JOURNAL OF ARTIFICIAL INTELLIGENCE AND COMPUTATIONAL THEORY
JAICT

Structural Limits of Guardrail Optimization in Resource-Constrained RAG

AMIRMASOUD MOHAMMADIAN[1], SHAHLA NEMATI[2,*]

[1]Computer Engineering Department, Faculty of Engineering, Shahrekord University, Shahrekord, Iran
[2]Computer Engineering Department, Faculty of Engineering, Shahrekord University, Shahrekord, Iran

aammiirrrmasoud@gmail.com
s.nemati@sku.ac.ir

Article Information

Received: February 7, 2022
Revised: March 31, 2022
Accepted: May 8, 2022

Keywords

Retrieval-Augmented Generation, Question Answering, Language Models, Hallucination, Small Language Models

*Corresponding Author: Shahla Nemati

Abstract

Retrieval-Augmented Generation (RAG) grounds language model responses in retrieved documents, but effective retrieval does not necessarily ensure sufficient evidence to answer a query. This study investigates the limits of retrieval-threshold optimization and prompt engineering in a resource-constrained, closed-domain RAG system for question answering over the complete Harry Potter book corpus. The system combines dense semantic retrieval, a similarity-based guardrail, and instruction-tuned Qwen2.5 models. Eight Rule Set configurations were evaluated on a 175-question benchmark comprising 100 in-domain, 25 adjacent, and 50 out-of-domain questions. Results reveal a persistent trade-off between accepting relevant queries and rejecting unsupported ones: in-domain guardrail acceptance increased from 38% to 82%, while out-of-domain blocking decreased from 94.7% to 65.3%. Prompt refinements improved lexical-overlap metrics but had limited impact on adjacent questions, with answer rates remaining approximately 56–64%. Exact McNemar tests showed that increasing model size from 0.5B to 1.5B parameters significantly improved appropriate refusal of non-book-content questions (69.3% to 80.0%, $p = 0.0078$), but reduced human-evaluated in-domain answer quality (17% to 6%, $p = 0.0034$). These findings show that semantic relevance alone is insufficient to establish answerability and that retrieval and generation interventions address distinct failure modes, motivating explicit answerability or evidence-sufficiency mechanisms in closed-domain RAG systems.

Supplementary information:

Please cite this paper as:

1 Introduction

Retrieval-Augmented Generation (RAG) grounds language model outputs in external knowledge sources by combining a retrieval step over a document corpus with a generative language model, an approach first formalized for knowledge-intensive NLP tasks [1] and closely related to earlier retrieval-augmented pre-training methods such as REALM [2] and dense passage retrieval [3]. Since the introduction of the Transformer architecture [4] and large pretrained language models such as BERT [5], RAG has become a standard tool for reducing hallucination — the tendency of language models to produce fluent but unsupported content [6] — by conditioning generation on retrieved passages rather than relying solely on parametric memory. This is especially attractive for practitioners without large-scale compute, since it allows small, resource-efficient models to answer domain-specific questions without the cost of fine-tuning or the memory footprint of much larger models such as those trained under modern scaling regimes [7, 8].

Retrieval in such systems is typically implemented via dense embedding similarity, following approaches such as Sentence-BERT [9] and its distilled variants like MiniLM [10], often paired with efficient similarity search infrastructure [11] and evaluated using standardized retrieval benchmarks [12]. Downstream generation quality is conventionally assessed with lexical-overlap metrics originally developed for machine translation and summarization, such as BLEU [13] and ROUGE [14]. More sophisticated RAG variants incorporate iterative retrieval, self-critique, or fusion-in-decoder generation [15, 16] to improve groundedness; the system studied in this paper instead isolates a simpler, more common architecture — a single dense retrieval step augmented with an explicit relevance guardrail — to study a specific failure mode in isolation.

A common assumption in the design of such systems is that answer reliability can be controlled primarily through pipeline tuning: adjusting the similarity threshold that determines which retrieved passages are "relevant enough" to answer from, and refining prompts to discourage the model from answering outside retrieved context, an approach conceptually related to instruction tuning and human-feedback-based alignment methods used to shape model behavior more generally [17], and to reasoning-eliciting prompting strategies such as chain-of-thought [18]. This assumption implies that, given enough tuning, a small RAG system can be made to reliably distinguish between questions it can answer from its source material and questions it cannot, regardless of model size. However, it remains unclear whether this holds when a question is topically related to the source domain but not actually answerable from it — a failure mode closely tied to hallucination more broadly [6], and one that embedding-based retrieval is not obviously equipped to detect. This distinction has a direct precedent in extractive question answering: the original SQuAD benchmark [19] was later extended with adversarially constructed unanswerable questions closely resembling answerable ones [20], requiring systems to determine when no answer is supported and abstain accordingly — a challenge also reflected in broader knowledge-intensive task benchmarks [21]. Separately, prior work on large, general-purpose language models has found that calibration and self-evaluation tend to improve with model scale [22], and that explicit training can teach models to abstain when uncertain [23]; this has primarily been studied in large, non-retrieval-augmented models rather than sub-2-billion-parameter models operating inside a RAG pipeline under hardware constraints — the gap this paper addresses.

In this work, we experimentally investigate this assumption using a Retrieval-Augmented Generation pipeline built with standard open-source tooling [24] over the Harry Potter book series, with generation handled by a small instruction-tuned model from the Qwen2.5 family [25]; we iteratively tune both the retrieval guardrail and the generation prompt across eight configurations, and show that guardrail failures and generation failures are governed by different underlying mechanisms. Specifically, our experiments indicate that embedding similarity cannot reliably separate "topically related" questions from "answerable" questions at any threshold setting, and that a model's ability to recognize and

refuse such questions is a function of model scale rather than prompt design — a distinction that has direct implications for anyone building similar systems under hardware constraints.
We evaluate this pipeline on a 175-question benchmark spanning in-domain, out-of-domain, and topically-adjacent-but-unanswerable questions, using standard lexical-overlap metrics [13, 14] alongside guardrail- and model-level refusal rates, and compare results across two model sizes (Qwen2.5-0.5B and Qwen2.5-1.5B-Instruct [25]).

2 Related Work

2.1 Retrieval-Augmented Generation

Retrieval-Augmented Generation (RAG) has emerged as one of the most effective approaches for improving the factual reliability of large language models by augmenting text generation with external knowledge sources rather than relying solely on parametric memory. The original RAG framework proposed by Lewis et al. demonstrated that combining dense document retrieval with sequence generation substantially improves performance on knowledge-intensive natural language processing tasks while providing greater factual grounding than conventional language models [1]. A typical RAG pipeline consists of document indexing, semantic retrieval, and response generation, and subsequent surveys have identified RAG as a key direction for trustworthy language model deployment [26, 21].
Advances in dense retrieval methods, particularly Dense Passage Retrieval (DPR), have significantly improved the ability to retrieve semantically relevant documents from large knowledge collections [3]. More recent systems, such as Self-RAG, further integrate retrieval with self-reflection mechanisms, enabling language models to evaluate retrieved evidence before producing a final response [16]. Despite these improvements, retrieval quality remains the primary factor determining downstream generation performance, as inaccurate or incomplete retrieval directly limits the quality of generated answers.

2.2 Hallucination in Large Language Models

Hallucination remains one of the most significant challenges affecting the reliability of modern large language models. Ji et al. define hallucinations as generated content that is unsupported by either the provided context or verifiable external knowledge and provide a comprehensive taxonomy of hallucination sources in natural language generation systems [6]. Even highly capable transformer-based models remain susceptible to confidently generating incorrect information, particularly when contextual evidence is incomplete or ambiguous [4, 7].
The scaling of language models has substantially improved reasoning and language understanding capabilities [8], while instruction tuning and reinforcement learning from human feedback have further enhanced alignment with user intentions [17]. Nevertheless, increasing model size alone does not eliminate hallucinations. Recent work has shown that language models often produce fluent yet factually unsupported responses even when uncertainty exists, motivating research into mechanisms that encourage models to recognize and express uncertainty rather than fabricate answers [22, 23].

2.3 Hallucination Mitigation in Retrieval-Augmented Generation

A variety of approaches have been proposed to improve the reliability of Retrieval-Augmented Generation systems specifically. At the embedding level, Sentence-BERT introduced semantically meaningful sentence embeddings that significantly improve dense retrieval performance [9], while MiniLM demonstrated that compact transformer models can achieve competitive semantic representations with substantially lower computational requirements [10]. Efficient vector indexing techniques such as FAISS further enable scalable similarity search over large document collections [11].
Beyond retrieval improvements, prompt engineering and guardrail mechanisms have emerged as complementary strategies for reducing hallucinations. Prompting techniques encourage language models to abstain from answering when supporting evidence is insufficient, while instruction-based approaches such as R-Tuning explicitly train models to respond with "I don't know" when appropriate [23]. Self-RAG extends this concept by allowing models to critique their own retrieved evidence before

generating an answer [16]. Collectively, these approaches seek to improve factual reliability without requiring modifications to the underlying language model architecture.

2.4 Evaluation of Retrieval-Augmented Systems

Reliable evaluation remains essential for measuring the effectiveness of Retrieval-Augmented Generation systems. Traditional language generation metrics, including BLEU and ROUGE, continue to provide useful measures of lexical similarity between generated and reference answers [13, 14]. However, these metrics alone are insufficient for evaluating retrieval-based systems because they do not assess retrieval quality or evidence grounding.

Consequently, recent studies increasingly combine generation metrics with retrieval-specific measures such as precision, recall, context relevance, and answer groundedness [26, 12]. Benchmark datasets including SQuAD and KILT have also become widely used for evaluating question-answering systems under both answerable and unanswerable conditions [20, 19, 21]. These benchmarks emphasize that successful retrieval alone does not guarantee factual correctness, highlighting the need to evaluate both retrieval performance and generation behavior within an integrated framework.

2.5 Research Gap

Although Retrieval-Augmented Generation has become an active area of research, several limitations remain in the existing literature. Most studies evaluate either retrieval performance or generation quality independently, rather than jointly analyzing whether failures originate at the retrieval stage or the generation stage of the pipeline. Comparatively fewer investigate how lightweight guardrail mechanisms influence the overall behavior of complete RAG pipelines. Furthermore, many existing approaches rely on large proprietary language models or computationally intensive verification strategies, limiting their applicability in resource-constrained environments.

Another limitation is that previous evaluations primarily emphasize overall answer accuracy while providing comparatively little analysis of domain-adjacent questions, retrieval ambiguity, refusal behavior, and computational trade-offs. These aspects are particularly important when deploying compact open-source language models, where balancing factual reliability against computational efficiency becomes a practical concern.

To address these gaps, the present study performs a systematic evaluation of a Retrieval-Augmented Generation system built upon the Harry Potter book corpus. Multiple guardrail configurations are assessed across several small open-source language models using benchmark datasets containing in-domain, out-of-domain, and domain-adjacent questions. In addition to conventional lexical evaluation metrics, the study examines retrieval effectiveness, hallucination mitigation, refusal behavior, and response latency to provide practical guidelines for designing efficient and trustworthy Retrieval-Augmented Generation systems.

The remainder of this paper is organized as follows. Section 3 presents the proposed Retrieval-Augmented Generation framework, including the retrieval pipeline, guardrail configurations, and experimental methodology. Section 4 presents the experimental results and evaluation, including the analysis of retrieval performance, hallucination mitigation, refusal behavior, and computational trade-offs. Finally, Section 5 summarizes the main findings, discusses the limitations of the study, and outlines directions for future work.

3 Proposed Method

This study develops and evaluates a Retrieval-Augmented Generation (RAG) system for closed-domain question answering over the complete Harry Potter book corpus. The methodology consists of four stages: system construction, experimental environment definition, baseline implementation, and iterative optimization through controlled experiments. Figure 1 provides an overview of the resulting RAG pipeline and the main stages of the experimental methodology.

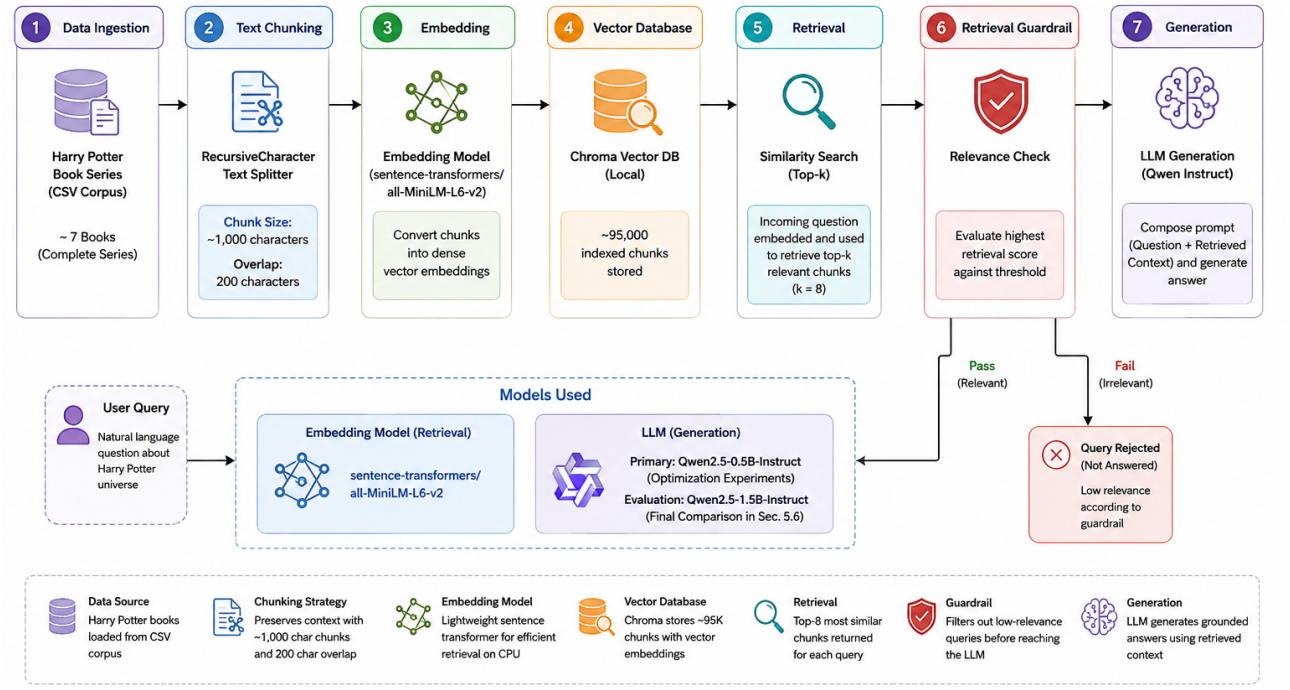


Figure 1: Overview of the proposed Retrieval-Augmented Generation (RAG) pipeline for closed-domain question answering over the Harry Potter book corpus.

3.1 System Architecture

The system follows a retrieve–then–generate Retrieval-Augmented Generation (RAG) architecture comprising seven sequential stages. An overview of the pipeline is presented in Figure 1, while the corpus preprocessing procedure is described in detail in Section 3.2.1.

3.2 Experimental Environment

3.2.1 Dataset Description

Two datasets were employed throughout this study: (1) a knowledge corpus used by the Retrieval-Augmented Generation (RAG) system for semantic retrieval and (2) an evaluation benchmark used to assess answer quality and the system's ability to handle unsupported queries. The knowledge corpus was derived from a publicly available Harry Potter text dataset [27], while the evaluation benchmark combined questions sampled from an existing Harry Potter question-answering dataset with manually curated questions designed to evaluate the system's behavior on unsupported queries [28].

The retrieval corpus consists of the complete Harry Potter book series represented as a structured CSV dataset published by Gaston Sanchez [27]. The dataset contains the text of all seven novels together with metadata identifying the corresponding book and chapter. Prior to indexing, the corpus was segmented into overlapping text chunks using LangChain's Recursive Character Text Splitter [24], with a chunk size of approximately 1,000 characters and an overlap of 200 characters. This preprocessing strategy preserves contextual continuity while producing retrieval units of manageable size.

Each text chunk was encoded using the Sentence-Transformers implementation of the all-MiniLM-L6-v2 embedding model [9, 10]. The resulting vector representations were indexed within a ChromaDB vector database to support semantic similarity search throughout the experiments.

System performance was evaluated using a benchmark consisting of three complementary categories of questions, as illustrated in Figure 2. The benchmark contains 175 questions in total: 100 in-domain questions, 25 adjacent questions, and 50 out-of-domain questions.

In-domain questions (n = 100): These questions were randomly sampled from the publicly available vapit/HarryPotterQA dataset hosted on Hugging Face [28]. Preliminary inspection revealed numerous duplicated entries and templated answers reused across unrelated questions. Consequently,

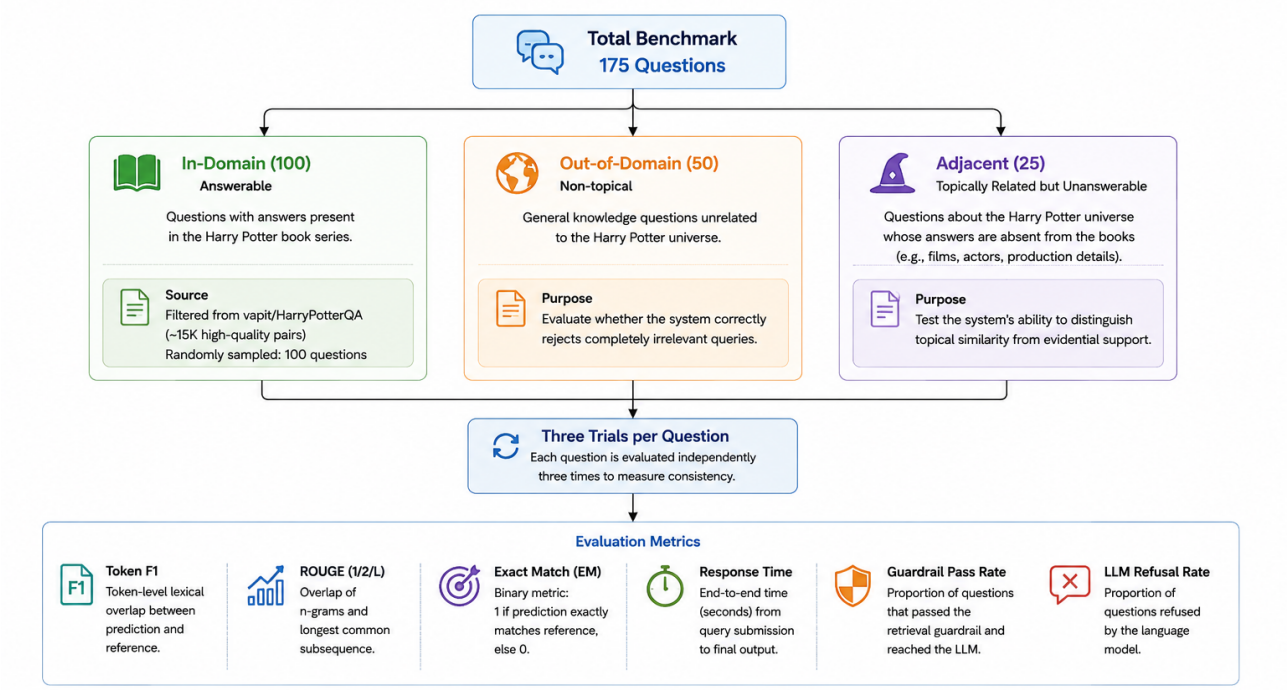


Figure 2: Overview of the evaluation benchmark.

approximately half of the original dataset (around 15,000 of roughly 30,000 question–answer pairs) was removed during preprocessing to improve data quality before randomly sampling the evaluation subset.

Out-of-domain questions (n = 50): These questions consist of general factual questions unrelated to the Harry Potter universe, including topics such as geography, science, and general knowledge. This category evaluates whether the retrieval guardrail appropriately rejects queries that fall outside the indexed knowledge corpus.

Adjacent questions (n = 25): These questions were specifically designed for this study to evaluate a failure mode that is less directly captured by conventional closed-domain RAG benchmarks: queries that are semantically related to the target domain but whose answers are absent from the indexed corpus. The questions focus on Harry Potter-related information not contained in the book corpus, such as film, actor, and production details. This category enables evaluation of the system's ability to distinguish domain-related questions from questions that can actually be answered using the available corpus, thereby probing its susceptibility to generating unsupported responses.

Together, the three categories provide a multi-dimensional evaluation framework that measures both answer quality on questions supported by the corpus and the system's ability to identify and appropriately handle queries for which sufficient evidence is unavailable.

3.2.2 Hardware and Software Environment

All experiments were conducted on consumer CPU-only hardware with approximately 10 GB of system memory and no GPU acceleration. These hardware constraints influenced model selection and the experimental design by limiting the size of models that could be executed reliably alongside the embedding model and vector database.

Initial experiments were conducted using Qwen2.5-1.5B-Instruct. However, simultaneous execution of the language model, embedding model, and vector database exceeded the available memory budget and resulted in repeated segmentation faults associated with memory exhaustion. To ensure stable execution throughout the iterative optimization process, subsequent development and rule-set experiments were therefore performed using the smaller Qwen2.5-0.5B-Instruct model.

The computational constraints were incorporated into the experimental setting rather than treated solely as an implementation limitation. Accordingly, the primary objective of the optimization process was not to maximize absolute performance, but to investigate the extent to which guardrail configura-

tion and prompt engineering could influence grounded generation within a resource-constrained RAG environment.
After the optimization process was completed, Qwen2.5-1.5B-Instruct was reintroduced to evaluate the effect of model scale while keeping the optimized retrieval and prompting configuration fixed. Specifically, the best-performing rule set identified during the optimization experiments was applied to both model sizes. This comparison provides a controlled assessment of whether increasing model capacity improved answer quality and guardrail behavior under the same RAG configuration.

3.3 Baseline Configuration

Before the iterative optimization process began, an initial baseline configuration, referred to as the No Length Constraint (NLC) baseline, was established to characterize the behavior of the untuned pipeline. The retrieval guardrail employed a relevance threshold of 0.85 with top-k retrieval ($k = 8$), while the language model was configured with a maximum generation length of 300 new tokens and a sampling temperature of 0.3. The prompt instructed the model to answer questions exclusively using information contained in the Harry Potter corpus and to explicitly refuse questions that could not be answered from the retrieved context.
The full results of the NLC baseline, including guardrail behavior, refusal rates, response characteristics, and inference latency, are reported in Section 4.2. These results motivated the first optimization stage (Rule Set 1), which focused on constraining generation behavior through prompt refinement and more deterministic decoding.
Between the NLC baseline and Rule Set 1, the intentional modifications recorded in the project's version-control history were limited to the generation configuration. Specifically, the maximum generation length was reduced from 300 to 20 tokens, the sampling temperature was reduced from 0.3 to 0.2, and top-p sampling ($p = 0.9$) was introduced. No intentional changes were recorded to the retrieval pipeline or guardrail logic.
Despite the absence of documented changes to the retrieval configuration, the measured in-domain guardrail pass rate differed between the NLC baseline and Rule Set 1. Because generation parameters are applied only after a query has passed the guardrail, the observed difference cannot be directly attributed to the documented generation changes. An untracked rebuild or modification of the vector database during the same development period is one possible explanation; however, this could not be verified retrospectively. The discrepancy therefore represents a limitation of the early experimental tracking process.
Following this observation, configuration management and experiment logging were made substantially stricter from Rule Set 3 onward. The subsequent experiments were conducted under more explicitly tracked configurations, and comparisons among these later rule sets are interpreted within the documented experimental sequence. The NLC–Rule Set 1 discrepancy is consequently treated as an early-stage reproducibility limitation rather than as evidence of a causal change in retrieval behavior.

3.4 Iterative Optimization Strategy

The optimization process followed a controlled, incremental methodology in which successive pipeline configurations, referred to as "Rule Sets," introduced documented changes to either the retrieval guardrail or the generation configuration. Where possible, the remaining components were kept fixed so that the effect of individual interventions could be examined in isolation. After each configuration change, the complete evaluation benchmark was rerun before the next intervention was introduced. This procedure enabled performance differences to be examined systematically while reducing the potential influence of uncontrolled simultaneous changes.
From the earliest stages of development, guardrail-level rejections and language-model refusals were recorded as separate outcomes. A guardrail rejection occurred when a query failed the retrieval relevance criterion and was therefore prevented from reaching the generation stage. In contrast, a language-model refusal occurred after the query had passed the retrieval guardrail but the model declined to provide an answer based on the retrieved context. Maintaining this distinction enabled

retrieval-level and generation-level behavior to be analyzed separately rather than combining both outcomes into a single refusal measure.

As discussed in Section 3.3, an unexplained shift in guardrail behavior occurred between the NLC baseline and Rule Set 1. Because the underlying cause of this change could not be conclusively identified, comparisons involving these earliest configurations should be interpreted with appropriate caution. From Rule Set 3 onward, the experimental configurations and corresponding changes were more systematically tracked. Within this later sequence, guardrail behavior changed consistently with the documented retrieval-threshold adjustments, while prompt-only modifications did not produce measurable changes in retrieval outcomes. These observations support interpreting the later Rule Set comparisons as controlled evaluations of the corresponding documented configuration changes, while avoiding unsupported causal claims about the undocumented early-stage discrepancy.

The iterative optimization procedure is summarized in Algorithm 1. Starting from the initial configuration, each Rule Set was evaluated on the complete benchmark and its resulting answer quality, guardrail behavior, refusal outcomes, and response characteristics were recorded. The process continued while further configuration changes produced meaningful performance improvements. The best-performing configuration identified during this process was subsequently retained as the reference configuration for the model-scaling experiment.

Algorithm 1 Iterative Optimization of the RAG Configuration

Require: Initial RAG configuration C_0, evaluation benchmark B
Ensure: Selected optimized configuration C^*
Evaluate C_0 on the complete benchmark B
Record answer quality, guardrail behavior, refusal outcomes, and response metrics
Set $C_{\text{current}} \leftarrow C_0$
Set $P_{\text{best}} \leftarrow$ performance of C_0
while further configuration changes improve performance **do**
  Select a documented configuration intervention
  Modify the retrieval guardrail or generation configuration
  Define the resulting configuration as C_{new}
  Evaluate C_{new} on the complete benchmark B
  Record answer quality, guardrail behavior, refusal outcomes, and response metrics
  Compare the performance of C_{new} with C_{current}
  if C_{new} provides improved performance **then**
    $C_{\text{current}} \leftarrow C_{\text{new}}$
    Update P_{best}
  else
    Stop the optimization process
  end if
end while
$C^* \leftarrow C_{\text{current}}$
Retain C^* for subsequent model-scale evaluation
return C^*

The incremental optimization strategy enabled the contribution of different retrieval and generation interventions to be examined throughout the development process. Rather than reporting only the final configuration, the sequence of intermediate Rule Sets documents both successful and unsuccessful optimization attempts and provides a transparent account of the progression toward the selected configuration. In the subsequent experiments, Rule Set 7 was retained as the reference configuration for evaluating the effect of model scale under a fixed retrieval and generation setup.

3.5 Statistical Significance Analysis

To determine whether differences between experimental configurations were statistically significant, paired comparisons were performed using McNemar's test [29, 30]. McNemar's test is appropriate for

paired categorical outcomes because each configuration was evaluated on the same set of benchmark questions, allowing the performance of two configurations to be compared on a question-by-question basis. The test focuses on discordant pairs, namely questions for which the two configurations produced different evaluation outcomes [29].

Two complementary analyses were conducted. First, *answer quality* was evaluated on the 100 in-domain questions whose answers were expected to be supported by the Harry Potter book corpus. For this analysis, a response was classified as either correct or incorrect, with ambiguous responses treated as incorrect. Refusals on in-domain questions were likewise treated as incorrect because the system failed to provide an answer to a question that was answerable from the target corpus.

Second, *guardrail behavior* was evaluated on the 75 questions that were not directly answerable from the indexed book corpus, comprising 25 Harry Potter-adjacent questions and 50 out-of-domain questions. An appropriate refusal was considered correct guardrail behavior, whereas answering an unsupported question was considered a guardrail failure. Ambiguous responses were conservatively treated as guardrail failures.

For each paired comparison, we report the number and proportion of correct outcomes for each configuration, the absolute difference in performance, the 2×2 contingency table of paired outcomes, and the exact two-sided McNemar p-value. An exact test was used rather than relying solely on the asymptotic chi-square approximation because the number of discordant observations can be relatively small in some comparisons. Statistical significance was assessed at $\alpha = 0.05$. A p-value below 0.05 was interpreted as evidence that the two configurations differed significantly on the corresponding evaluation criterion. The statistical analysis was conducted separately for answer quality and guardrail behavior to distinguish improvements in grounded answering from changes in refusal behavior.

4 Results and Evaluation

4.1 Overall System Performance

The proposed Retrieval-Augmented Generation (RAG) system was evaluated using the benchmark described in Section 3.2.1. The results are organized into four thematic parts. First, the baseline configuration (NLC) establishes the behavior of the unoptimized system and provides a reference point for subsequent experiments. Second, the retrieval guardrail and generation prompt are iteratively modified across successive Rule Set configurations to examine their effects on system behavior. Third, the influence of language model scale is investigated by comparing Qwen2.5 instruction-tuned models with 0.5B and 1.5B parameters under the same optimized Rule Set configuration. Finally, the evaluation of adjacent and out-of-domain questions examines the system's ability to reject queries for which appropriate evidence is unavailable in the indexed corpus.

In addition to the aggregate system-level metrics reported for each configuration, a manual evaluation was conducted on the complete set of 100 in-domain questions and 75 non-book-content questions for the first trial. The manual evaluation distinguishes answer quality from guardrail behavior and enables paired statistical comparisons between configurations using McNemar's exact test. This additional analysis provides a more direct assessment of whether observed differences between configurations represent statistically significant changes in individual question-level outcomes.

4.2 Baseline System Performance (NLC)

The NLC baseline establishes the performance of the RAG pipeline before the introduction of the subsequent guardrail and prompt optimizations. As described in Section 3.3, the baseline employed a retrieval threshold of 0.85 with $k = 8$ retrieved chunks and a maximum generation length of 300 new tokens.

Figure 3 presents the response-time distribution for the baseline configuration. Questions rejected by the retrieval guardrail were processed almost instantaneously, whereas questions reaching the language model required substantially longer inference times. Model-processed questions exhibited an average latency of approximately 41 s, with several responses approaching 100 s.

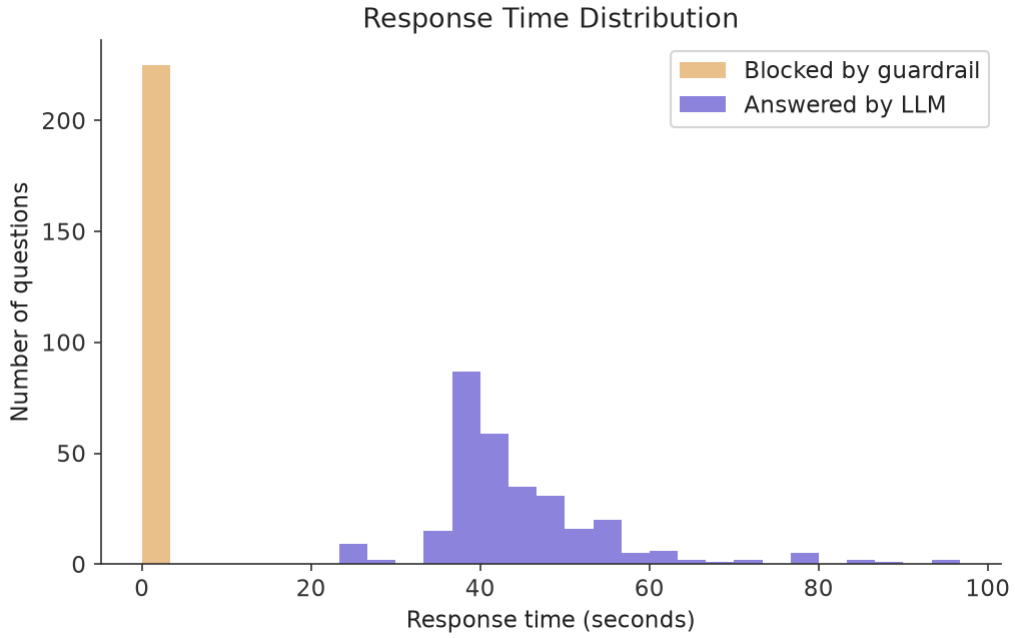


Figure 3: Response-time distribution for the NLC baseline.

Figure 4 summarizes the distribution of retrieval guardrail rejections, language-model refusals, and completed responses across the three evaluation categories. The retrieval guardrail accepted 96% of in-domain questions, while rejecting all generic out-of-domain questions and 84% of adjacent questions. Among the in-domain questions that passed the retrieval guardrail, approximately 31% were subsequently refused by the language model. This distinction between retrieval-level rejection and language-model refusal is important because the two mechanisms represent different failure modes within the RAG pipeline.

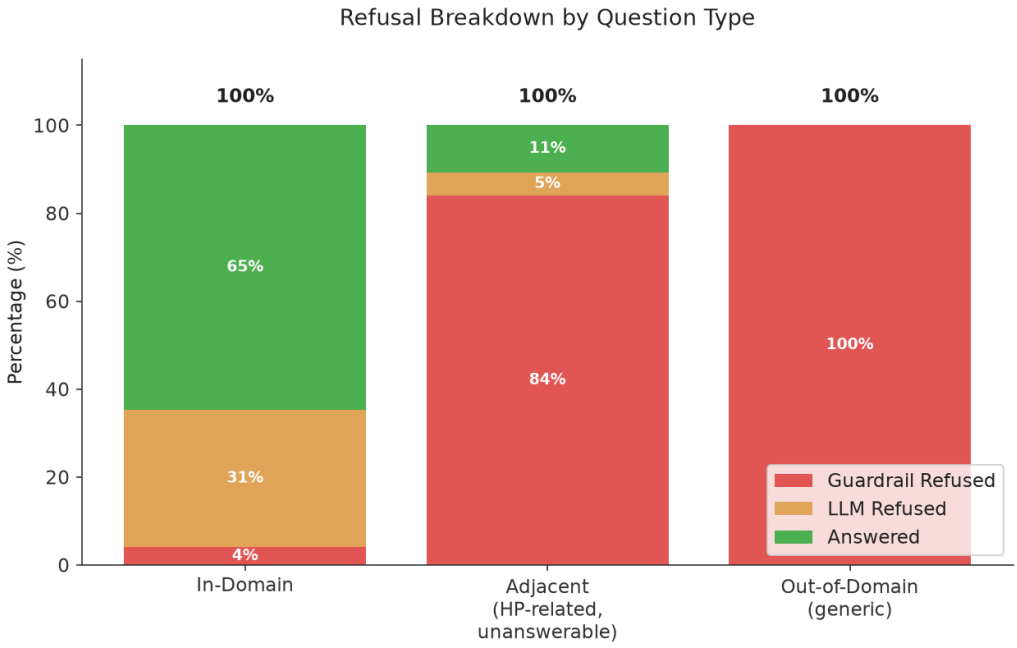


Figure 4: Refusal breakdown by question type for the NLC baseline.

In addition to the observed refusal behavior, generated responses averaged 42.4 words, compared with an average reference-answer length of 19.8 words. The baseline achieved a Token F1 score of 0.21 and a ROUGE-L score of 0.19. Across the repeated trials, the system exhibited a consistency of 91.4%. The combination of relatively long responses, substantial language-model latency, and frequent in-

domain language-model refusals motivated the first optimization stage (Rule Set 1). This stage focused primarily on reducing response length through prompt refinement and more deterministic generation settings.

Figure 5 presents the category-level outcome distribution after applying Rule Set 1. As discussed in Section 3.3, an unexpected change in guardrail behavior was observed between the NLC baseline and Rule Set 1. Because the recorded code changes between these configurations affected only generation parameters, this change cannot be attributed to the documented experimental intervention. It is therefore treated as an early-stage experimental tracking limitation rather than evidence of a causal change in retrieval behavior. Accordingly, the subsequent analysis focuses primarily on Rule Sets 3–8, for which configuration tracking and experiment logging were more strictly controlled.

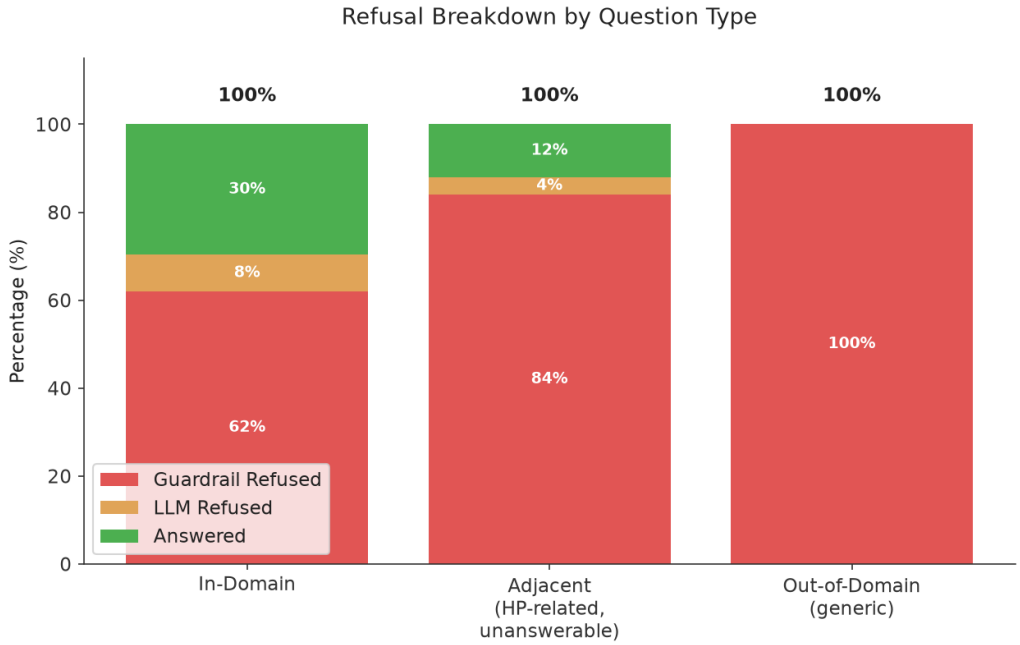


Figure 5: Refusal breakdown by question type for Rule Set 1. The apparent change in guardrail behavior relative to the baseline is discussed in Section 3.3.

4.3 Prompt and Guardrail Optimization

The effects of retrieval-guardrail configuration and prompt design were investigated across Rule Sets 2–8. The experiments were conducted incrementally, with changes to the retrieval threshold and generation configuration introduced in a controlled sequence while the remaining components were kept fixed whenever possible. Table 1 summarizes the configurations and their corresponding evaluation results. Rule Set 7 is highlighted as the reference configuration selected for the subsequent model-scale comparison.

Table 1: Summary of the optimization process across Rule Sets 2–8. Rule Set 7 is used as the reference configuration in the subsequent model-scale comparison.

Rule Set	Ans. Len (ref ∼19.8)	Consistency (%)	In-Domain Passed (%)	OOD Blocked (%)	Token F1	ROUGE-L	Avg. Resp. Time (answered)
2	14.5	96.6	38	94.7	0.15	0.14	∼24s
3	13.4	100	38	94.7	0.16	0.16	∼20–22s
4	12.8	100	55	90.7	0.19	0.18	∼21–22s
5	12.4	100	67	78.7	0.20	0.20	∼21–22s
6	15.3	98.9	67	78.7	0.24	0.22	∼21–21.5s
7	15.2	98.9	82	65.3	0.26	0.25	∼21s
8	15.9	98.3	82	65.3	0.26	0.24	∼26s

Figure 6 illustrates the relationship between the retrieval configuration and the resulting in-domain pass rate and out-of-domain block rate. Across the evaluated configurations, the two measures exhibited a clear trade-off: configurations associated with higher in-domain acceptance also exhibited lower rejection of non-book-content questions.

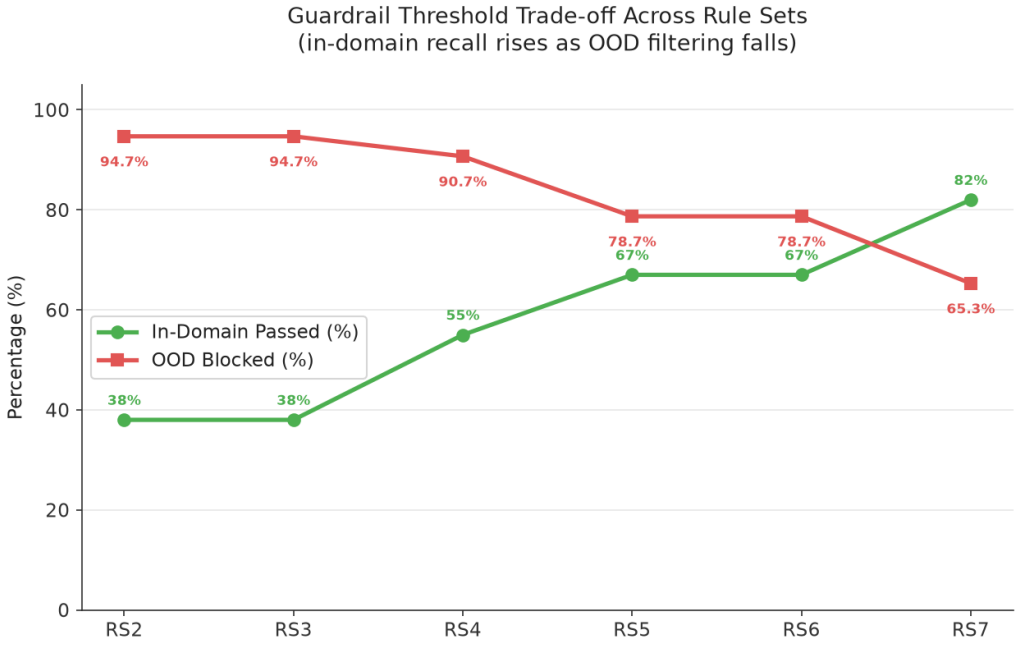


Figure 6: Guardrail threshold trade-off across evaluated rule sets.

Across Rule Sets 3 through 7, the in-domain pass rate increased from 38% to 82%, while the corresponding out-of-domain block rate decreased from 94.7% to 65.3%. This trade-off demonstrates that the retrieval guardrail threshold substantially affects the balance between accepting potentially answerable questions and rejecting unsupported queries. Rule Set 7 provided the highest in-domain pass rate observed during this sequence while retaining a non-negligible level of rejection for out-of-domain questions.

Prompt-only modifications introduced between successive configurations did not produce observable changes in guardrail behavior when the retrieval configuration remained unchanged. For example, Rule Sets 2 and 3, Rule Sets 5 and 6, and Rule Sets 7 and 8 exhibit identical in-domain pass and out-of-domain block rates despite differences in prompt wording and generation parameters. This separation between retrieval and generation behavior provides evidence that the observed guardrail changes were primarily associated with the retrieval configuration rather than prompt modifications.

Alongside the changes in retrieval behavior, lexical-overlap metrics generally improved during the optimization process. Token F1 increased from 0.15 in Rule Set 2 to 0.26 in Rule Set 7, while ROUGE-L increased from 0.14 to 0.25 over the same sequence. These metrics indicate improved lexical agreement between generated and reference answers; however, they do not by themselves establish factual correctness or groundedness, which are evaluated separately through the manual question-level analysis presented later.

These results establish the empirical performance trends observed during the retrieval-configuration experiments and motivate the subsequent prompt-engineering analysis. In particular, Rule Set 7 provided the highest combination of in-domain acceptance and lexical-overlap performance among the evaluated configurations and was therefore selected as the reference configuration for the model-scale experiment.

Following the retrieval-configuration experiments, a series of prompt refinements was evaluated while keeping the guardrail configuration fixed. These experiments were designed to determine whether generation behavior could be modified independently of retrieval-stage behavior.

Figure 7 compares the average generated answer length across the evaluated configurations with the average reference-answer length. The initial Rule Set 1 configuration substantially reduced response length relative to the NLC baseline, producing answers averaging 12.1 words. Similar response lengths were maintained throughout Rule Sets 2–5, ranging from 12.4 to 14.5 words. Following the introduction of a more permissive prompt in Rule Set 6, the average response length increased to approximately 15–16 words while remaining below the verbosity observed in the NLC baseline.

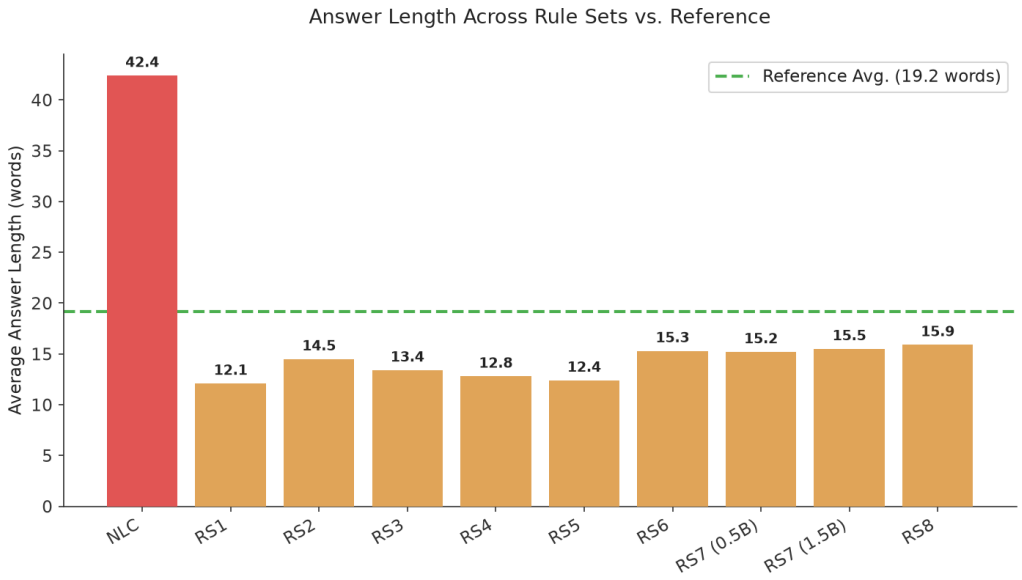


Figure 7: Average generated answer length across evaluated rule sets compared with the NLC baseline and the reference-answer average.

To examine whether prompt modifications were associated with changes in answer-generation metrics independently of retrieval behavior, Rule Sets 5 and 6 were compared directly. These configurations shared identical retrieval behavior, with a 67% in-domain pass rate and a 78.7% out-of-domain block rate, while differing in prompt design.

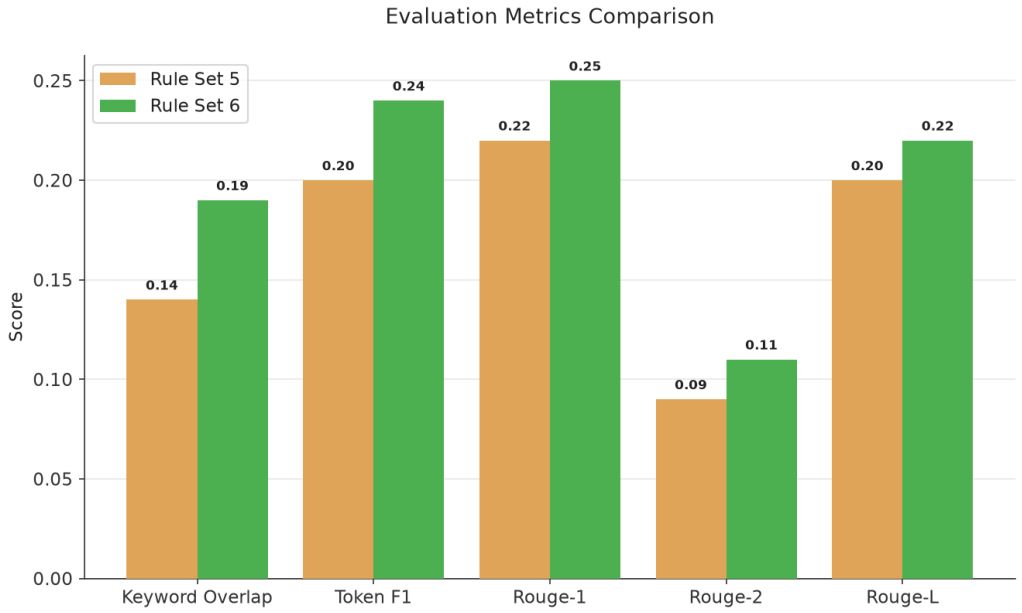


Figure 8: Comparison of evaluation metrics between Rule Sets 5 and 6. The retrieval configuration remained unchanged, allowing the effect of prompt refinement on generation metrics to be examined independently of guardrail behavior.

Figure 8 summarizes the corresponding evaluation metrics. Despite identical guardrail behavior in both configurations, Rule Set 6 achieved higher lexical-overlap scores, with Token F1 increasing from 0.20 to 0.24 and ROUGE-L increasing from 0.20 to 0.22. The average response length also increased from 12.4 to 15.3 words.

The improvement in lexical-overlap metrics occurred without an accompanying change in retrieval-stage behavior, indicating that the prompt modification affected the characteristics of generated responses while leaving the retrieval guardrail unchanged. These results motivated the adoption of the

resulting prompt formulation in the subsequent experiments.

4.4 Model Scaling Evaluation

The preceding experiments suggested that neither retrieval-threshold optimization nor prompt refinement substantially reduced failures on adjacent questions. To determine whether this limitation was instead related to model capacity, the reference Rule Set 7 configuration was evaluated using both Qwen2.5-0.5B-Instruct and Qwen2.5-1.5B-Instruct while keeping the retrieval pipeline, guardrail threshold, prompt, and evaluation benchmark identical.
Tables 2 and 3 summarize the aggregate evaluation metrics and category-level outcomes.

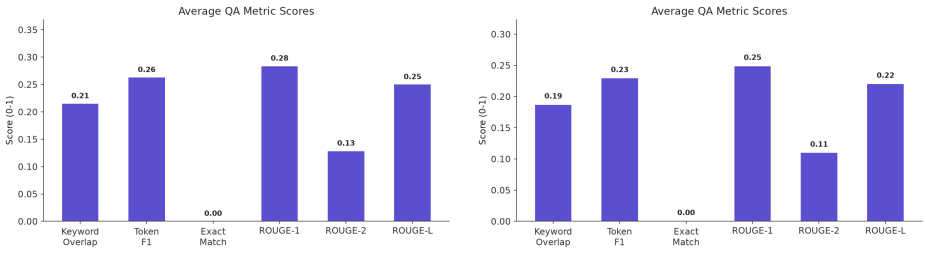
Table 2: Aggregate evaluation metrics for the two evaluated language models under the Rule Set 7 configuration.

Model	Ans. Len (ref ∼19.8)	Consistency (%)	Token F1	ROUGE-L	Avg. Resp. Time (answered)
Qwen2.5-0.5B	15.2	98.9	0.26	0.25	∼21s
Qwen2.5-1.5B	15.5	97.7	0.23	0.22	∼61s

Table 3: Category-level outcome comparison for Qwen2.5-0.5B and Qwen2.5-1.5B under the Rule Set 7 configuration.

Category	Outcome	Qwen2.5-0.5B (%)	Qwen2.5-1.5B (%)
In-Domain	Answered	81	66
In-Domain	LLM Refused	1	16
In-Domain	Guardrail Refused	18	18
Adjacent	Answered	64	35
Adjacent	LLM Refused	12	41
Adjacent	Guardrail Refused	24	24
Generic OOD	Answered	14	12
Generic OOD	LLM Refused	0	2
Generic OOD	Guardrail Refused	86	86

Tables 2 and 3 demonstrate that increasing model size from Qwen2.5-0.5B to Qwen2.5-1.5B produced only modest improvements in overall evaluation metrics while substantially increasing average response latency. More importantly, the category-level analysis reveals that the larger model was considerably more likely to refuse adjacent-domain questions rather than generate unsupported responses, suggesting improved calibration when supporting evidence was uncertain. Nevertheless, both models exhibited similar guardrail rejection rates, indicating that retrieval-stage decisions remained largely unaffected by model scale. These observations motivate a closer examination of the adjacent-question benchmark, where retrieval and generation limitations become most apparent.
Figures 9(a) and 9(b) compare the principal lexical-overlap metrics obtained by the two models.



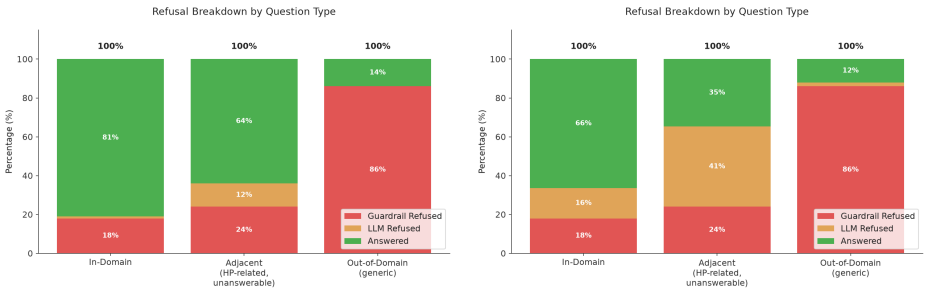
(a) Qwen2.5-0.5B (b) Qwen2.5-1.5B

Figure 9: Aggregate evaluation metrics for Qwen2.5-0.5B and Qwen2.5-1.5B.

Despite the larger model containing approximately three times as many parameters, aggregate lexical-overlap metrics remained broadly comparable. Qwen2.5-1.5B achieved a Token F1 of 0.23 compared

with 0.26 for Qwen2.5-0.5B, while ROUGE-L decreased slightly from 0.25 to 0.22. Consistency likewise remained high for both models, differing by only a small margin. Although these aggregate metrics suggest relatively similar overall performance, the category-level analysis reveals a substantially different pattern.

Figures 10(a) and 10(b) compare the distribution of answered and refused questions across benchmark categories.

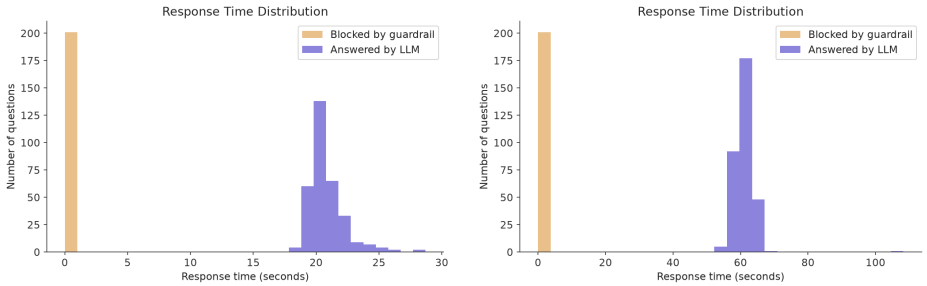


(a) Qwen2.5-0.5B (b) Qwen2.5-1.5B

Figure 10: Refusal breakdown by question type for Qwen2.5-0.5B and Qwen2.5-1.5B.

On adjacent questions, the larger model refused approximately 41% of queries after passing the retrieval guardrail, compared with only 12% for the smaller model. Correspondingly, the proportion of adjacent questions receiving generated answers decreased from approximately 64% to 35%. A smaller but consistent increase in refusal behavior was also observed on in-domain questions, indicating that the larger model adopted a generally more conservative answering strategy rather than selectively refusing only unsupported queries.

This improvement in adjacent-question handling was accompanied by a substantial increase in computational cost, as illustrated by the response latency comparison in Figures 11(a) and 11(b).



(a) Qwen2.5-0.5B (b) Qwen2.5-1.5B

Figure 11: Average response time for Qwen2.5-0.5B and Qwen2.5-1.5B under the Rule Set 7 configuration.

Average inference time increased from approximately 21 seconds to approximately 61 seconds per question, representing a latency increase of roughly 190%. Figure 12 summarizes this trade-off between computational cost and adjacent-question refusal performance.

To place the model comparison within the broader optimization process, Figure 13 presents all evaluated configurations on the same latency–refusal plane.

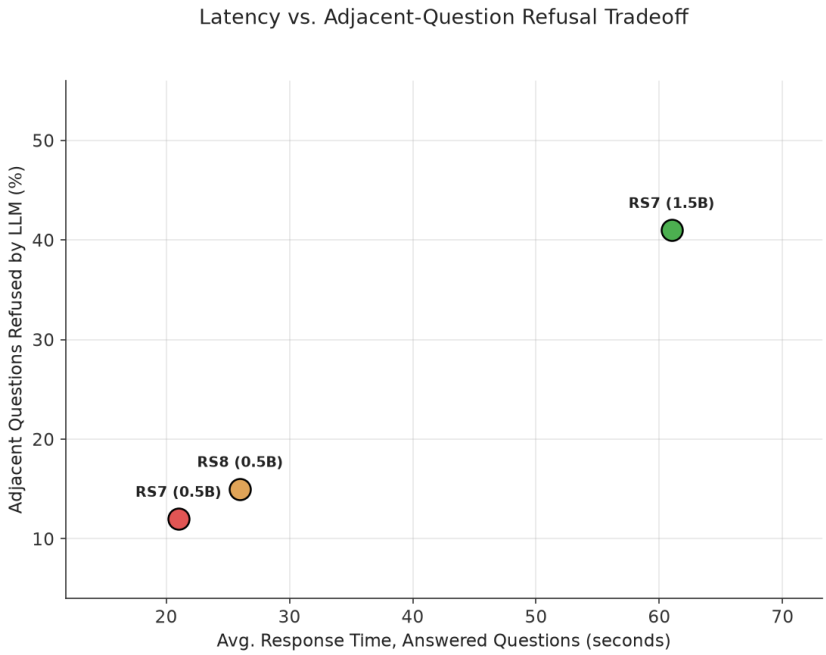


Figure 12: Latency versus adjacent-question refusal trade-off for representative configurations.

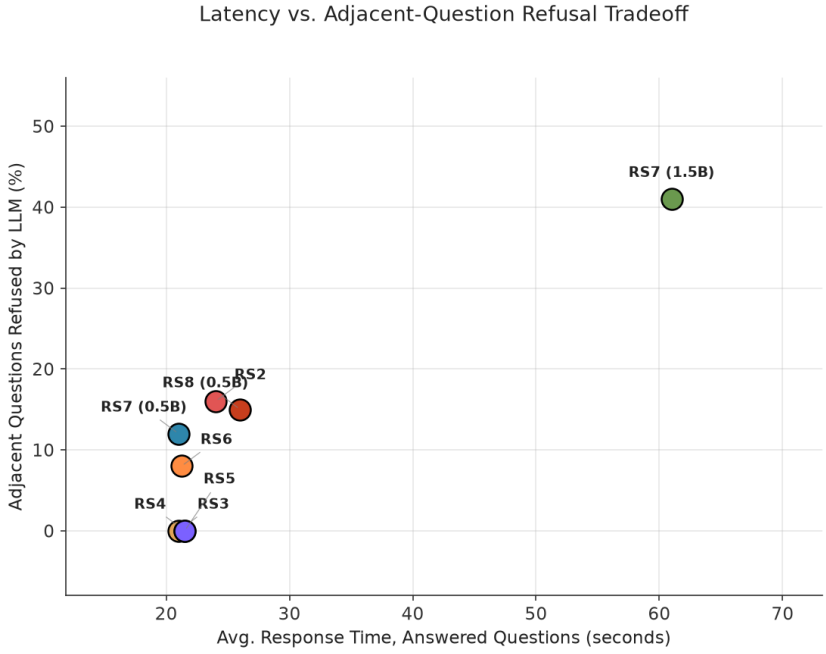


Figure 13: Latency versus adjacent-question refusal across all evaluated rule sets and both model sizes.

Across Rule Sets 2 through 8, prompt refinement and retrieval-threshold tuning produced relatively small changes in response latency while leaving adjacent-question refusal within a comparatively narrow range. In contrast, increasing model size shifted the operating point substantially, simultaneously increasing refusal behavior on adjacent questions and inference time. Within the design space explored in this study, model capacity therefore emerged as the only intervention producing a pronounced improvement on the adjacent-question failure mode.

Finally, an earlier discrepancy between preliminary chart images and summary statistics for Rule Sets 2 and 3 was resolved during manuscript preparation by verifying the original evaluation logs. The inconsistency was traced to an outdated draft figure rather than the experimental results themselves, while the early NLC-to-Rule Set 1 logging anomaly described in Section 3.3 remains a documented limitation of the study.

4.5 Statistical Significance Analysis

To determine whether the observed differences between configurations were statistically meaningful, paired statistical analyses were conducted using McNemar's exact two-sided test [29]. McNemar's test is appropriate for comparing two systems evaluated on the same set of binary outcomes because it focuses on discordant cases, i.e., questions for which the two configurations produced different outcomes. Statistical significance was assessed at $\alpha = 0.05$.

The statistical analysis was performed separately for answer quality and guardrail behavior. For answer quality, the analysis considered the 100 in-domain book-content questions. Each response was classified as correct or incorrect through human evaluation, with ambiguous responses and model refusals treated as incorrect. For guardrail behavior, the analysis considered the 75 non-book-content questions, consisting of 25 Harry Potter-adjacent questions and 50 out-of-domain questions. An appropriate refusal was considered a successful guardrail outcome, whereas an unsupported answer was considered a guardrail failure. Ambiguous responses were treated as guardrail failures. Because the original experiments demonstrated approximately 96% consistency across repeated trials, the statistical analysis was conducted using the first trial rather than requiring human evaluation of all three trials.

Table 4 summarizes the resulting paired comparisons. For answer quality, the reported rate represents the proportion of book-content questions classified as correct. For guardrail quality, the reported rate represents the proportion of non-book-content questions for which the system exhibited appropriate refusal behavior. The discordant-pair counts are reported as "first configuration only / second configuration only."

Table 4: Statistical significance analysis using exact two-sided McNemar's tests.

Analysis	**Measure**	**Comparison**	**First (%)**	**Second (%)**	**Δ (pp)**	**Discordant Pairs**	***p*-value**
Answer Quality	Correct	NLC vs. 0.5B	13.0	17.0	+4.0	9 / 13	0.5235
Answer Quality	Correct	0.5B vs. 1.5B	17.0	6.0	-11.0	12 / 1	0.0034
Guardrail Quality	Appropriate refusal	NLC vs. 0.5B	97.3	69.3	-28.0	21 / 0	< 0.001
Guardrail Quality	Appropriate refusal	0.5B vs. 1.5B	69.3	80.0	+10.7	0 / 8	0.0078

For answer quality, the comparison between the NLC baseline and Rule Set 7 showed a modest increase in human-evaluated accuracy from 13% to 17%. However, this difference was not statistically significant ($p = 0.5235$). Nine questions were answered correctly only by the baseline, whereas 13 were answered correctly only by Rule Set 7. Thus, although Rule Set 7 achieved a higher observed answer accuracy, the available evidence does not support a statistically significant improvement over the baseline.

In contrast, the comparison between Rule Set 7 and Qwen2.5-1.5B revealed a statistically significant difference in answer quality. Human-evaluated accuracy decreased from 17% for Qwen2.5-0.5B to 6% for Qwen2.5-1.5B, with 12 questions answered correctly only by the smaller model and one question answered correctly only by the larger model. The exact McNemar test yielded $p = 0.0034$, indicating a statistically significant difference at both the 0.05 and 0.01 significance levels. Therefore, despite the larger model exhibiting substantially more conservative refusal behavior, increased model capacity did not translate into improved human-evaluated answer quality on the book-content questions.

The guardrail analysis revealed statistically significant differences in both model comparisons. The NLC baseline achieved an appropriate refusal rate of 97.3%, compared with 69.3% for Rule Set 7. The difference was statistically significant ($p < 0.001$), with 21 questions appropriately refused only by the baseline and no questions appropriately refused only by Rule Set 7. This result demonstrates that the improved in-domain acceptance achieved through retrieval-threshold optimization was accompanied by a significant reduction in the system's ability to reject unsupported queries.

Increasing the model size from Qwen2.5-0.5B to Qwen2.5-1.5B produced the opposite effect on guardrail behavior. The appropriate refusal rate increased from 69.3% to 80.0%, with eight questions appropriately refused only by the larger model and no questions showing the opposite pattern. The exact McNemar test yielded $p = 0.0078$, indicating a statistically significant improvement in guardrail behavior. Thus, the larger model demonstrated significantly more conservative behavior on non-book-content questions, although this improvement was accompanied by a significant reduction in human-evaluated

answer quality on book-content questions.
Overall, the statistical analysis indicates that the effects of the evaluated interventions were not uniformly beneficial. Retrieval-threshold optimization substantially altered the balance between question acceptance and rejection, but the corresponding improvement in answer quality relative to the baseline was not statistically significant. Increasing model capacity significantly improved guardrail behavior while simultaneously producing a significant decrease in human-evaluated answer quality. These findings reinforce the importance of evaluating RAG systems using separate measures of answer quality and refusal behavior rather than relying solely on aggregate lexical-overlap metrics.

4.6 Hallucination Analysis

The preceding experiments revealed a persistent failure mode on the adjacent-question category despite improvements in several aggregate evaluation metrics. This section examines whether retrieval-threshold tuning and prompt refinement were able to reduce unsupported generation when questions were semantically related to the Harry Potter corpus but their answers were absent from the indexed books.
The category-level outcomes for the reference configuration (Rule Set 7) under Qwen2.5-0.5B are summarized in the Qwen2.5-0.5B column of Table 3, presented previously in Section 4.4 alongside the corresponding results for Qwen2.5-1.5B. Figure 14 visualizes the corresponding distribution of guardrail rejections, language-model refusals, and answered questions.

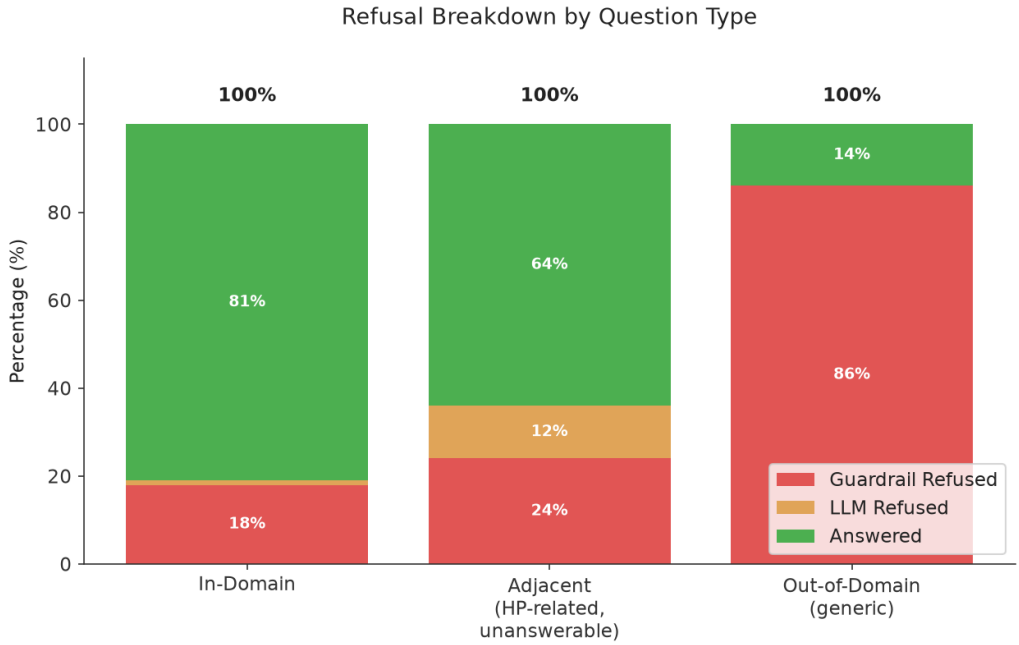


Figure 14: Refusal breakdown by question type for Rule Set 7 using Qwen2.5-0.5B Instruct.

The retrieval guardrail successfully rejected 86% of generic out-of-domain questions. In contrast, only 24% of adjacent questions were rejected at the retrieval stage, allowing the remaining 76% to reach the language model. Among the adjacent questions that passed the guardrail, approximately 12% were subsequently refused by the language model, while the remaining questions received generated answers. These results demonstrate that the primary difficulty was not detecting completely unrelated queries, but distinguishing semantically related queries for which the corpus nevertheless lacked sufficient evidence.
Figure 15 illustrates the mechanism underlying this behavior. Because adjacent questions remain semantically close to the source corpus, the retrieval stage can identify highly similar passages and allow the query to proceed. However, semantic similarity does not necessarily indicate that the retrieved passages contain evidence supporting the requested information. Consequently, the language model may receive context that is topically relevant while lacking the specific facts required to answer the

question. This creates conditions in which a fluent response can be generated despite the absence of adequate supporting evidence, thereby increasing the risk of hallucination.

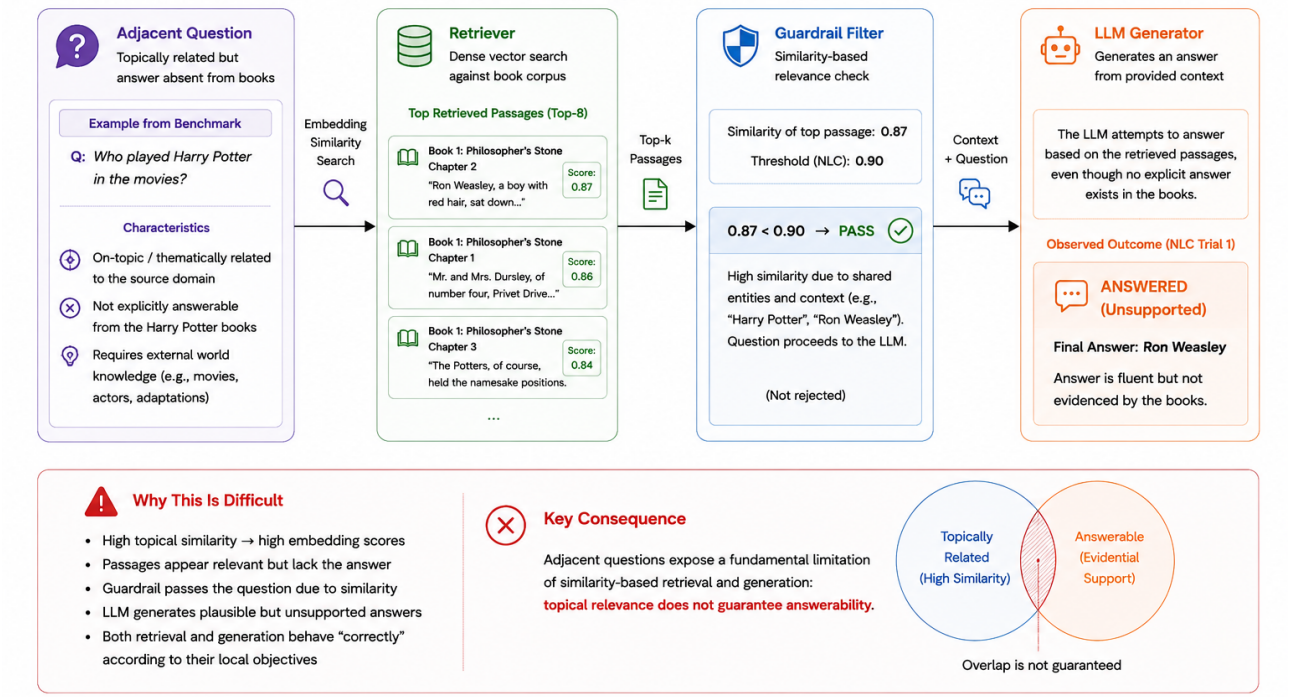


Figure 15: Illustration of the adjacent-question failure mode, in which semantically similar retrieved context does not necessarily provide evidence for the requested information.

To determine whether this behavior depended on the particular optimization stage, the proportion of adjacent questions receiving generated answers was tracked across successive rule sets.
Figure 16 presents the corresponding trend. The proportion of adjacent questions answered increased between Rule Sets 3 and 5, rising from 16% to approximately 64%. Across Rule Sets 5 through 8, however, the answered rate remained within a relatively narrow range of approximately 56–64%, despite continued retrieval-threshold optimization and multiple prompt refinements.

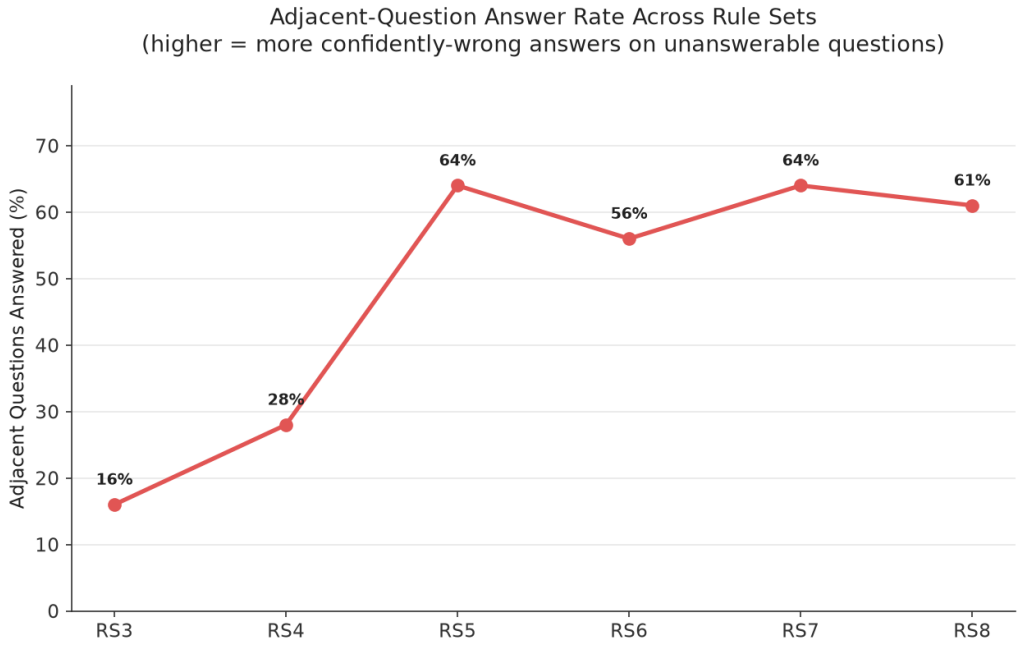


Figure 16: Adjacent-question answered rate across Rule Sets 3–8.

Although the later rule sets introduced progressively refined prompting strategies, the adjacent-question answered rate remained broadly stable during the final stages of optimization. This persistence contrasts with the improvements observed in lexical-overlap metrics and in-domain acceptance, suggesting that conventional answer-quality metrics alone do not capture the specific failure mode associated with unsupported but semantically related questions.

The persistent adjacent-question failure raises an important question: can more explicit prompting encourage the language model to rely more strictly on retrieved evidence? To investigate this possibility, Rule Set 8 introduced a stronger groundedness instruction while leaving the retrieval configuration unchanged. This experiment was designed to isolate the effect of prompt wording on the model's willingness to answer when the retrieved context did not provide sufficient evidence.

Figure 17 summarizes the category-level outcomes for Rule Set 8, while Figure 18 shows the corresponding response time.

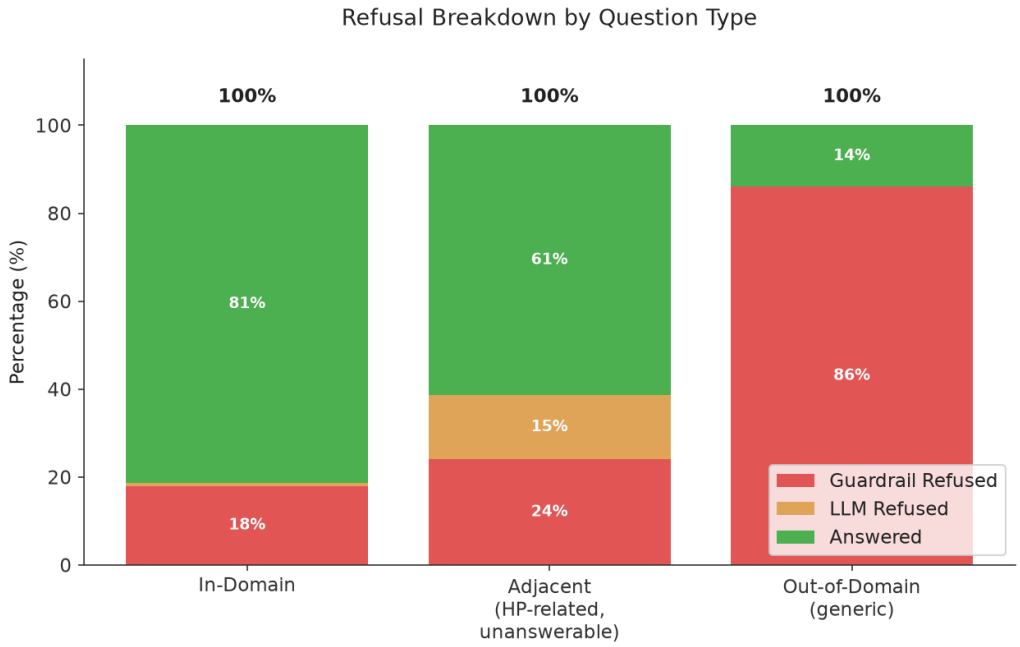


Figure 17: Refusal breakdown by question type for Rule Set 8 using Qwen2.5-0.5B Instruct.

Compared with Rule Set 7, the stricter groundedness instruction produced only a modest increase in adjacent-question refusals, from 12% to approximately 15%. At the same time, aggregate lexical-overlap metrics remained effectively unchanged, with Token F1 remaining at 0.26 and ROUGE-L decreasing slightly from 0.25 to 0.24.

The stricter prompt also increased average response latency from approximately 21 seconds to 26 seconds per question. Because the retrieval configuration was identical between the two rule sets, the observed difference in latency and the modest change in refusal behavior are attributable to the modified generation configuration rather than to changes in retrieval-stage behavior.

Overall, the additional groundedness instruction provided only limited improvement in handling adjacent questions while introducing a measurable latency penalty. More importantly, the persistence of unsupported answers across the optimization sequence indicates that prompt refinement and retrieval-threshold tuning alone were insufficient to resolve this failure mode. This observation motivated the subsequent model-scaling experiment, which examined whether increased language-model capacity could improve the model's ability to recognize when semantically relevant context was insufficient to support an answer. Consequently, Rule Set 7 was retained as the reference configuration for the model-scale comparison.

4.7 Discussion

The experiments presented in this study indicate that retrieval-stage and generation-stage failures arise from different mechanisms and therefore respond differently to system interventions. Although

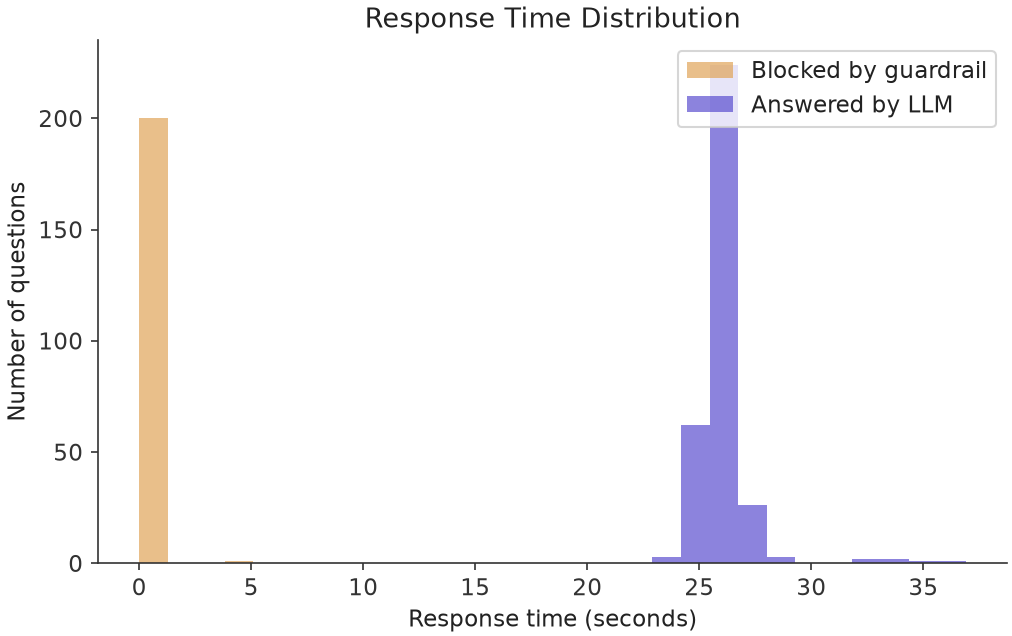


Figure 18: Response time for Rule Set 8.

both can ultimately result in incorrect system behavior, the results suggest that retrieval relevance and generation-level answerability should be treated as distinct aspects of a RAG system rather than as manifestations of a single limitation.

The retrieval guardrail operates primarily on semantic similarity between the input question and the indexed document chunks. Across the evaluated thresholds, modifying the guardrail threshold changed the balance between accepting in-domain questions and rejecting non-book-content questions, but no evaluated threshold simultaneously optimized both objectives. This trade-off was also reflected in the statistical analysis: compared with the NLC baseline, Rule Set 7 produced a statistically significant reduction in appropriate refusal behavior on non-book-content questions ($p < 0.001$), despite increasing the proportion of in-domain questions accepted by the retrieval stage. Adjacent questions expose this limitation particularly clearly because they are designed to remain semantically close to the source corpus while lacking evidential support within it. The retrieval stage can therefore identify highly similar passages even when those passages do not contain the information required to answer the question. From the perspective of the retrieval model, topical similarity and answerability are consequently not equivalent properties.

The language model exhibits a different pattern. Once an adjacent question passes the retrieval stage, successful refusal depends on the model recognizing that the retrieved context does not provide sufficient support for the requested information. Prompt refinement produced only modest changes in this behavior, whereas increasing model size under an otherwise identical retrieval pipeline, guardrail configuration, and prompt substantially increased adjacent-question refusal. The guardrail analysis confirmed that this difference was statistically significant: the appropriate refusal rate increased from 69.3% for Qwen2.5-0.5B to 80.0% for Qwen2.5-1.5B ($p = 0.0078$). These observations suggest that the evaluated pipeline contains at least two distinct decision boundaries: one determining whether retrieved content is sufficiently similar to the query and another determining whether that content provides sufficient evidence to justify an answer. Improving the first boundary through threshold tuning does not necessarily improve the second.

The persistence of adjacent-question failures across successive optimization configurations provides further evidence that this behavior was not confined to a single unsuccessful configuration. After Rule Set 5, several prompt revisions motivated by different hypotheses about the source of the failure, including modifications intended to encourage shorter responses, stronger reliance on retrieved evidence, and explicit refusal when supporting information was unavailable, left the proportion of adjacent questions receiving generated answers within a relatively narrow range for the remainder of the

optimization process. The final prompt evaluated in Rule Set 8 represented the most direct attempt to address this failure mode by explicitly instructing the model to avoid information unsupported by the retrieved context. Nevertheless, only a modest change in refusal behavior was observed, accompanied by increased latency and no meaningful improvement in aggregate lexical-overlap metrics. Within the evaluated design space, these results suggest diminishing returns from further threshold and prompt optimization for this particular failure mode.

The comparison between Qwen2.5-0.5B and Qwen2.5-1.5B provides a complementary perspective. With the retrieval pipeline, guardrail configuration, and prompt held constant, the larger model refused a substantially greater proportion of adjacent questions while incurring a considerable increase in inference latency. However, this improvement in refusal behavior did not translate into improved answer quality on the in-domain benchmark. Human evaluation showed that answer accuracy decreased from 17% for Qwen2.5-0.5B to 6% for Qwen2.5-1.5B, and the difference was statistically significant according to the exact McNemar test ($p = 0.0034$). Thus, increased model capacity improved conservative behavior on unsupported questions but was accompanied by a significant reduction in human-evaluated accuracy on answerable book-content questions. This trade-off is consistent with the broader observation that improved refusal behavior does not necessarily imply improved task performance.

The model-scale result should nevertheless be interpreted cautiously. Only two models from the same Qwen2.5 model family were evaluated, and the experiment therefore does not establish a general scaling law. Rather, it provides evidence that model capacity can influence answerability decisions within the particular resource-constrained RAG setting examined in this study. The observed increase in refusal behavior is consistent with prior work investigating model uncertainty and the ability of language models to recognize the limits of their knowledge [22], but broader conclusions would require comparisons across additional model families and parameter scales.

These findings are broadly consistent with previous research on hallucination and answerability in language models. Retrieval augmentation improves factual grounding by conditioning generation on external evidence [1], but retrieval alone cannot guarantee that generated responses are supported by that evidence [6]. The adjacent-question benchmark used in this study is conceptually related to the unanswerable-question setting introduced by SQuAD 2.0 [20], in which systems must distinguish answerable questions from questions for which no answer is supported by the available evidence. The present setting extends this distinction to a retrieval-augmented pipeline: rather than evaluating only whether a retrieved passage contains an answer, the system must determine whether the retrieved evidence collectively provides sufficient support for generation. This highlights a practical distinction between retrieval relevance and evidential sufficiency that is particularly important for RAG systems.

From a practical perspective, the results suggest that retrieval-threshold optimization should be viewed primarily as a mechanism for selecting an operating point on the retrieval trade-off rather than as a complete solution to unsupported answer generation. The statistical analysis reinforces this interpretation: although Rule Set 7 achieved higher in-domain acceptance, its guardrail behavior on non-book-content questions was significantly worse than that of the NLC baseline. Similarly, prompt engineering remains useful for controlling response characteristics such as answer length and lexical-overlap metrics, but within the evaluated design space it did not substantially resolve the adjacent-question failure mode. Increasing model size improved appropriate refusal behavior significantly, but this benefit came with a substantial computational cost, with average response latency increasing from approximately 21 seconds to approximately 61 seconds under the evaluated hardware conditions. Consequently, deployment decisions for resource-constrained RAG systems require balancing computational cost, answer quality, and the desired level of grounded refusal behavior.

The conclusions of this study are subject to several limitations. First, the experiments evaluated a single domain-specific corpus based on the Harry Potter book series, one embedding model (all-MiniLM-L6-v2), and two language models from the Qwen2.5 family. The findings therefore should not be interpreted as universal properties of RAG systems or language models. Second, the adjacent-question category contained only 25 questions, meaning that individual examples can have a noticeable effect on the reported percentages. The benchmark was also constructed around a single fictional domain, and the extent to which the observed behavior transfers to domains such as biomedical, legal, or enterprise

knowledge bases remains unknown. Third, the primary automatic metrics measure lexical agreement rather than factual groundedness and therefore cannot directly determine whether a generated answer is fully supported by the retrieved evidence. Human evaluation or explicit evidence-grounding metrics would provide a more direct assessment. Finally, latency measurements were obtained on a single consumer CPU-only system, so the absolute response times should be interpreted as representative of the evaluated resource-constrained environment rather than as hardware-independent measurements. The early-stage experimental anomaly discussed in Section 3.3 also represents a reproducibility limitation. The transition between the NLC baseline and Rule Set 1 exhibited an unexpected change in guardrail behavior despite no corresponding tracked modification to the retrieval logic. Although subsequent version control and experimental logging eliminated similar inconsistencies, this early discrepancy could not be reconstructed retrospectively and is therefore reported for transparency.

Overall, the experiments support a distinction between retrieval relevance and evidential sufficiency as separate aspects of grounded generation. Improving semantic retrieval can change which questions are admitted to the generation stage, but it does not necessarily provide the language model with the ability to determine whether the retrieved evidence is sufficient to answer those questions. Within the design space evaluated in this study, this distinction helps explain why retrieval-threshold and prompt optimization exhibited diminishing returns on adjacent questions, while increased model capacity produced a statistically significant improvement in refusal behavior at the cost of lower human-evaluated answer accuracy and substantially higher inference latency.

5 Conclusion

This study investigated the effects of retrieval-threshold optimization, prompt refinement, and model capacity on grounded answer generation in a resource-constrained Retrieval-Augmented Generation system. A lightweight RAG pipeline based on the Harry Potter book series was evaluated using a 175-question benchmark containing in-domain, out-of-domain, and topically adjacent but unanswerable questions.

The results demonstrate that retrieval-threshold optimization primarily changes the operating point between accepting in-domain questions and rejecting unsupported queries. Although increasing the threshold improved in-domain acceptance, it also reduced appropriate refusal of non-book-content questions, and the difference in guardrail behavior between the NLC baseline and Rule Set 7 was statistically significant ($p < 0.001$). Prompt refinement improved response characteristics such as answer length and lexical-overlap metrics, but produced only limited changes in the persistent adjacent-question failure mode. These findings indicate that retrieval relevance alone is insufficient to determine whether retrieved evidence is adequate to support an answer.

Model scaling produced a different trade-off. Increasing the model size from Qwen2.5-0.5B to Qwen2.5-1.5B significantly increased appropriate refusal behavior on non-book-content questions, from 69.3% to 80.0% ($p = 0.0078$), and substantially increased refusal of adjacent questions. However, this improvement was accompanied by a significant decrease in human-evaluated in-domain answer accuracy, from 17% to 6% ($p = 0.0034$), as well as an increase in average response latency from approximately 21 seconds to 61 seconds. Thus, greater model capacity did not provide a uniformly better RAG system; instead, it shifted the system toward more conservative behavior at the cost of answer quality and computational efficiency.

The study is limited by its use of a single domain-specific corpus, one embedding model, two models from the same model family, and a relatively small adjacent-question benchmark. The results should therefore be interpreted as evidence within the evaluated design space rather than as a universal characterization of RAG systems. Future work should evaluate additional domains, retrieval architectures, embedding models, and language model families, while also considering larger benchmarks, human groundedness evaluation, reranking, confidence estimation, and dedicated answerability verification mechanisms.

Overall, the findings support a distinction between retrieval relevance and evidential sufficiency as separate challenges in grounded generation. Improving semantic retrieval can determine which queries reach the generation stage, but it does not guarantee that the retrieved evidence is sufficient to answer

those queries. Similarly, increasing model capacity can improve conservative refusal behavior but may introduce substantial computational costs and trade-offs in answer quality. Reliable RAG systems may therefore require explicit mechanisms for determining evidence sufficiency in addition to conventional retrieval optimization and prompt-based control.

Acknowledgments

The author would like to acknowledge the use of Claude (Anthropic) for assistance with codebase development and debugging, as well as for editing and refining the text of this manuscript. Additionally, ChatGPT (OpenAI) was used to generate three figures included in this paper: the system architecture overview, the benchmark overview, and the illustration of adjacent questions. The codebase developed for this study is available at `https://github.com/AmirmasoudCS/Harry-Potter-RAG.git` (private repository, to be made public upon acceptance).

References

[1] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S. and Kiela, D., et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks", Advances in Neural Information Processing Systems, Vol. 33, pp. 9459-9474, 2020.

[2] Guu, K., Lee, K., Tung, Z., Pasupat, P. and Chang, M., "REALM: Retrieval-Augmented Language Model Pre-Training", Proceedings of the 37th International Conference on Machine Learning, pp. 3929-3938, 2020.

[3] Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D. and Yih, W., "Dense Passage Retrieval for Open-Domain Question Answering", Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pp. 6769-6781, 2020.

[4] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L. and Polosukhin, I., "Attention Is All You Need", Advances in Neural Information Processing Systems, Vol. 30, pp. 5998-6008, 2017.

[5] Devlin, J., Chang, M., Lee, K. and Toutanova, K., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, pp. 4171-4186, 2019.

[6] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Madotto, A. and Fung, P., "Survey of Hallucination in Natural Language Generation", ACM Computing Surveys, Vol. 55, No. 12, pp. 1-38, 2023.

[7] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G. and Askell, A., et al., "Language Models are Few-Shot Learners", Advances in Neural Information Processing Systems, Vol. 33, pp. 1877-1901, 2020.

[8] Kaplan, J., McCandlish, S., Henighan, T., Brown, T.B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J. and Amodei, D., "Scaling Laws for Neural Language Models", arXiv preprint, arXiv:2001.08361, 2020.

[9] Reimers, N. and Gurevych, I., "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks", Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, pp. 3982-3992, 2019.

[10] Wang, W., Wei, F., Dong, L., Bao, H., Yang, N. and Zhou, M., "MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-trained Transformers", Advances in Neural Information Processing Systems, Vol. 33, pp. 5776-5788, 2020.

[11] Johnson, J., Douze, M. and Jégou, H., "Billion-Scale Similarity Search with GPUs", IEEE Transactions on Big Data, Vol. 7, No. 3, pp. 535-547, 2019.

[12] Thakur, N., Reimers, N., Rücklé, A., Srivastava, A. and Gurevych, I., "BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models", Proceedings of the 35th Conference on Neural Information Processing Systems, Datasets and Benchmarks Track, 2021.

[13] Papineni, K., Roukos, S., Ward, T. and Zhu, W., "BLEU: a Method for Automatic Evaluation of Machine Translation", Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 311-318, 2002.

[14] Lin, C., "ROUGE: A Package for Automatic Evaluation of Summaries", Text Summarization Branches Out: Proceedings of the ACL-04 Workshop, pp. 74-81, 2004.

[15] Izacard, G. and Grave, E., "Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering", Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pp. 874-880, 2021.

[16] Asai, A., Wu, Z., Wang, Y., Sil, A. and Hajishirzi, H., "Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection", arXiv preprint, arXiv:2310.11511, 2023.

[17] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K. and Ray, A., et al., "Training Language Models to Follow Instructions with Human Feedback", Advances in Neural Information Processing Systems, Vol. 35, pp. 27730-27744, 2022.

[18] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q. and Zhou, D., "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models", Advances in Neural Information Processing Systems, Vol. 35, pp. 24824-24837, 2022.

[19] Rajpurkar, P., Zhang, J., Lopyrev, K. and Liang, P., "SQuAD: 100,000+ Questions for Machine Comprehension of Text", Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 2383-2392, 2016.

[20] Rajpurkar, P., Jia, R. and Liang, P., "Know What You Don't Know: Unanswerable Questions for SQuAD", Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Vol. 2, pp. 784-789, 2018.

[21] Petroni, F., Piktus, A., Fan, A., Lewis, P., Yazdani, M., De Cao, N., Thorne, J., Jernite, Y., Karpukhin, V. and Maillard, J., et al., "KILT: a Benchmark for Knowledge Intensive Language Tasks", Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics, pp. 2523-2544, 2021.

[22] Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Dodds, Z.H., DasSarma, N. and Tran-Johnson, E., et al., "Language Models (Mostly) Know What They Know", arXiv preprint, arXiv:2207.05221, 2022.

[23] Zhang, H., Diao, S., Lin, Y., Fung, Y., Lian, Q., Wang, X., Chen, Y., Ji, H. and Zhang, T., "R-Tuning: Instructing Large Language Models to Say 'I Don't Know"', arXiv preprint, arXiv:2311.09677, 2023.

[24] Chase, H., "LangChain: Building Applications with LLMs through Composability", GitHub repository, available at: https://github.com/langchain-ai/langchain, 2022.

[25] Qwen Team, Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D. and Huang, F., et al., "Qwen2.5 Technical Report", arXiv preprint, arXiv:2412.15115, 2025.

[26] Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J. and Wang, H., "Retrieval-Augmented Generation for Large Language Models: A Survey", arXiv preprint, arXiv:2312.10997, 2023.

[27] Sanchez, G., "Harry Potter Books Dataset", GitHub repository, available at: https://github.com/gastonstat/harry-potter-data, 2024.

[28] vapit, "HarryPotterQA", Hugging Face Datasets, available at: https://huggingface.co/datasets/vapit/HarryPotterQA, 2024.

[29] McNemar, Q., "Note on the Sampling Error of the Difference Between Correlated Proportions or Percentages", Psychometrika, Vol. 12, No. 2, pp. 153-157, 1947.

[30] Dietterich, T.G., "Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms", Neural Computation, Vol. 10, No. 7, pp. 1895-1923, 1998.