A Comprehensive Evaluation of Large Language Models for Retrieval-Augmented Generation under Noisy Conditions

Anonymous ACL submission

Abstract

Retrieval-Augmented Generation (RAG) has emerged as an effective strategy to ground Large Language Models (LLMs) with reliable, real-time information. This paper investigates the trade-off between cost and performance by evaluating 13 LLMs within a RAG pipeline for the Question Answering (Q&A) task under noisy retrieval conditions. We assess four extractive and nine generative models—spanning both open- and closed-source ones of varying sizes—on a journalistic benchmark specifically designed for RAG. By systematically varying the level of noise injected into the retrieved context, we analyze not only which models perform best, but also their robustness to noisy input. Results show that large open-source generative models (approx. 70B parameters) achieve performance and noise tolerance on par with top-tier closed-source models. However, their computational demands limit their practicality in resource-constrained settings. In contrast, medium-sized open-source models (approx. 7B parameters) emerge as a compelling compromise, balancing efficiency, robustness, and accessibility.

1 Introduction

Large Language Models (LLMs) have experienced a notable surge in development and adoption in recent years. They have been achieving exceptional results across a wide range of tasks, especially in natural language generation tasks such as summarization, conversation, and translation, but also in natural language understanding tasks such as sentiment analysis, text classification, and linguistic inference, among others (Chang et al., 2024). For tasks like question answering, two primary approaches have been established. The first is extractive, where models operate with precision by identifying and returning exact spans of text from a given context (Ai et al., 2024). The second is generative, where LLMs are leveraged for their capability to produce novel text based on input prompts (Sun et al., 2023).

Although LLMs encode substantial parametric knowledge—acquired through the optimization of transformer-based neural networks—this knowledge is typically sufficient only for answering open-domain questions whose answers were present during training. In closed-domain settings, where the required information is domain-specific and often absent from the training data, parametric knowledge alone is often insufficient (Tonmoy et al., 2024). To bridge this gap, Retrieval-Augmented Generation (RAG) systems have emerged as a promising and effective architecture. RAG strategies enhance LLMs by integrating external document retrieval into the generation process, enabling models to produce more grounded and factual outputs based on up-to-date or domain-specific information (Gao et al., 2023).

The primary motivation for RAG is to address one of the key challenges in deploying LLMs in real-world applications: the issue of hallucination. This issue becomes particularly pressing in corporate environments, where language models often handle sensitive information and tend to generate non-factual content (Gao et al., 2023). In this context, RAG systems are especially suitable for the Question Answering (Q&A) task, as they operate under the assumption that reliable information resides in external databases. Consequently, the generative model is instructed to rely solely on retrieved documents as the source of truth, bypassing its internal parametric knowledge on the target topic (Lin et al., 2023; Tonmoy et al., 2024).

Generative LLMs are frequently employed in RAG systems due to their high accuracy and their ability to abstain from answering when the provided context is insufficient. However, these models often entail substantial computational and financial costs, and retrieval components seldom achieve perfect accuracy. This underscores the importance

of evaluating how varying levels of noise in the retrieved context affect the performance of LLMs in RAG settings.

In this study, we conduct a comprehensive evaluation of LLMs for Retrieval-Augmented Generation (RAG) by:

1. assessing the Accuracy, F1-Score, Response Relevancy and Faithfulness of language models under varying levels of noise in the retrieved context;

2. analyzing the associated costs, including the computational demands of open-source models and the financial implications of closed-source commercial alternatives;

3. and examining how model size impacts the trade-off between robustness, hallucination, and resource efficiency.

The source code for this study will be made publicly available in a GitHub repository (after the review process).

We compare the performance of four extractive and six generative open-source models of varying sizes (ranging from 3.8 billion to 70 billion parameters), and three closed-source generative models, using the Retrieval-Augmented Generation Benchmark (RGB) dataset (Chen et al., 2024). Our results show that it is possible to replace generative models with smaller extractive ones when the retrieval procedure is sufficiently accurate. Additionally, we show that replacing closed-source models with open-source alternatives—when computational resources allow for 70B parameter models—yields comparable accuracy and noise robustness. In scenarios with more limited resources, 7B parameter models emerge as a promising alternative, offering competitive accuracy at the expense of reduced robustness to noise.

2 Related Work

The quality of retrieved documents is an important factor in the performance of RAG systems. As demonstrated by Perçin et al. (2025), if the retriever fails to locate correct information, the LLM lacks relevant context, likely resulting in an incorrect answer. The effect of noise—defined as passages that are superficially relevant but lack the correct answer (Fang et al., 2024)—is particularly significant.[1] Recent work shows that RAG systems are vulnerable to the effects of distraction from noisy contexts, where the LLM component can be easily misled into generating an incorrect answer (Amiraz et al., 2025).

In response, several benchmark datasets have been developed to introduce realistic, noisy scenarios for evaluating RAG systems. Notable examples include CRAG (Comprehensive RAG Benchmark) (Yang et al., 2024), MIRAGE (Metric-Intensive Benchmark for Retrieval-Augmented Generation Evaluation) (Park et al., 2025), and RGB (Retrieval-Augmented Generation Benchmark) (Chen et al., 2024).

Among these, the RGB dataset is distinguished by its inclusion of questions, one or more gold-standard answers, and a collection of documents categorized as either positive (containing relevant information) or negative (containing distractors or unrelated content). Consequently, the RGB dataset not only allows for the evaluation of RAG systems in the presence of noise but also enables control over the level of noise introduced to the model by altering the proportion of positive and negative documents provided.

The choice of evaluation metrics is particularly critical when assessing RAG performance in noisy settings. RAG systems are typically evaluated using metrics such as accuracy and F1-score. However, in noisy contexts, it is important to include metrics that can quantify the performance degradation caused by noise. Park et al. (2025) propose a custom metric, Noise Vulnerability, to measure the performance difference of the entire RAG system between noisy and noise-free contexts.

Furthermore, metrics from the RAGAS (Retrieval-Augmented Generation Assessment) framework are well-suited for evaluating performance in noisy environments. Unlike traditional methods that rely on gold-standard answers, RAGAS leverages large language models to evaluate generated responses based on criteria such as Response Relevancy—how thoroughly the answer addresses the user's question—and Faithfulness—how well the answer remains grounded in the retrieved context (Es et al., 2024; Roychowdhury et al., 2024).

These dimensions are especially important in noisy settings, where different failure modes can

[1] Fang et al. (2024) distinguish between three types of noise: relevant noise, where passages are superficially relevant but lack the correct answer; irrelevant noise, where passages are on entirely different topics; and counterfactual noise, where passages contain misleading information. In this study, we focus on relevant noise.

emerge. For example, responses may seem topically appropriate but lack grounding in the retrieved evidence, indicating that the model is relying on its internal, parametric knowledge rather than the provided documents (Zhang et al., 2024; Longpre et al., 2022). In other cases, a model might generate responses that are faithful to the retrieved context but fail to answer the question because the retrieved passages themselves are irrelevant or off-topic (Amiraz et al., 2025). By capturing both the alignment with context (Faithfulness) and the relevance to the user's query (Response Relevancy), RAGAS makes these distinct failure patterns visible, offering a nuanced picture of system behavior under noisy retrieval.

Many studies have explored the comparative performance of extractive models, open-source generative models, and closed-source generative models in Q&A (Pearce et al., 2021; Gaikwad et al., 2022; Luo et al., 2022; Mallick et al., 2023; Jayakumar et al., 2023; Cadena et al.; Tan et al., 2023; Ai et al., 2024). However, these studies typically rely on standard Q&A benchmark datasets such as the Stanford Question Answering Dataset (SQuAD), MultiSpanQA, or domain-specific datasets like COVIDQA. Consequently, they do not account for the effect of noise in their performance evaluations.

While a body of recent literature does address the effect of noise within RAG systems (Park et al., 2025; Liang et al., 2025; Fang et al., 2024; Yang et al., 2024), these studies often have a narrow scope, evaluating a limited number of language models—predominantly closed-source generative models—and treating noise as a dichotomous variable instead of a graded factor. This limitation is a direct consequence of using datasets such as CRAG or MIRAGE, which, unlike RGB, do not permit granular control over noise levels.

Additionally, the choice of metrics presents a similar limitation, as most studies rely on traditional or task-specific scores (e.g., RAGQuestEval from Lyu et al. (2024)) that are not designed to capture the nuanced effects of noise on generation.[2] Metrics from the RAGAS framework, such as Faithfulness and Response Relevancy, offer a more fine-grained evaluation by using LLMs to assess the relevance and consistency of generated responses in noisy contexts.

Finally, a critical gap in the existing literature is the lack of analysis of computational costs associated with varying model sizes in RAG systems under noisy conditions. Prior studies fail to assess the computational resources required to process noisy contexts across different model architectures. This omission hinders a comprehensive understanding of the practical trade-offs for deploying RAG systems in resource-constrained settings.

[2]The Noise Vulnerability metric from Park et al. (2025) is a notable exception, though it focuses on binary noise presence at the system level.

3 Experimental Setup

Models: This study evaluates a diverse set of models for question answering, including extractive (all of them open-source), open-source generative models, and closed-source generative models, as detailed in Table 1. Model sizes are shown in millions (M) or billions (B) of parameters. The size of the closed-source models is not publicly disclosed.

Model	Type	Size	Reference
DistilBERT	Extractive	65 M	(Sanh et al., 2019)
BERT Multicased	Extractive	178 M	(Romero, 2020)
BERT Uncased	Extractive	335 M	(Devlin et al., 2018)
RoBERTa	Extractive	560 M	(Pietsch et al., 2019)
Phi-3 Mini	Gen. Open	3.8 B	(Microsoft, 2024)
GPT4All	Gen. Open	13 B	(Anand et al., 2023)
Nous Hermes 2	Gen. Open	7 B	(NousResearch, 2024)
Nous Hermes 3	Gen. Open	70 B	(Teknium et al., 2024)
Meta LLaMA 3	Gen. Open	8 B	(Meta, 2024a)
Meta LLaMA 3.1	Gen. Open	70 B	(Meta, 2024b)
GPT-3.5 Turbo	Gen. Closed	N/A	–
GPT-4o Mini	Gen. Closed	N/A	–
GPT-4o	Gen. Closed	N/A	–

Table 1: Models evaluated in this study. Gen. = Generative. Size in parameters (M = million, B = billion).

This selection of closed-source models was based on those currently made available by our company, reflecting the options effectively accessible within our institutional environment.

Dataset: To evaluate performance, we utilized the Retrieval-Augmented Generation Benchmark (RGB) dataset (Chen et al., 2024), which comprises 300 questions, each accompanied by a list of correct answers and a set of positive (relevant) and negative (irrelevant) context documents. As stated in Section 2, this dataset allows for the assessment of model robustness under varying levels of noise, where noise is defined as the proportion of negative documents included in the context. Five noise levels were tested: 0%, 20%, 40%, 60%, and 80%.

Since some open-source models accept a maximum of 1,500 tokens as context, this limit was imposed across all models to ensure fairness. Document ordering within each context was random-

ized using a fixed, reproducible seed. Each model was evaluated across five independent runs, with different randomized distributions of positive and negative documents in each run. Within a single run, all models shared the same randomized context. Final performance values are reported as the mean across these five runs, along with standard deviation values to reflect variability.

Hardware: The models were evaluated using different hardware configurations based on their computational requirements. The extractive models and smaller generative models were tested using a single NVIDIA V100 GPU (32 GB RAM). The larger generative models (Meta LLaMA 3.1 70B and Nous Hermes 3 70B) were evaluated using eight NVIDIA V100 GPUs (32 GB RAM each) to accommodate their higher computational demands. In contrast, the closed-source models (GPT-3.5 Turbo, GPT-4o Mini, and GPT-4o) were accessed via an external Azure endpoint provided by our company, whose specifications are undisclosed. It is important to note that the inference times of these closed-source models may be affected by external factors, such as Azure's rate limits and network latency.

Metrics: We evaluated model performance using a combination of traditional and modern Question-Answering (Q&A) metrics. To assess basic correctness, we employed Accuracy and F1-score. Accuracy is computed at the answer level by normalizing the predicted and gold responses—removing punctuation, lowercasing, and tokenizing on whitespace. A prediction is marked as correct only if it contains all substrings that are required given the dataset correct answer, regardless of order, following the method described in (Mallen et al., 2023). F1-score captures partial correctness by computing the harmonic mean of precision (the proportion of relevant tokens in the prediction) and recall (the proportion of relevant tokens recovered from the gold answer), as defined in (Chhablani et al., 2021). In this case, tokens are the substrings required to make a correct answer given the dataset reference answer.

For a more nuanced assessment, we incorporated the *Response Relevancy* and *Faithfulness* metrics from the RAGAS framework (Es et al., 2024; Amiraz et al., 2025). These metrics are particularly crucial in noisy contexts, as they can distinguish between answers that are relevant but not factually grounded in the source context and those that are faithful to the context but fail to fully address the question. The RAGAS metrics require a complete gold-standard answer for comparison. However, the RGB dataset provides only a list of required strings rather than a full reference answer. To overcome this limitation, we adopted an LLM-as-a-judge approach (Snell et al., 2022; Wang et al., 2023; Muller et al., 2025) and used the answers generated by GPT-4o as the reference for each question. Consequently, all other models were evaluated against the GPT-4o responses. A necessary implication of this methodology is that GPT-4o's own performance on these specific metrics could not be assessed. The RAGAS scores were computed using GPT-4o Mini as the evaluator model, which judged the quality and factual alignment of each generated output against the GPT-4o reference.

To quantify the impact of noise on LLM performance, we introduce the Δ Accuracy metric, inspired by prior work on noise sensitivity in Q&A models (Havrilla and Iyer, 2024). This metric measures performance degradation by calculating the difference in accuracy between the baseline (0% noise) and maximum (80% noise) conditions. A smaller Δ Accuracy value signifies greater robustness against contextual noise.

Prompt: For generative models, a standardized prompt was employed to guide responses during inference. The prompt instructed the model to read the provided context carefully and generate the most accurate and concise answer to the given question:

```
You are an AI assistant specializing in Question
     Answering. Your task is to read the
    provided context carefully and then generate
     the most accurate and concise answer to the
     question based on the context.

Context: {context}

Question: {question}

Answer:
```

4 Results

Table 2 and Figure 1 report the accuracy obtained by the three groups of models—extractive, open-source generative, and closed-source generative—under five noise conditions (0, 20, 40, 60 and 80%). Standard deviations, shown in parentheses, are consistently small, indicating that random re-samplings of positive and negative documents

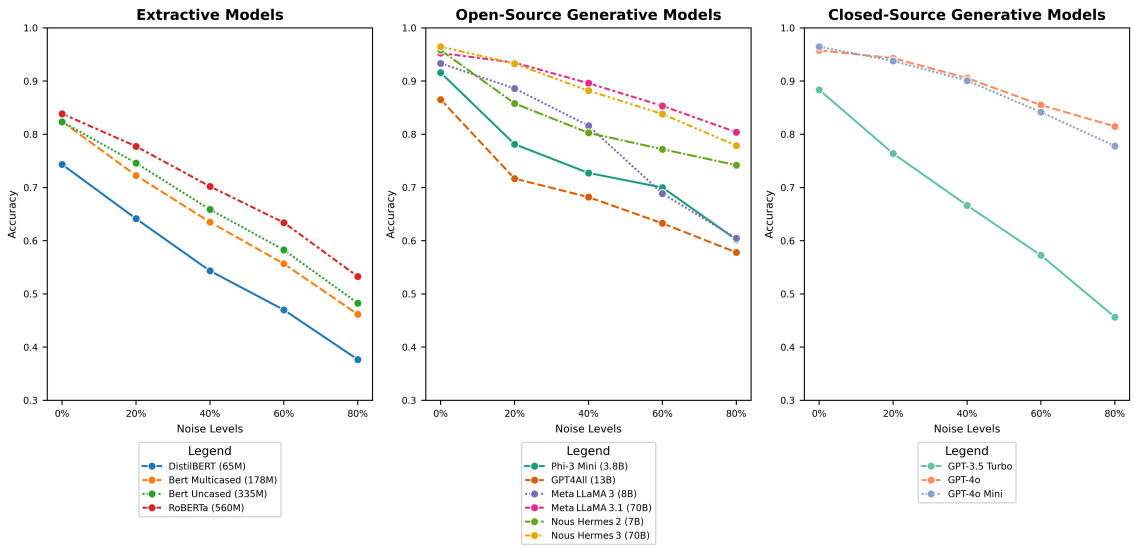

Figure 1: Accuracy across different noise levels

have little impact on the measurements. Figure 2 depicts the F1-score metric.

Figure 3 plots the two RAGAS metrics—Response Relevancy and Faithfulness—across the same noise spectrum.[3] Finally, Table 3 presents the average inference time for each model.

5 Discussion

A clear pattern emerges when comparing results for accuracy (Figure 1): closed-source generative models achieve the highest accuracy and are the most resilient to noise, followed by open-source generative models and, finally, extractive models. Within each group, a secondary but not universal trend is visible—larger parameter counts generally lead to higher accuracy. Among closed-source models, GPT-4o stands out, while Meta LLaMA 3.1 (70B) and Nous Hermes 3 (70B) lead the open-source group, and RoBERTa is the strongest among extractive baselines. Notably, the large open-source models, such as Hermes 3 (70B) and Meta LLaMA 3.1 (70B), exhibit accuracy and noise resistance comparable to the best-performing closed-source models (GPT-4o and GPT-4o Mini).

The size–performance correlation is not absolute. GPT4All (13B) performs consistently worse, and degrades faster under noise, than smaller models such as Phi-3Mini (3.8B), Meta LLaMA-3 (8B) and Nous Hermes-2 (7B). Likewise, Meta LLaMA-3 (8B) exhibits lower robustness to noise than the lighter Phi-3 Mini and Nous Hermes-2 despite its larger size.

Extractive models earn higher F1-scores (Figure 2) at low noise but erode faster than generatives as noise increases. Because F1 is token-overlap-based, the longer outputs typical of generative models share fewer exact tokens with the reference answers, leading to lower values—even when their semantic content is correct. However, while the F1 metric may under-represent the performance of generative models, particularly in noisy contexts, it is notable that open-source generative models—specifically Meta LLaMA 3.1 (70B)—demonstrate performance and noise robustness comparable to that of closed-source models like GPT-4o and GPT-4o Mini.

For Response Relevancy (Figure 3), extractive models hover around 0.80 throughout, with a barely perceptible downward slope; RoBERTa is the lowest but follows the same flat profile. This behaviour is expected: span-prediction models return a literal substring, so as long as the gold answer remains in the passage, topical relevance is preserved.

Open-source generative models display a more heterogeneous pattern when using the Response Relevancy and Faithfulness metrics. They start above 0.80 but decline more sharply with noise; Meta LLaMA-3 (8B) and Phi-3 Mini (3.8B) fall to about 0.60 at 80% noise. Larger models (70B) mitigate this drop thanks to greater capacity for instruction-following and distraction filtering, whereas smaller models are prone to copying irrel-

[3]The figure reports scores for every model we tested except GPT-4o, because the models answers are evaluated against GPT-4o's answers as reference.

Model	Size	Noise Levels					Δ acc.
		0%	20%	40%	60%	80%	
Extractive Models							
DistilBERT	65M	0.743±0.012	0.641±0.009	0.543±0.034	0.470±0.022	0.377±0.015	0.367
BERT Multicased	178M	0.824±0.009	0.723±0.016	0.635±0.014	0.557±0.025	0.462±0.025	0.362
BERT Uncased	335M	0.823±0.020	0.746±0.016	0.659±0.019	0.583±0.022	0.483±0.011	0.341
RoBERTa	560M	**0.839**±0.006	**0.777**±0.030	**0.702**±0.012	**0.634**±0.019	**0.533**±0.030	**0.306**
Open-Source Generative Models							
Phi-3 Mini	3.8B	0.916±0.009	0.781±0.017	0.727±0.006	0.700±0.017	0.602±0.039	0.314
GPT4All	13B	0.865±0.008	0.717±0.012	0.682±0.012	0.633±0.015	0.578±0.051	0.287
Nous Hermes 2	7B	0.958±0.008	0.858±0.018	0.803±0.025	0.772±0.020	0.742±0.019	0.216
Nous Hermes 3	70B	**0.965**±0.006	0.933±0.008	0.882±0.009	0.838±0.016	0.779±0.010	0.186
Meta LLaMA 3	8B	0.933±0.014	0.886±0.017	0.816±0.018	0.689±0.018	0.605±0.033	0.329
Meta LLaMA 3.1	70B	0.953±0.006	**0.935**±0.011	**0.896**±0.008	**0.853**±0.012	**0.804**±0.014	**0.149**
Closed-Source Generative Models							
GPT-3.5 Turbo	N/A	0.884±0.013	0.764±0.012	0.666±0.032	0.573±0.028	0.456±0.026	0.427
GPT-4o Mini	N/A	**0.965**±0.003	0.938±0.004	0.901±0.007	0.842±0.021	0.778±0.013	0.186
GPT-4o	N/A	0.958±0.005	**0.943**±0.007	**0.906**±0.025	**0.855**±0.015	**0.815**±0.017	**0.143**

Table 2: Accuracy comparison across noise levels (mean ± standard deviation).

Model	Size	Device	Count	Exec. Time (sec.)
Extractive Models				
DistilBERT	65M	GPU	1	**0.08** (± 0.01)
BERT Multicased	178M	GPU	1	0.10 (± 0.01)
BERT Uncased	335M	GPU	1	0.20 (± 0.01)
RoBERTa	560M	GPU	1	0.23 (± 0.01)
Open-Source Generative Models				
Phi-3 Mini	3.8B	GPU	1	4.52 (± 0.09)
GPT4All	13B	GPU	1	9.26 (± 0.15)
Meta LLaMA 3	8B	GPU	1	5.28 (± 0.10)
Meta LLaMA 3.1	70B	GPU	8	**1.16** (± 0.44)
Nous Hermes 2	7B	GPU	1	5.64 (± 0.12)
Nous Hermes 3	70B	GPU	8	1.26 (± 0.43)
Closed-Source Generative Models				
GPT-3.5 Turbo	N/A	N/A	N/A	**0.67** (± 3.24)
GPT-4o Mini	N/A	N/A	N/A	**0.67** (± 0.39)
GPT-4o	N/A	N/A	N/A	0.86 (± 0.69)

Table 3: Average Query Execution Time by Model (mean ± standard deviation). "Device" indicates CPU or GPU and "Count" the number of units used.

evant fragments once attention is diluted.

Among the closed-source generative models that do appear in the figure, two distinct trends can be observed. On one hand, the GPT-4o Mini model maintains the highest and most stable relevance curve, consistently staying above 0.80. This performance is very similar to that of large open-source generative models (those with 70B parameters). On the other hand, GPT-3.5 Turbo shows a sharp decline, dropping to around 0.50, which reflects its smaller effective context window and weaker alignment.

Faithfulness reveals different group dynamics. Extractive models decline steadily and homogeneously from about 0.80 to 0.50 as noise reaches 80%. Under heavy noise and multi-span questions, they can select spans that no longer correspond lexically to the reference answer produced by GPT-4o, hence the steeper loss.

Among open-source generators, the large 70B variants (Meta LLaMA-3.1 and Nous Hermes 3) are the most stable (from 0.80 to 0.70). Mid-sized models such as Meta LLaMA-3 (8B) and GPT4All (13B) trace similar slopes but start from lower baselines (from 0.55 to 0.40). Small models (Nous Hermes 2 7B, Phi-3 Mini 3.8B) drop abruptly at the first noise level (20%), then continue a gentler decline—an effect also documented by Ming et al. (2025), who show that smaller LLMs hallucinate more readily when confronted with distractors.

Among open-source generative models, the large 70B variants (Meta LLaMA 3.1 and Nous Hermes 3) demonstrate the greatest stability, with their scores declining moderately from 0.80 to 0.70. Mid-sized models, such as Meta LLaMA 3 (8B) and GPT4All (13B), follow similar downward trends but start from lower baseline scores, ranging from 0.55 to 0.40. In contrast, small models (Nous Hermes 2 7B and Phi-3 Mini 3.8B) experience a sharp initial drop at the 20% noise level, followed by a more gradual decline. This behavior is consistent with prior findings (Ming et al., 2025), which indicate that smaller language models are more prone to hallucinations when exposed to distractors.

Closed-source generative models exhibit a similar trend in the Faithfulness metric as they do in Response Relevancy. GPT-4o Mini consistently main-

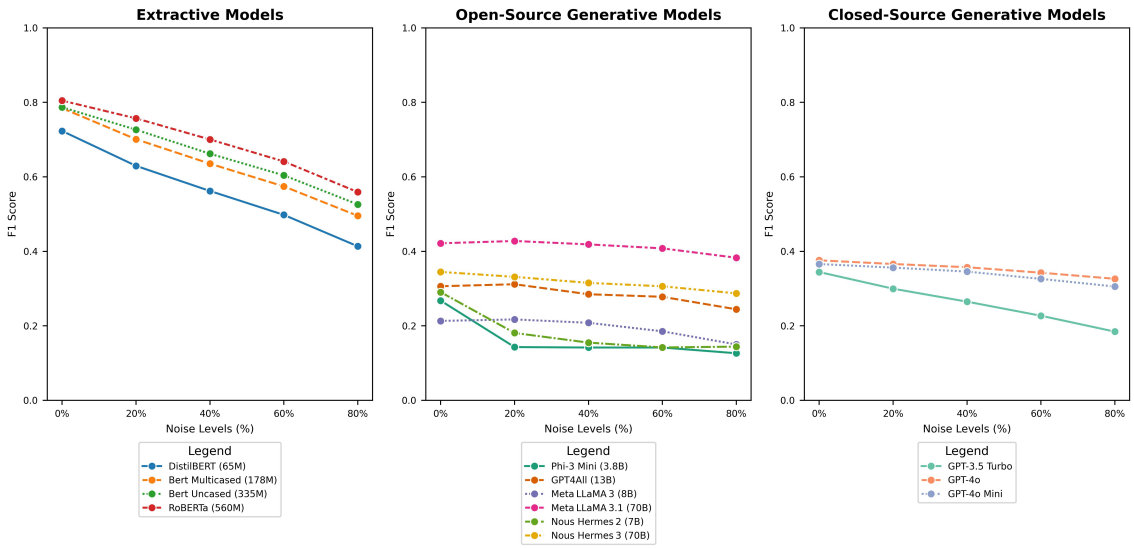

Figure 2: F1-score across different noise levels

tains a score above 0.70 across all scenarios, displaying a pattern comparable to that of large open-source generative models. In contrast, the Faithfulness score of GPT-3.5 Turbo declines rapidly, dropping to approximately 0.40 when noise reaches 80% noise.

Regarding inference time (Table 3), our results reveal three distinct resource profiles. Extractive models, such as RoBERTa, achieved the fastest inference times, requiring only a single GPU for efficient execution (e.g., 0.23 seconds for RoBERTa). In contrast, medium-sized models, such as Meta LLaMA 3 (8B) and Nous Hermes 2 (7B), operated efficiently with a single GPU, achieving inference times of 5.28 and 5.64 seconds, respectively. Finally, the largest open-source generative models, Meta LLaMA 3.1 (70B) and Nous Hermes 3 (70B), required eight GPUs to achieve competitive performance, with reduced inference times of 1.16 and 1.26 seconds, respectively.[4]

6 Concluding Remarks

This study offers three key insights into the performance, efficiency, and practical use of extractive and generative language models for Question Answering (Q&A) in Retrieval-Augmented Generation (RAG) systems.

First, regarding performance and noise robustness, large open-source generative models such as Meta LLaMA 3.1 (70B) and Nous Hermes 3 (70B) demonstrate performance comparable to the best closed-source models, including GPT-4o and GPT-4o Mini. These open-source models maintain high accuracy across various noise levels, indicating their resilience even in challenging conditions. This result highlights that open-source alternatives can deliver competitive performance without the constraints of proprietary solutions.

Second, achieving this performance with large open-source models comes with significant hardware requirements. In our environment, 70B-parameter open-source models necessitated their distribution across eight 32 GB NVIDIA V100 GPUs. In contrast, medium-sized open-source models (Meta LLaMA 3 8B, Nous Hermes 2 7B) can operate effectively on a single GPU, while closed-source APIs offload computational demands to external servers. This observation underscores the trade-off between model size and operational cost.

Third, the choice between large and medium-sized generative models should be guided by the available computational resources and budget. Large open-source models are well-suited for environments with ample computational infrastructure, offering a cost-efficient alternative to closed-source models. In contrast, medium-sized open-source models present a practical solution for resource-constrained settings, delivering strong accuracy with significantly lower hardware consumption.

[4]It should be noted that, as mentioned in Section 3, the closed-source generative models were executed via an external Azure endpoint. As a result, their inference times are affected by external factors such as network latency and Azure's rate limits, making them not directly comparable to the locally executed models.

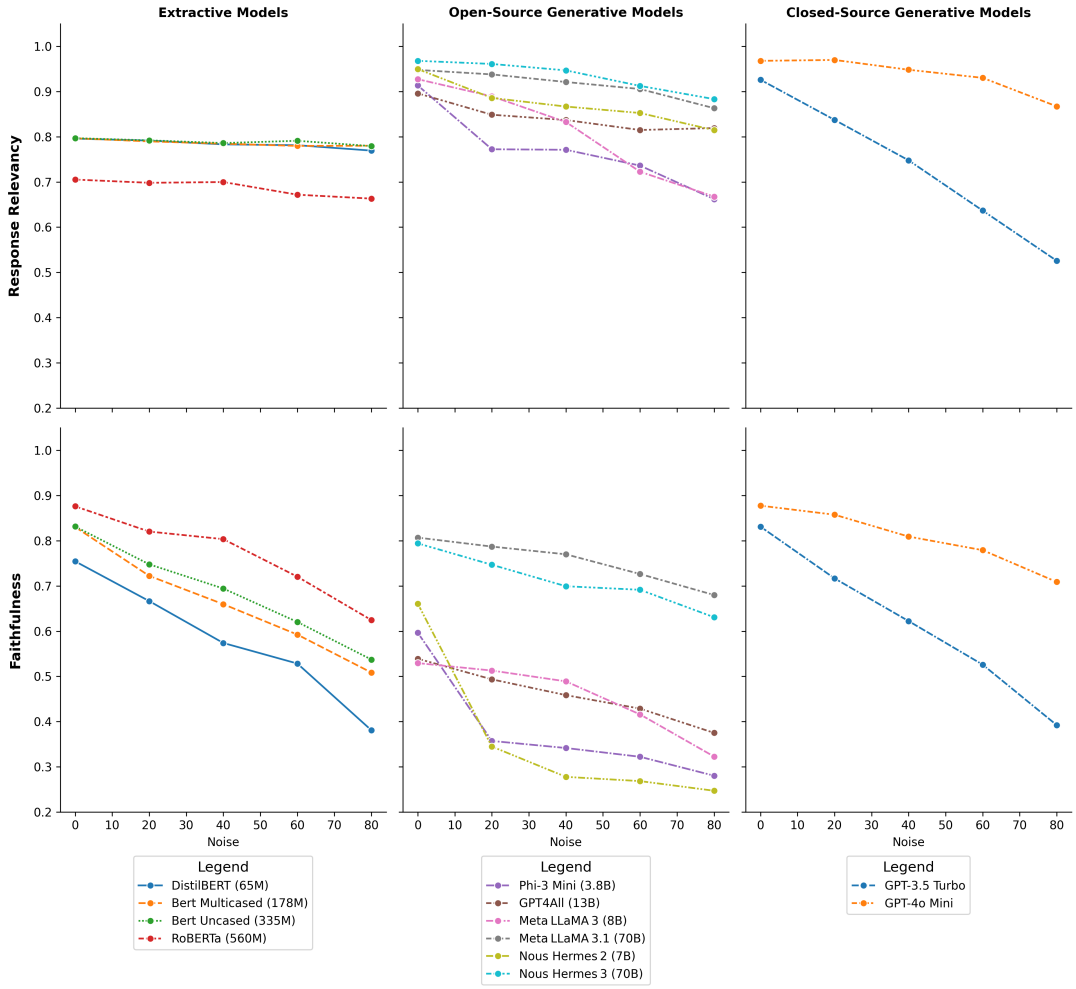

Figure 3: Response Relevancy and Faithfulness across different noise levels

Limitations

We acknowledge a few limitations in our study that should be considered when interpreting the results.

Our evaluation, while covering diverse model categories, was constrained in its selection of closed-source models. We analyzed three proprietary models, as these were the only ones accessible through our organization's internal model hub. A direct consequence of this constraint is the inability to conduct a financial cost analysis for these models, as granular usage metrics and associated pricing were not available to us. Therefore, our cost-related conclusions are primarily focused on the computational demands of open-source models.

A second limitation arises from the potential for data contamination in the benchmark dataset. The dataset, published in 2024, is composed of general-domain news articles. It is plausible that contemporary, continuously updated generative models (such as the closed-source models evaluated) may have encountered this data, or information related to it, during their training cycles. This could confer an unfair advantage, as this prior exposure might influence generation despite the instruction to rely solely on the provided context.

Finally, a specific methodological limitation pertains to the evaluation of the GPT-4o model on metrics of *Response Relevancy* and *Faithfulness*. This is because we utilize answers generated by GPT-4o itself as the ground-truth reference for responses, as the benchmark lacks independently verified or human-curated answers for each question. Evaluating GPT-4o against its own output would create a circular reference, leading to artificially perfect scores on these metrics. Consequently, we had to exclude GPT-4o from this portion of the analysis to maintain methodological validity.

References

Lin Ai, Zheng Hui, Zizhou Liu, and Julia Hirschberg. 2024. Enhancing pre-trained generative language models with question attended span extraction on machine reading comprehension. *Preprint*, arXiv:2404.17991.

Chen Amiraz, Florin Cuconasu, Simone Filice, and Zohar Karnin. 2025. The distracting effect: Understanding irrelevant passages in rag. *Preprint*, arXiv:2505.06914.

Yuvanesh Anand, Zach Nussbaum, Brandon Duderstadt, Benjamin Schmidt, and Andriy Mulyar. 2023. Gpt4all: Training an assistant-style chatbot with large scale data distillation from gpt-3.5-turbo. `https://github.com/nomic-ai/gpt4all`.

Angel Cadena, Gerardo Sierra, Jorge Lázaro, and Sergio-Luis Ojeda-Trueba. Information retrieval techniques for question answering based on pre-trained language models.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2024. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.*, 15(3).

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2024. Benchmarking large language models in retrieval-augmented generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):17754–17762.

Gunjan Chhablani, Abheesht Sharma, Harshit Pandey, Yash Bhartia, and Shan Suthaharan. 2021. Nlrg at semeval-2021 task 5: Toxic spans detection leveraging bert-based token classification and span prediction techniques. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2024. RAGAs: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–158, St. Julians, Malta. Association for Computational Linguistics.

Feiteng Fang, Yuelin Bai, Shiwen Ni, Min Yang, Xiaojun Chen, and Ruifeng Xu. 2024. Enhancing noise robustness of retrieval-augmented language models with adaptive adversarial training. *Preprint*, arXiv:2405.20978.

Arya Gaikwad, Palash Rambhia, and Sarthak Pawar. 2022. An extensive analysis between different language models: Gpt-3, bert and macaw. *Research Square*.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Alex Havrilla and Maia Iyer. 2024. Understanding the effect of noise in llm training data with algorithmic chains of thought. *Preprint*, arXiv:2402.04004.

Thanmay Jayakumar, Fauzan Farooqui, and Luqman Farooqui. 2023. Large language models are legal but they are not: Making the case for a powerful legalllm. *Preprint*, arXiv:2311.08890.

Xun Liang, Simin Niu, Zhiyu Li, Sensen Zhang, Hanyu Wang, Feiyu Xiong, Jason Zhaoxin Fan, Bo Tang, Shichao Song, Mengwei Wang, and Jiawei Yang. 2025. Saferag: Benchmarking security in retrieval-augmented generation of large language model. *Preprint*, arXiv:2501.18636.

Xi Victoria Lin, Xilun Chen, Mingda Chen, Weijia Shi, Maria Lomeli, Rich James, Pedro Rodriguez, Jacob Kahn, Gergely Szilvasy, Mike Lewis, and 1 others. 2023. Ra-dit: Retrieval-augmented dual instruction tuning. *arXiv preprint arXiv:2310.01352*.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2022. Entity-based knowledge conflicts in question answering. *Preprint*, arXiv:2109.05052.

Man Luo, Kazuma Hashimoto, Semih Yavuz, Zhiwei Liu, Chitta Baral, and Yingbo Zhou. 2022. Choose your qa model wisely: A systematic study of generative and extractive readers for question answering. *Preprint*, arXiv:2203.07522.

Yuanjie Lyu, Zhiyu Li, Simin Niu, Feiyu Xiong, Bo Tang, Wenjin Wang, Hao Wu, Huanyong Liu, Tong Xu, and Enhong Chen. 2024. Crud-rag: A comprehensive chinese benchmark for retrieval-augmented generation of large language models. *Preprint*, arXiv:2401.17043.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. *Preprint*, arXiv:2212.10511.

Prabir Mallick, Tapas Nayak, and Indrajit Bhattacharya. 2023. Adapting pre-trained generative models for extractive question answering. *Preprint*, arXiv:2311.02961.

Meta. 2024a. Llama 3 model card.

Meta. 2024b. Llama 3.1 70b.

Microsoft. 2024. Phi-3-mini-4k.

Yifei Ming, Senthil Purushwalkam, Shrey Pandit, Zixuan Ke, Xuan-Phi Nguyen, Caiming Xiong, and Shafiq Joty. 2025. Faitheval: Can your language model stay faithful to context, even if "the moon is made of marshmallows". *Preprint*, arXiv:2410.03727.

Sacha Muller, António Loison, Bilel Omrani, and Gautier Viaud. 2025. Grouse: A benchmark to evaluate evaluators in grounded question answering. *Preprint*, arXiv:2409.06595.

NousResearch. 2024. Nous hermes 2 mistral 7b dpo.

Chanhee Park, Hyeonseok Moon, Chanjun Park, and Heuiseok Lim. 2025. Mirage: A metric-intensive benchmark for retrieval-augmented generation evaluation. *Preprint*, arXiv:2504.17137.

Kate Pearce, Tiffany Zhan, Aneesh Komanduri, and Justin Zhan. 2021. A comparative study of transformer-based language models on extractive question answering. *Preprint*, arXiv:2110.03142.

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. *Preprint*, arXiv:2507.06956.

Malte Pietsch, Timo Möller, Bogdan Kostic, Julian Risch, Massimiliano Pippi, Mayank Jobanputra, Sara Zanzottera, Silvano Cerza, Vladimir Blagojevic, Thomas Stadelmann, Tanay Soni, and Sebastian Lee. 2019. Haystack: the end-to-end NLP framework for pragmatic builders.

Manuel Romero. 2020. Multilingual XLM-RoBERTa large for QA on various languages.

Sujoy Roychowdhury, Sumit Soman, H G Ranjani, Neeraj Gunda, Vansh Chhabra, and Sai Krishna Bala. 2024. Evaluation of rag metrics for question answering in the telecom domain. *Preprint*, arXiv:2407.12873.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. In *NeurIPS EMC^{2}Workshop*.

Charlie Snell, Dan Klein, and Ruiqi Zhong. 2022. Learning by distilling context. *Preprint*, arXiv:2209.15189.

Kaiser Sun, Peng Qi, Yuhao Zhang, Lan Liu, William Yang Wang, and Zhiheng Huang. 2023. Tokenization consistency matters for generative models on extractive nlp tasks. *Preprint*, arXiv:2212.09912.

Yiming Tan, Dehai Min, Yu Li, Wenbo Li, Nan Hu, Yongrui Chen, and Guilin Qi. 2023. Can chatgpt replace traditional kbqa models? an in-depth analysis of the question answering performance of the gpt llm family. *Preprint*, arXiv:2303.07992.

Ryan Teknium, Jeffrey Quesnelle, and Chen Guang. 2024. Hermes 3 technical report. *Preprint*, arXiv:2408.11857.

SM Tonmoy, SM Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint arXiv:2401.01313*.

Tianlu Wang, Ping Yu, Xiaoqing Ellen Tan, Sean O'Brien, Ramakanth Pasunuru, Jane Dwivedi-Yu, Olga Golovneva, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. 2023. Shepherd: A critic for language model generation. *Preprint*, arXiv:2308.04592.

Xiao Yang, Kai Sun, Hao Xin, Yushi Sun, Nikita Bhalla, Xiangsen Chen, Sajal Choudhary, Rongze Daniel Gui, Ziran Will Jiang, Ziyu Jiang, Lingkun Kong, Brian Moran, Jiaqi Wang, Yifan Ethan Xu, An Yan, Chenyu Yang, Eting Yuan, Hanwen Zha, Nan Tang, and 8 others. 2024. Crag - comprehensive rag benchmark. In *Advances in Neural Information Processing Systems*, volume 37, pages 10470–10490. Curran Associates, Inc.

Ruizhe Zhang, Yongxin Xu, Yuzhen Xiao, Runchuan Zhu, Xinke Jiang, Xu Chu, Junfeng Zhao, and Yasha Wang. 2024. Knowpo: Knowledge-aware preference optimization for controllable knowledge selection in retrieval-augmented language models. *Preprint*, arXiv:2408.03297.