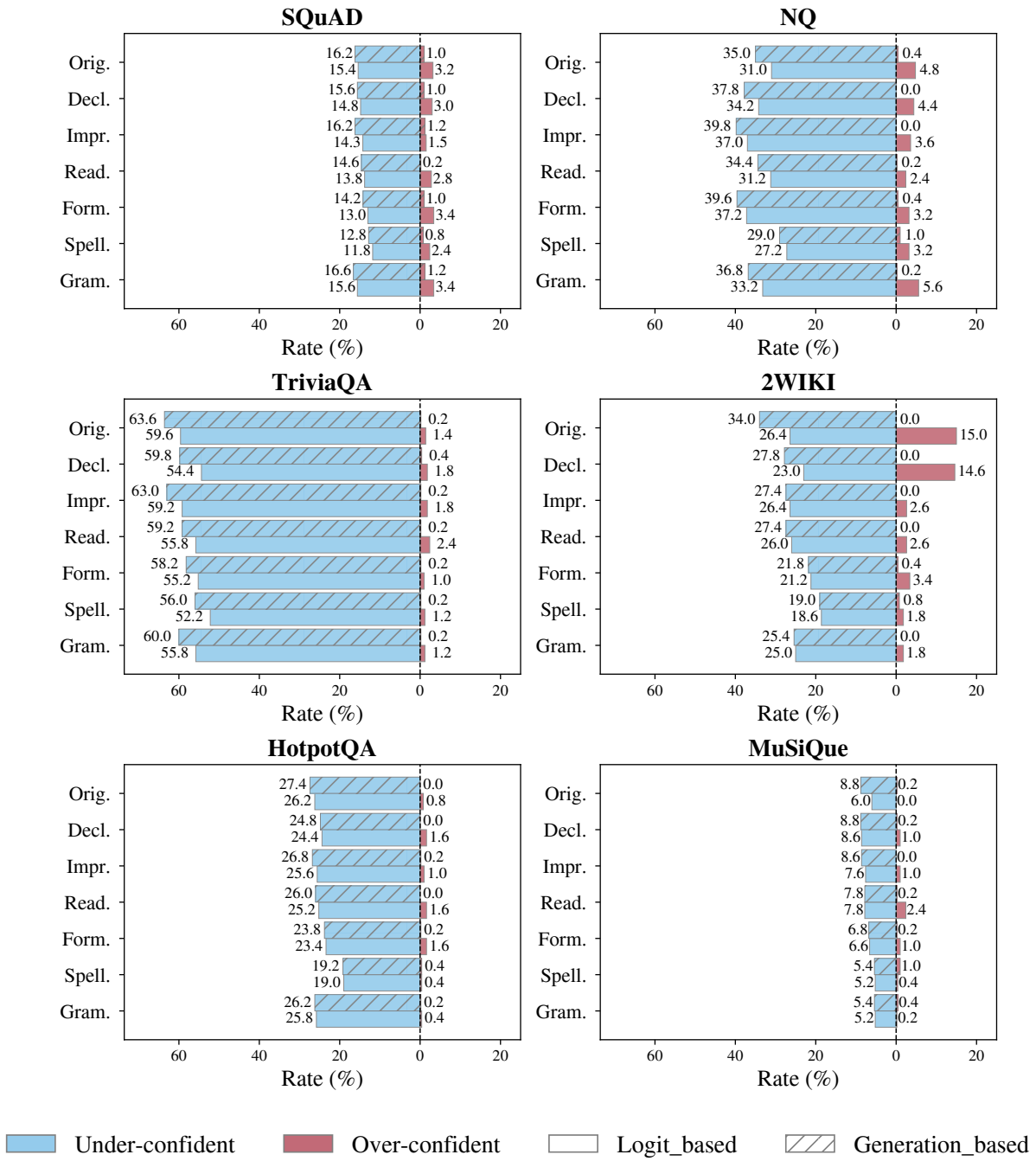


Figure 17: **Under- and Over-confidence results**on Llama-3.1-8B on model-generated queries.

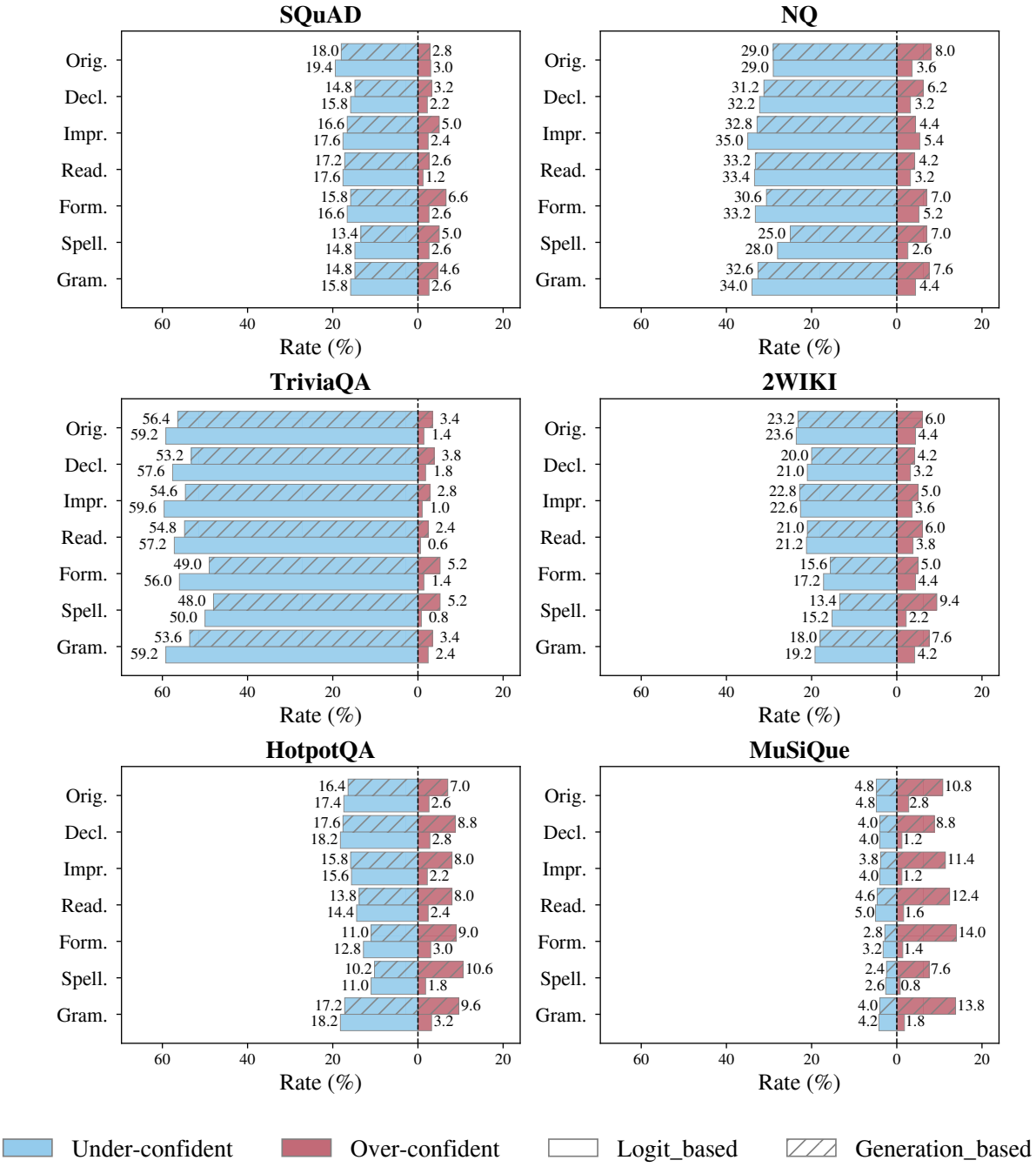


Figure 18: **Under- and Over-confidence results**on QwQ-32B on model-generated queries.

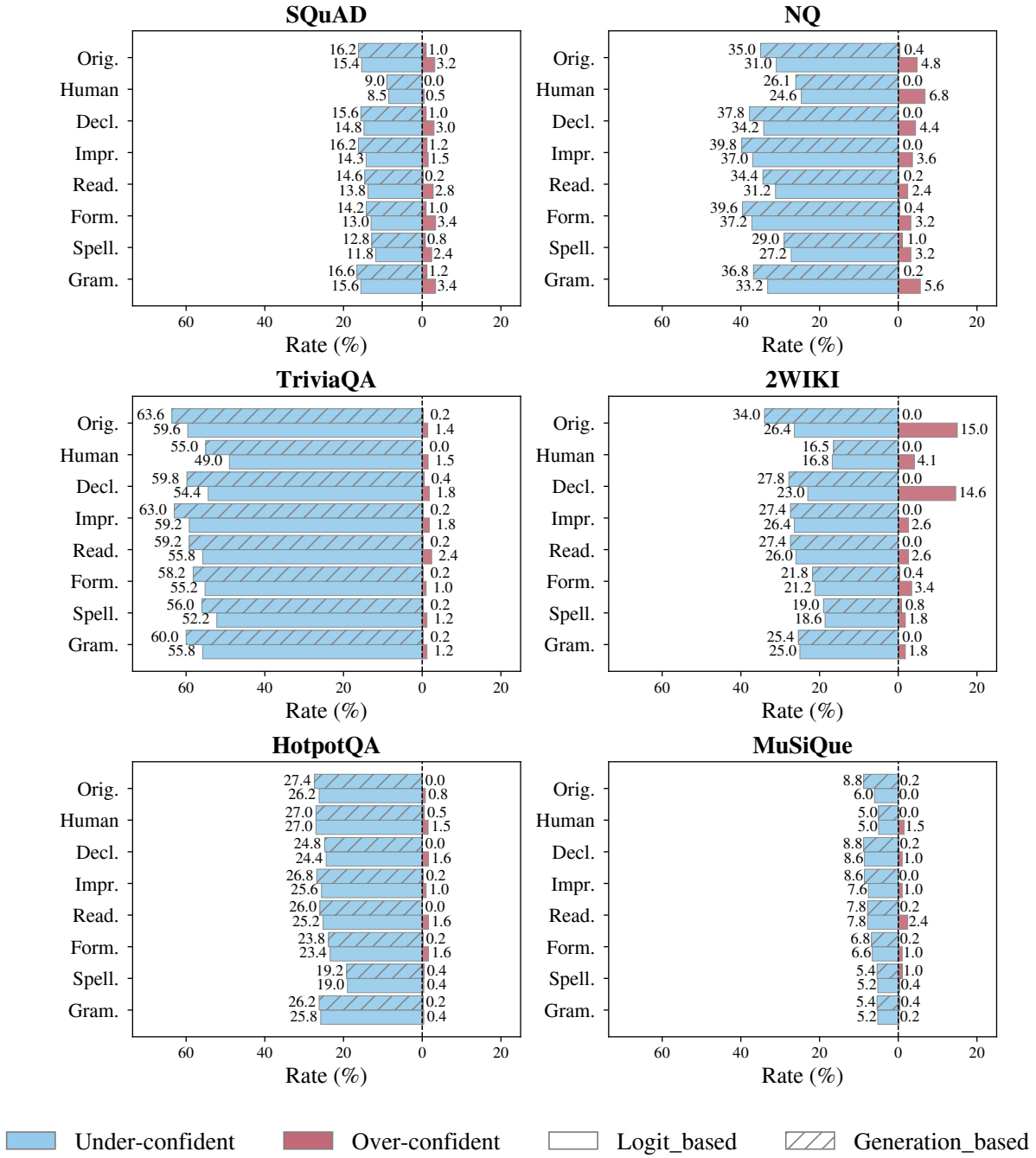


Figure 19: **Under- and Over-confidence results**on Llama-3.1-8B on including human-written queries.

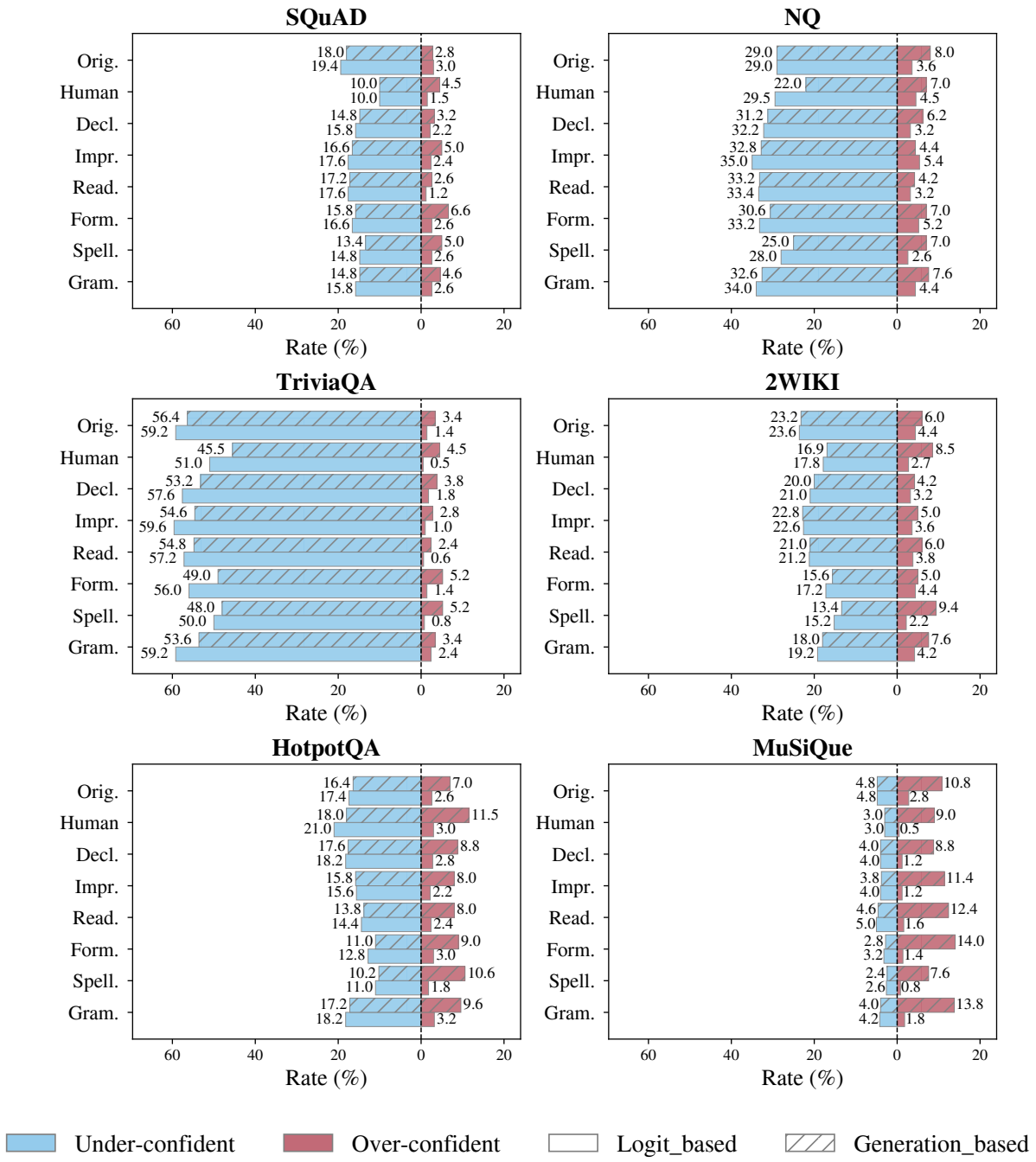


Figure 20: **Under- and Over-confidence results** on Llama-3.1-8B on including human-written queries.

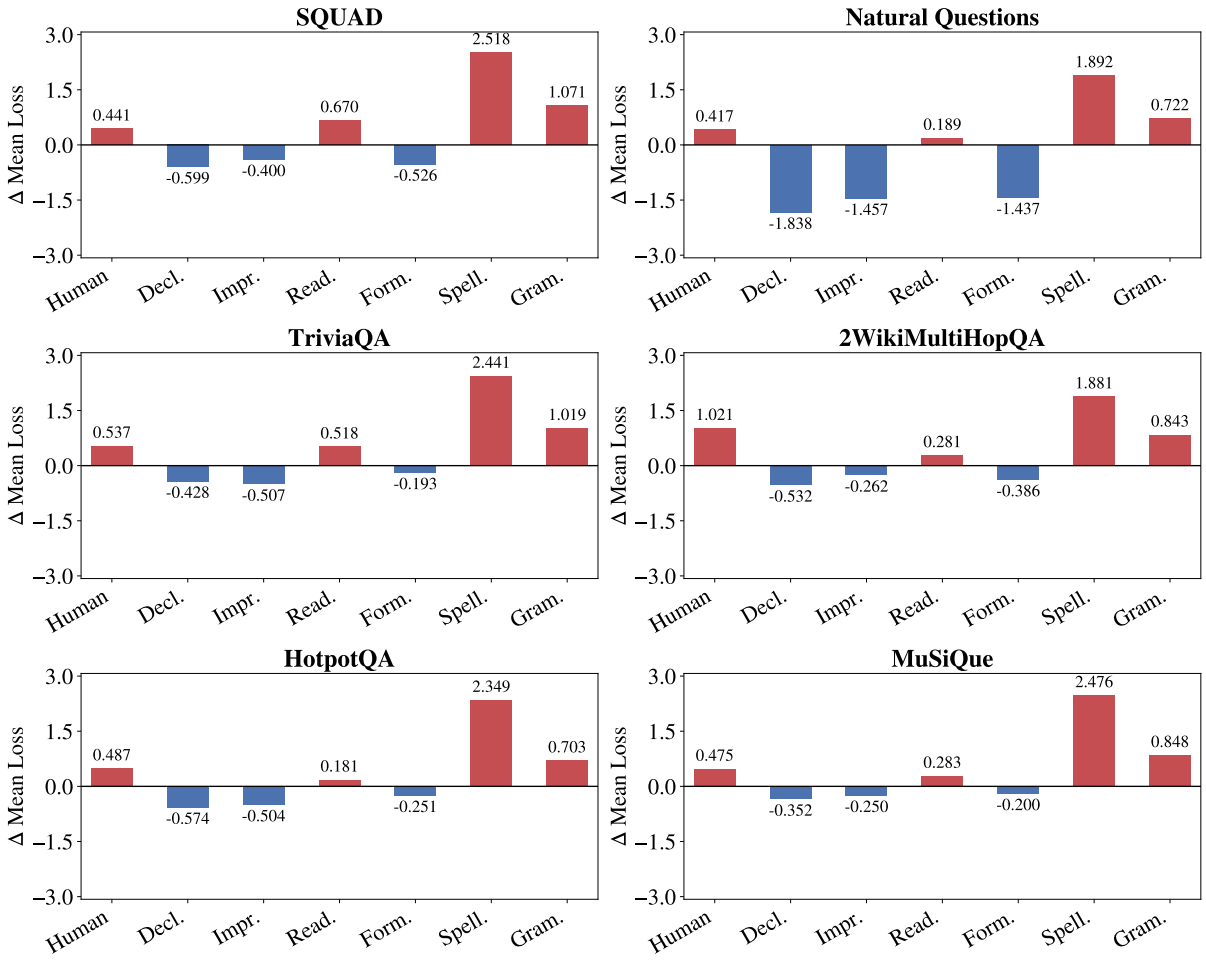


Figure 21: Loss score across all datasets

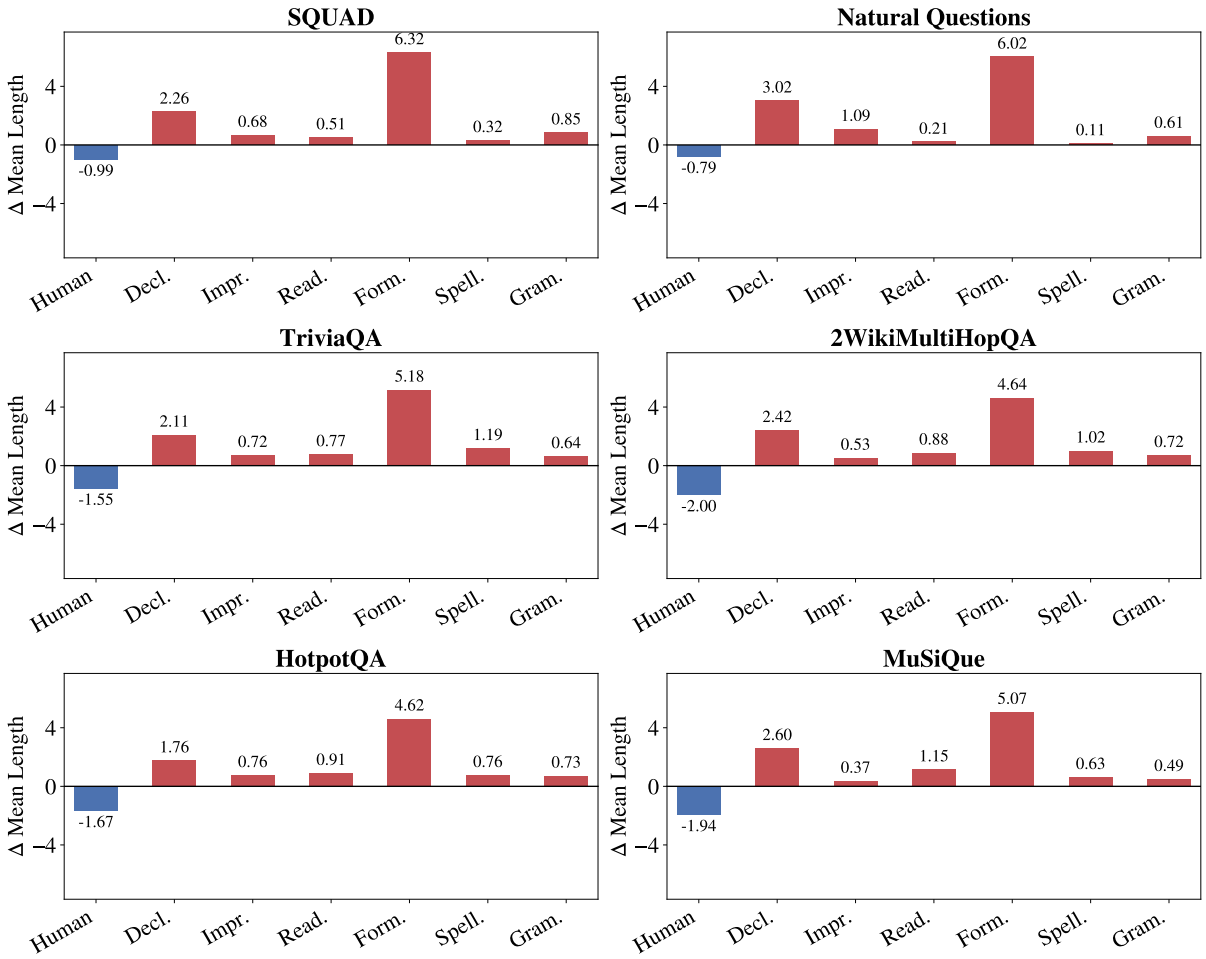


Figure 22: Query length across all datasets

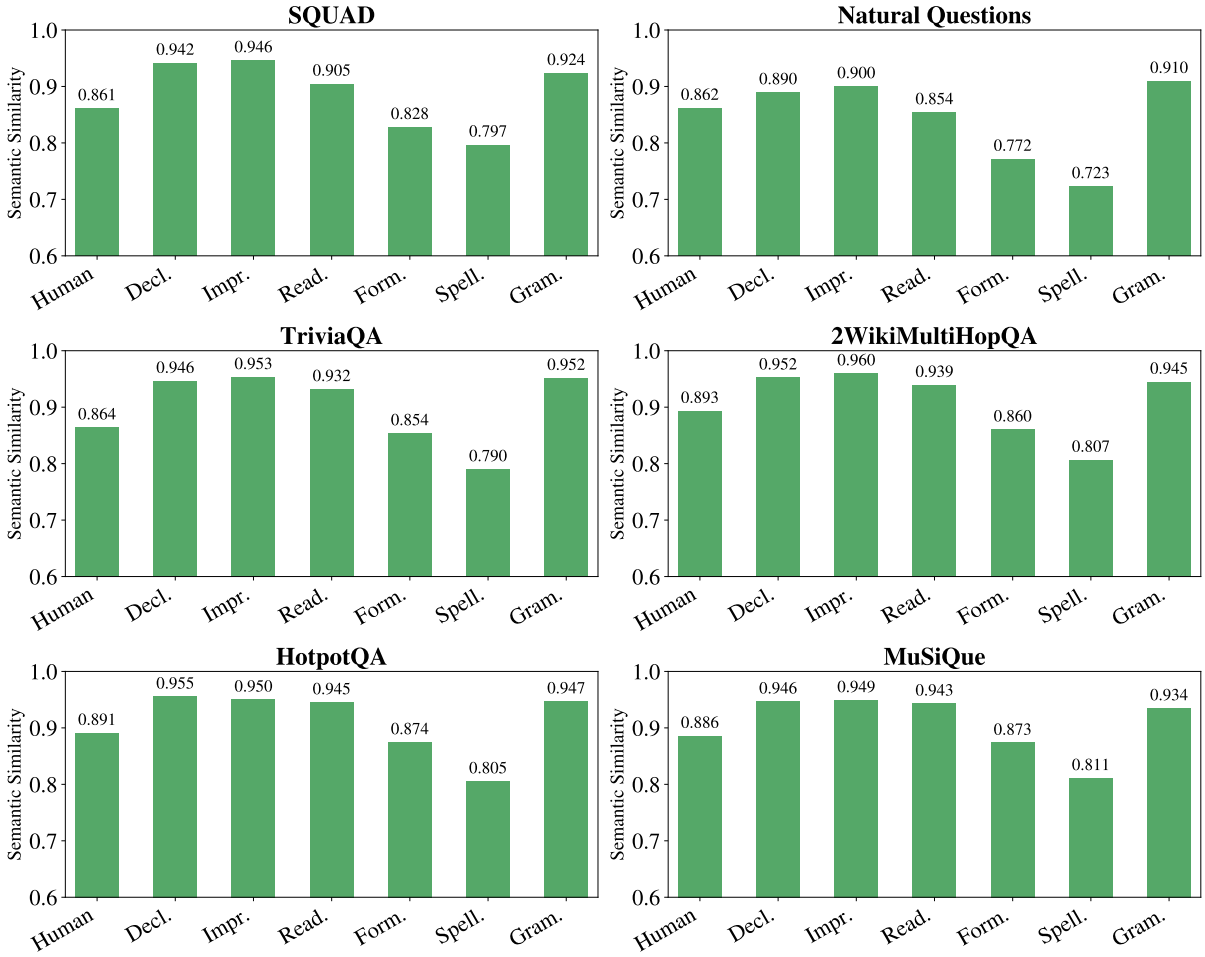


Figure 23: Semantic similarity to original query across all datasets