Pandora's Box or Aladdin's Lamp: A Comprehensive Analysis Revealing the Role of RAG Noise in Large Language Models

Anonymous ACL submission

Abstract

Retrieval-Augmented Generation (RAG) has emerged as a crucial method for addressing hallucinations in large language models (LLMs). While recent research has extended RAG models to complex noisy scenarios, these explorations often confine themselves to limited noise types and presuppose that noise is inherently detrimental to LLMs, potentially deviating from real-world retrieval environments and restricting practical applicability. In this paper, we define seven distinct noise types from a linguistic perspective and establish a Noise RAG Benchmark (NoiserBench), a comprehensive evaluation framework encompassing multiple datasets and reasoning tasks. Through empirical evaluation of eight representative LLMs with diverse architectures and scales, we reveal that these noises can be further categorized into two practical groups: noise that is beneficial to LLMs (aka beneficial noise) and noise that is harmful to LLMs (aka harmful noise). While harmful noise generally impairs performance, beneficial noise may enhance several aspects of model capabilities and overall performance. Our analysis offers insights for developing robust RAG solutions and mitigating hallucinations across diverse retrieval scenarios.

1 Introduction

Large language models (LLMs) (OpenAI, 2023; Meta, AI, 2024) have demonstrated remarkable proficiency across various tasks (Bubeck et al., 2023). Despite impressive capabilities, LLMs face challenges such as reliance on outdated knowledge and hallucination (Huang et al., 2023; Kandpal et al., 2023). Retrieval-Augmented Generation (RAG) has recently emerged as a promising approach to mitigate these limitations (Lewis et al., 2020; Gao et al., 2024). RAG enhances LLMs' performance by augmenting inputs with additional information retrieved from external sources during inference.

However, external sources often contain various non-standard noises, including fake news, outdated content, spelling errors, and data contamination, which may potentially influence model performance (Shi et al., 2023; Xie et al., 2024). It is crucial to explore how noise affects RAG systems and understand the underlying mechanisms.

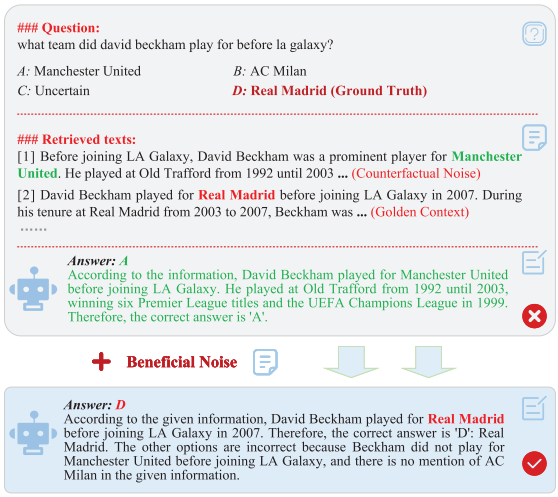

Figure 1: An example from NoiserBench illustrating effects of different RAG noises. Initially, the model is misled by counterfactual noise. Interestingly, upon introducing beneficial noise, it successfully discriminates between correct and incorrect information and produces the accurate answer 'D'.

Recently, several studies (Chen et al., 2024; Xiang et al., 2024) have attempted to extend RAG systems to complex real-world scenarios, investigating the impact of noisy documents and strategies to enhance the system's robustness. For example, Cuconasu et al. (2024) defines three types of noise in retrieved documents and examines their impacts on LLMs. Despite highlighting one noise's positive effect, the study lacks a comprehensive noise definition and in-depth investigation of underlying principles. Fang et al. (2024) applies adversarial training to dynamically adjust the model's training process in response to retrieval noises. RobustRAG (Xiang et al., 2024) proposes a defense framework

to improve the robustness of RAG models against retrieval corruption attacks. Nevertheless, these investigations typically focus on a limited number of noise types (usually no more than three) and lack clear classification, which fails to fully capture the complexity of real-world noise environments. Additionally, these studies often assume that noise is harmful, neglecting its potential positive effects and lacking systematic evaluation datasets. As shown in Figure 1, introducing beneficial noise allows the LLMs to avoid the harmful effects of counterfactual noise, concentrate on the golden context, and produce accurate responses. Thus, there's an urgent need to redefine and describe noisy scenarios in RAG, and systematically explore the specific impacts of retrieval noises.

In this paper, we conduct a comprehensive analysis to reveal the role of RAG noises in LLMs. We first define seven types of noise from a linguistic perspective. Based on this definition, we propose a systematic framework to create diverse noisy documents and establish NoiserBench, a novel noise RAG benchmark. Then, we evaluate eight representative LLMs with different architectures and scales. Extensive results show that RAG noises can be categorized into two practical groups: *beneficial noise* (semantic, datatype, illegal sentence) and *harmful noise* (counterfactual, supportive, orthographic, prior). While harmful noise impairs performance, beneficial noise surprisingly enhances model capabilities and leads to improved performance. Further analysis reveals that beneficial noise facilitates more standardized answer formats, clearer reasoning paths, and increases confidence in responses with golden context. These contrasting effects are analogous to *opening Pandora's Box* (harmful noise) versus *unlocking Aladdin's Lamp* (beneficial noise). We hope this study will advance efforts to mitigate harmful noise and leverage the positive effects of beneficial noise in future research. Our main contributions are:

- We define seven types of noise and categorize them into two groups: beneficial and harmful noise. This is the first comprehensive study to define and assess RAG noises from both linguistic and practical perspectives.

- We introduce a novel framework for constructing diverse retrieval documents and create NoiserBench, a benchmark that effectively simulates real-world noise in RAG models.

- Evaluated on multiple datasets and LLMs, our results reveal that while some RAG noises (e.g. counterfactual) can open Pandora's Box and cause errors, beneficial noise (e.g. datatype) has the potential to unlock the power of Aladdin's Lamp and deliver positive effects.

- Our findings redefine retrieval noise and encourage researchers to explore methods that harness its beneficial properties while addressing its harmful effects.

2 Related Work

Retrieval-Augmented Generation By integrating external information, RAG methods enhance reasoning and generation process (Gao et al., 2024; Zhao et al., 2024). Early works primarily focus on improving retrieval model performance to obtain relevant documents for subsequent generation (Qu et al., 2021; Wang et al., 2023; Zheng et al., 2024). Recent research has expanded RAG framework to real-world noisy scenarios, aiming to build robust RAG systems by enhancing the generator (Fang et al., 2024; Xiang et al., 2024). For instance, Self-RAG (Asai et al., 2024) employs four specialized tokens and GPT-4-generated instruction-tuning data to fine-tune the Llama2 model. RobustRAG (Xiang et al., 2024) proposes an isolate-then-aggregate defense framework to enhance model robustness against retrieval corruption attacks. However, these investigations are constrained by their narrow focus on specific noise types and the inherent assumption that noise is harmful, potentially hindering method generalization. This paper aims to analyze RAG noise and reveal its roles systematically.

Noise Injection in LLMs Noise injection (Grandvalet et al., 1997) in LLMs involves adding noise to inputs during training or inference, such as data augmentation (Ye et al., 2024), adversarial training (Fang et al., 2024), and prompt perturbation (Zhu et al., 2024). Recently, researchers have focused on noise injection in RAG systems (Chen et al., 2024). For example, Cuconasu et al. (2024) classifies three retrieval noises and explores their effects on LLMs. Fang et al. (2024) leverages adversarial training to dynamically adjust LLMs' training process in response to retrieval noises. However, these noise types are limited and may not reflect complex real-world scenarios. A comprehensive framework that simulates real-world noise is necessary.

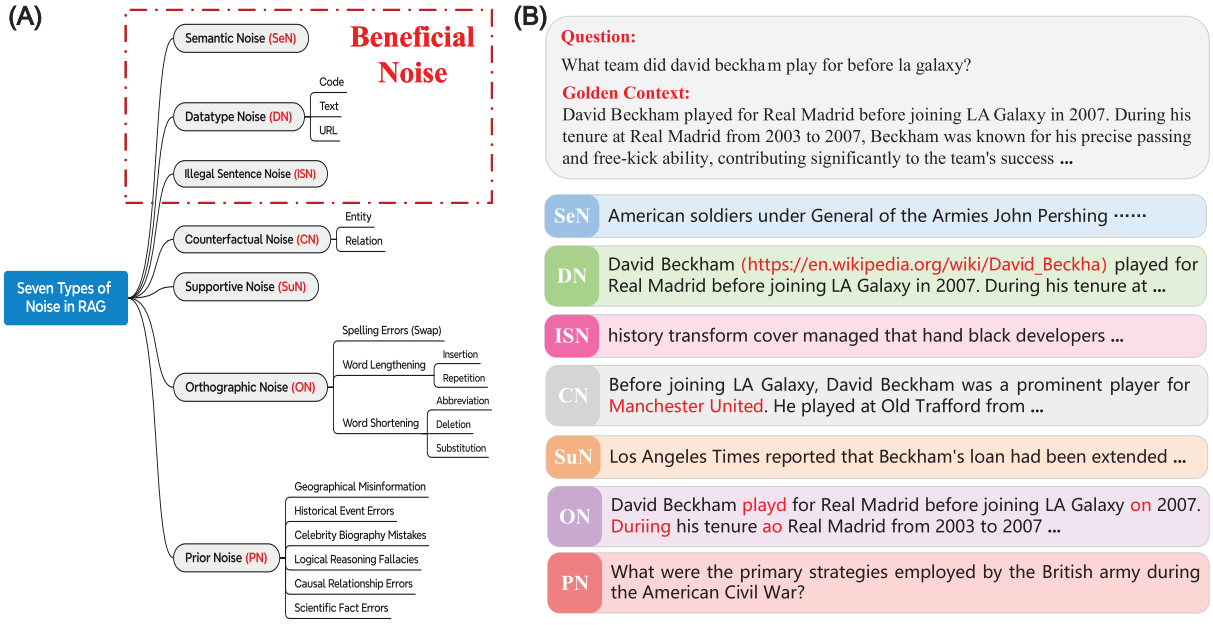

Figure 2: **(A)** Our seven RAG noise types comprehensively capture real-world retrieval challenges. **(B)** This detailed illustration intuitively depicts the diverse RAG noise landscape, with noise injection regions marked in red.

3 A Taxonomy of RAG Noise

As shown in Figure 2, we categorize RAG noise into seven linguistic types. They are further divided into beneficial (semantic, datatype, and illegal sentence) and harmful noise (counterfactual, supportive, orthographic, and prior) for practical applications. We will explain the reason behind this classification in *5 Experiment Setup*.

Semantic Noise (SeN) Retrieval documents may contain content with low semantic relevance to the query, often being off-topic or deviating from the intended meaning. Given that Warren Weaver originally defined semantic noise as "the perturbations or distortions of sentence meaning" (Shannon et al., 1961), we classify off-topic, low-semantic-relevance documents as *semantic noise*.

Datatype Noise (DN) This type of noise refers to the mixing of different data types on the web, such as the blending of links and text on Wikipedia. In this paper, we consider three data types: text, URLs, and code.

Illegal Sentence Noise (ISN) Web content may include fragments that do not form grammatically correct sentences, such as “history transform cover managed that hand black”. We define this type of noise as *illegal sentence noise*.

Counterfactual Noise (CN) The internet contains abundant false information, including fake news and outdated knowledge (Tumarkin and Whitelaw, 2001; Olan et al., 2024), presenting critical challenges to RAG systems. Drawing from linguistics, where “counterfactual" denotes statements contrary to fact (Feng and Yi, 2006), we introduce the term “*counterfactual noise*" to characterize factual errors. This concept aligns with prior research (Fang et al., 2024).

Supportive Noise (SuN) Supportive evidence, known as positive evidence, is highly semantically relevant to a hypothesis and provides necessary information to support it (Kertész and Rákosi, 2012). We introduce the term “*supportive noise*” to describe documents that exhibit high semantic relevance but lack corresponding answer information.

Orthographic Noise (ON) The word “orthography” originates from the Greek *orthós* (meaning “correct”) and *gráphein* (meaning “to write”), and refers to the way words are written in linguistics (Skeat, 1993; Aloufi, 2021). *Orthographic noise*, on the other hand, can refer to writing errors such as spelling mistakes and word lengthening.

Prior Noise (PN) In linguistics, prior knowledge refers to what a learner already knows before solving a problem (Chafe, 1971). Our study defines *prior noise* as questions based on false assumptions or premises. For example, the question “Who was the CEO of Google when they were restructured into Alphabet in 2017?” contains prior noise because the restructuring occurred in 2015, not 2017.

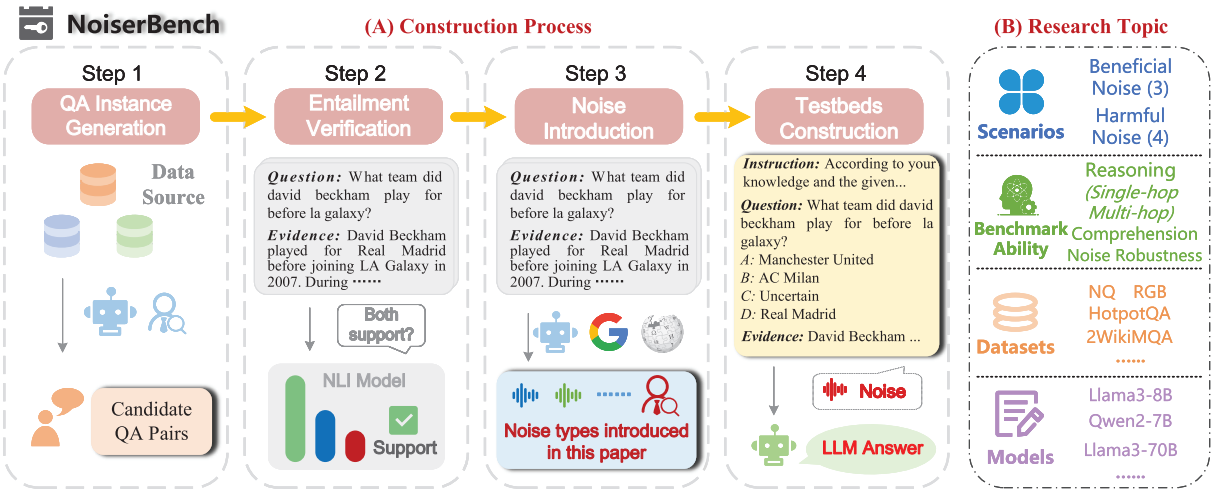

Figure 3: The overall framework for simulating the impact of real-world noise on RAG models. Initially, we generate and obtain QA instances, utilizing ChatGPT to filter out ambiguous examples (***Step 1***). Then, we perform entailment verification using NLI models to maintain evidence quality (***Step 2***). After that, we use tools like search engines to create noisy documents (***Step 3***). Finally, we transform the free-form QA into a multiple-choice QA format by providing several answer options for convenient automatic evaluation (***Step 4***). All experiments are conducted in a zero-shot setting to avoid bias from demonstrations.

4 Noise RAG Benchmark Construction

We discuss the data construction and evaluation metrics. The overall framework is illustrated in Figure 3.

4.1 Data Construction

As shown in Figure 3 (A), our framework comprises four essential steps, including QA Instance Generation, Entailment Verification, Noise Introduction and Testbeds Construction.

Step 1: QA Instance Generation For prior noise, we collect article snippets from mainstream media and Wikipedia, covering various time periods and domains such as sports, politics, and finance. We then design prompts for ChatGPT to generate relevant events, questions, and answers for each snippet. Note that the generated questions contain prior noise (factual errors), which we manually review to ensure that they are reasonably answerable by LLMs. For the remaining seven types of noise (SeN, DN, ISN, CN, SuN, ON, PN), we obtain question-answering (QA) pairs from existing datasets, following previous work (Fang et al., 2024; Cuconasu et al., 2024; Yoran et al., 2024). After obtaining candidate QA pairs, we employ ChatGPT to remove ambiguous or difficult-to-assess pairs, followed by a manual review. For example, questions like "How many companies have a market capitalization of over $25 billion and pledged to reduce greenhouse gas emissions?" should be excluded due to their broad potential answers and the dynamic market values of companies. Similar criteria are applied to other instances.

Step 2: Entailment Verification As illustrated in Xie et al. (2024); Yoran et al. (2024), effective evidence should strongly support its answer. For example, golden evidence about David Beckham should support that he played for Real Madrid before joining LA Galaxy. Therefore, we employ the natural language inference model, bart-large-mnli-407M (Lewis et al., 2019) to ensure evidence properly entails the answer. We only keep those examples with an entailment probability $p \geq 0.8$.

Step 3: Noise Introduction We construct diverse retrieval documents for noise testbeds. For counterfactual noise, we extract related entities and relations from Google search results to create counterfactual answers. ChatGPT is then employed to construct corresponding supportive evidence, followed by entailment verification. For Supportive and semantic noise, we utilize the 2018 English Wikipedia dump (Karpukhin et al., 2020) as source documents, with off-the-shelf Contriever-MS MARCO model (Izacard et al., 2022) for retrieval and the lightweight text embedding model all-MiniLM-L6-v2 (Wang et al., 2021) for semantic relevance filtering. To simulate illegal sentence noise, we construct meaningless sentences by randomly combining words from model vocabulary, mimicking real-world garbled text. Datatype noise

is created by prompting ChatGPT to insert URLs or code snippets while preserving key answer information. Finally, orthographic noise is generated using the open-source textnoisr package (Preligens Lab, 2023). This pipeline enables a comprehensive assessment of model performance across a range of noise scenarios.

Step 4: Testbeds Construction After obtaining high-quality QA instances and diverse retrieval documents, we build testbeds to evaluate model performance under various noise conditions. Given the challenges in automatically assessing LLMs' responses to open-ended QA tasks (Xie et al., 2024), we convert free-form QA into a multiple-choice format. This constrains the response space and facilitates more accurate evaluation. Specifically, for each QA pair, LLMs choose from 4 options: the correct answer, two counterfactual alternatives, and "Uncertain". The order of the golden option remains entirely random to avoid LLMs' sensitivity to option order (Wu et al., 2024).

Finally, eight datasets are obtained for Noiser-Bench. Following (Yoran et al., 2024; Wang et al., 2024), we randomly select 500 samples from each dataset as test cases or use all samples if the size of this dataset is smaller than 500.

4.2 Evaluation Metrics

This benchmark aims to reveal the role that RAG noise plays on LLMs. We use accuracy as the primary metric and also report the weighted average accuracy across datasets.

5 Experiment Setup

5.1 Datasets

We experiment with multiple QA datasets, which are categorized into four types based on the required reasoning skills:

- **Single-hop**: Questions requiring one-step reasoning. We evaluate using the Natural Questions (NQ) (Kwiatkowski et al., 2019) and RGB (Chen et al., 2024) datasets.
- **Explicit Multi-hop**: Questions where multiple reasoning steps are explicitly expressed. We utilize HotpotQA (Yang et al., 2018), 2WIKIMQA (Welbl et al., 2018) and Bamboogle dataset (Press et al., 2022).
- **Implicit Multi-hop**: Questions where intermediate steps are not explicitly stated, often requiring commonsense knowledge for implicit reasoning. We use StrategyQA (Geva et al., 2021) and TempQA (Jia et al., 2018).
- **Mixed-Hop**: Questions requiring single- or multi-hop reasoning. We use our constructed dataset, PriorQA.

5.2 Baseline Models

We evaluate eight LLMs of different architectures and scales: Llama3-Instruct (8B, 70B) (Meta, AI, 2024), Qwen2-7B-Instruct (Yang et al., 2024), Mistral (7B, 8x7B) (Jiang et al., 2023, 2024), Vicuna-13B-v1.5 (Chiang et al., 2023), Llama2-13B (Touvron et al., 2023), and Baichuan2-13B (Yang et al., 2023). This enables a comprehensive assessment of noise across various dimensions.

6 Results and Analysis

First, we examine the roles of RAG noise (6.1). While prior work has analyzed its harmful effects, we focus on its beneficial aspects (6.2). We evaluate these benefits across four dimensions: (1) Generalization across Models, (2) Noise Robustness Across Scenarios, (3) Noise Ratio Impact, and (4) Statistical Validation. Finally, we investigate the underlying mechanisms of these phenomena (6.3).

6.1 Roles of RAG Noise

Table 1 illustrates the impact of diverse noise types (the first six) on two open-source models: Llama3-8B-Instruct and Qwen2-7B-Instruct. We observe consistent performance trends across multiple datasets and retrieval noises. Based on these trends, we can categorize retrieval noises into two types: *harmful noise* (counterfactual, supportive, and orthographic) and *beneficial noise* (semantic, datatype, and illegal sentence). We find that:

(1) For harmful noise, counterfactual noise impacts model performance most significantly by disrupting accurate fact discernment and answer generation. As shown in Figure 1, the false statement "Beckham was a prominent player for Manchester United" leads the model to disregard correct information and respond erroneously.

(2) For beneficial noise, illegal sentence noise exhibits the most notable improvement in model performance. It improves accuracy by an average of 3.32% and 1.65% for two models, respectively, and consistently achieves powerful performance across diverse datasets.

Table 1: Impact of diverse noise types on accuracy (%) for Llama3-8B-Instruct and Qwen2-7b-Instruct across seven datasets. We assess performance across various retrieval scenarios: "Base" (no retrieval), "Golden Only" (only golden retrieval context), and "Golden & XXX" (golden context + specific retrieval noises, including Counterfactual, Supportive, Orthographic, Semantic, Datatype, Illegal Sentence Noise). The green and red values indicate the performance gap from "Golden Only". We also provide the weighted average accuracy for each noise type. The best two results are shown in bold and underlined.

Scenario	Single-hop		Multi-hop (Explicit)			Multi-hop (Implicit)		Average
	NQ	RGB	HotpotQA	2WikiMQA	Bamboogle	StrategyQA	TempQA	
Llama3-8B-Instruct								
Base	61.34	47.00	53.80	34.40	32.00	58.80	50.54	51.58
Golden Only	93.06	80.00	97.80	79.80	87.20	73.40	91.94	86.57
Golden & CN	58.86	36.33	44.20	21.20	61.60	43.20	67.74	45.58$_{-40.99}$
Golden & SuN	90.58	80.00	95.60	81.00	93.60	69.40	93.01	85.37$_{-1.20}$
Golden & ON	93.31	75.00	96.20	78.60	89.60	63.60	90.86	83.99$_{-2.58}$
Golden & SeN	96.53$_{+0.47}$	81.33$_{+1.33}$	98.40$_{+0.60}$	87.20$_{+7.40}$	93.60$_{+6.40}$	68.40	96.24$_{+4.30}$	88.73$_{+2.16}$
Golden & DN	93.19$_{+0.13}$	81.67$_{+1.67}$	95.00	82.00$_{+2.20}$	88.00$_{+0.80}$	**73.60**$_{+0.20}$	94.62$_{+2.68}$	86.91$_{+0.34}$
Golden & ISN	**96.65**$_{+0.65}$	**83.00**$_{+1.33}$	**98.80**$_{+1.00}$	**87.40**$_{+7.60}$	**94.40**$_{+7.20}$	72.60	**97.85**$_{+5.91}$	**89.89**$_{+3.32}$
Qwen2-7B-Instruct								
Base	58.24	31.33	50.20	22.60	31.20	42.40	40.86	43.01
Golden Only	**97.03**	76.33	98.40	78.00	94.40	67.00	94.62	86.46
Golden & CN	41.88	26.00	38.40	12.40	39.20	37.60	45.16	33.96$_{-52.50}$
Golden & SuN	90.46	74.00	96.40	80.40	92.00	64.00	90.32	83.65$_{-2.81}$
Golden & ON	95.66	74.00	97.80	80.00	91.20	54.60	94.62	83.82$_{-2.64}$
Golden & SeN	96.53	77.67$_{+1.34}$	98.80$_{+0.40}$	77.00	**96.80**$_{+2.40}$	66.80	97.31$_{+2.69}$	86.60$_{+0.14}$
Golden & DN	96.03	**84.33**$_{+9.00}$	98.20	79.60$_{+1.60}$	93.60	**71.80**$_{+4.80}$	95.70$_{+1.08}$	**88.11**$_{+1.65}$
Golden & ISN	96.65	80.00$_{+3.67}$	**99.00**$_{+0.60}$	**83.80**$_{+5.80}$	**96.80**$_{+2.40}$	66.80	**97.85**$_{+1.23}$	**88.11**$_{+1.65}$

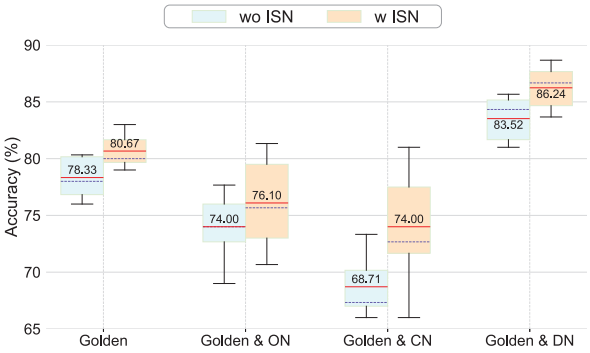

Figure 4: Impact of ISN on the average accuracy of eight representative LLMs on RGB. Red solid lines indicate means and purple dashed lines show medians.

For prior noise, we evaluate on our PriorQA dataset in Appendix Table 7. Questions in PriorQA contain factual errors, such as "Which country hosted 1980 FIFA World Cup?" (1980 FIFA World Cup was not held). Accuracy is measured by whether LLMs correctly identify and respond with "The question is factually incorrect". LLMs achieve 79.93% average accuracy in handling prior noise. However, when models fail to identify prior errors and continue retrieval, accuracy drops to 34.20%. his highlights the importance of detecting factual errors in queries before generating responses.

Table 2: Effects of beneficial noise on Self-RAG (13B). We report enhanced accuracy ratios (%), and the weighted average values (WA, %) are also provided.

Scenario	NQ	RGB	StrategyQA	WA
Golden only	+3.12	+1.74	+18.88	+7.77
Golden & DN	+1.84	+1.96	+13.50	+5.49
Golden & ON	+1.76	+3.63	+10.00	+4.67
Average	+2.24	+2.45	+14.13	+5.98

6.2 Additional Results on Beneficial Noise

Generalization across Models To demonstrate beneficial noise's broad applicability, we examine its effects across model architectures (Figure 4) and RAG configurations (Table 2). For brevity, we present illegal sentence noise results in the main text, with full results in the Appendix.

Results across various architectures and scales are shown in Figure 4, we evaluate the impact of illegal sentence noise (ISN) on eight LLMs by presenting average accuracy across scenarios with no noise, harmful noise (e.g. CN, ON), and beneficial noise (e.g. DN). We apply proportional scaling to CN data to make a clearer illustration within one figure while maintaining consistent conclusions. The results indicate that ISN significantly enhances

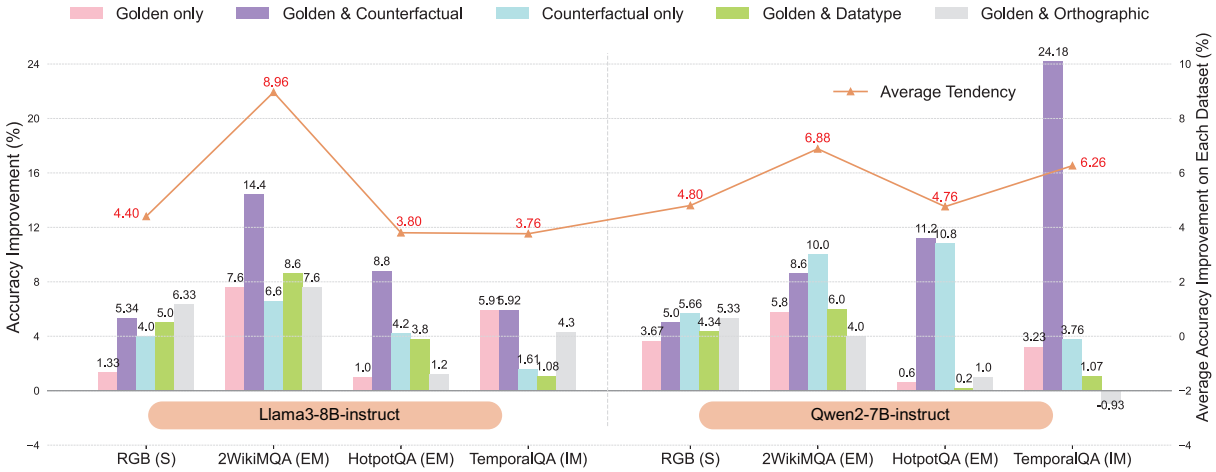

Figure 5: Results for the impact of illegal sentence noise on the Llama3-8B-instruct and Qwen2-7B-instruct models when exposed to five typical noise categories across four datasets, including both single-hop (S) and multi-hop (explicit: EM, implicit: IM) reasoning tasks. The bar charts show performance differences upon introducing illegal sentence noise. The line graphs illustrate the average accuracy improvement across noise types per dataset.

Table 3: Results for different illegal sentence noise (ISN) ratios on RGB. L2-13B, L3-8B, Q2-7B, M-7B, V-13B, B2-13B, L3-70B and M-8x7B represents Llama2-13B, Llama3-8B-Instruct, Qwen2-Instruct, Mistral-7B-Instruct-v0.2, Vicuna-13B-v1.5, Baichuan2-13B-chat, Llama3-70B-Instruct, Mixtral-8x7B-Instruct.

Scenario	Small						Large		Average
	L2-13B	L3-8B	Q2-7B	M-7B	V-13B	B2-13B	L3-70B	M-8x7B	
0	29.33	80.00	76.33	80.33	80.33	78.00	76.00	77.33	72.21
+ ISN	72.33	83.00	80.00	81.00	82.33	79.67	79.67	79.67	**79.71**$_{+7.50}$
0.2	18.67	77.33	75.33	76.00	79.33	73.33	76.67	73.67	68.79
+ ISN	73.67	82.67	80.33	76.67	80.00	72.33	80.33	73.67	**77.46**$_{+8.67}$
0.4	12.33	73.67	71.33	69.00	72.67	68.00	76.33	65.67	63.63
+ ISN	70.67	77.00	73.00	71.00	73.33	68.33	80.00	66.67	**72.50**$_{+8.87}$

model performance in all scenarios, with the most substantial improvement under harmful noise.

Noise effects on specialized RAG models are illustrated in Table 2. Introducing illegal sentence noise to the specialized RAG model Self-RAG (Asai et al., 2024) consistently enhances model performance across various datasets (NQ, RGB, and StrategyQA) and scenarios (without noise, with harmful or beneficial noise). This further validates the positive effects of beneficial noise.

Noise Robustness Across Scenarios We analyze the effect of illegal sentence noise (ISN) in 5 scenarios: no noise (i.e., Golden only), harmful noise (i.e., Golden & Counterfactual, Counterfactual only and Golden & Orthographic), and beneficial noise (i.e., Golden & Datatype). Figure 5 shows accuracy gains with ISN introduction. Results indicate consistent improvements across datasets, especially when combined with harmful noise like counterfactual, leading to an average accuracy increase of over 10%. This highlights the potential significance of beneficial noise in RAG applications.

Noise Ratio Impact To demonstrate the positive effects at different harmful noise ratios, we present results for orthographic noise disturbances with ratios ranging from 0 to 0.4. As shown in Table 3, we see that the introduction of illegal sentence noise (beneficial noise) consistently enhances model performance, thereby further illustrating the generalizability of beneficial noise.

Statistical Validation To statistically evaluate the differences between scenarios with and without beneficial noise, we apply the nonparametric Wilcoxon signed-rank test (Kotz and Johnson, 1992). This method effectively measures the magnitudes of differences and detects statis-

Table 4: Statistical significance of differences between scenarios with and without beneficial noises.

Noise	Llama3-8B-Instruct	Qwen2-7B-Instruct
ISN	4.10e-5	4.88e-3
DN	1.71e-4	9.59e-4

tical significance between two conditions. We test the null hypothesis of no significant difference ($H_0 : difference = 0$) against the alternative hypothesis of a significant difference ($H_1 : difference \neq 0$). Following (Seth et al., 2023; Wu et al., 2023), we use a significance level of 0.05. As shown in Table 4, all p-values are below 0.05, leading us to reject the null hypothesis (H_0). These results provide strong statistical evidence that beneficial noise improves model performance.

6.3 Analysis of Noise Phenomena

We propose three hypotheses regarding how beneficial noise may enhance performance, which we confirm through case study and statistical analysis.

- ***H1: Clearer reasoning process***
- ***H2: More standardized response formats***
- ***H3: Increased confidence with gold context***

Illustrative Case Study Table 13 in the appendix presents the reasoning process of Llama3-8B-instruct on the multi-hop dataset Bamboogle. Without beneficial noise, the model ignores correct information and exhibits logical flaws under counterfactual noise influence. This is exemplified by its erroneous statement: "The other options are incorrect, as they provide different birth dates for the author." However, upon introducing beneficial noise, the model exhibits heightened attention to the golden context and successfully distinguishes between correct and incorrect information (***H1***). We hypothesize that beneficial noise enhances the LLMs' ability to integrate its parameterized knowledge with retrieved information, thus improving its capacity to discern truth from falsehood. Furthermore, by comparing model outputs under two conditions, we observe that beneficial noise contributes to more standardized answer formats (***H2***).

Statistical Characterization To verify three hypotheses statistically, we use a two-step process. We first gather model outputs from multiple datasets before and after introducing beneficial noise. Then, we randomly sample 100 examples per dataset to manually assess which condition produces more standardized output formats and clearer reasoning processes. Outputs are deemed similar if no significant difference exists between conditions with and without beneficial noise. Results across seven datasets show that, on average, 37 samples with beneficial noise exhibit clearer reasoning compared to 31 without (***H1***), while 26 samples with beneficial noise demonstrate better output formats versus 23 without (***H2***).

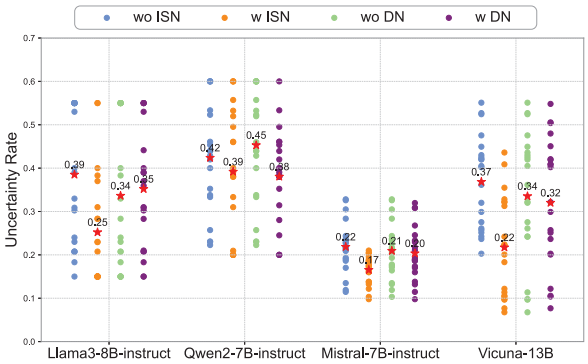

Figure 6: Impact of beneficial noise on LLM output uncertainty (anti-confidence). ISN and DN represent Illegal Sentence Noise and Datatype Noise, with ⋆ indicating mean uncertainty rate (μ).

Second, as shown in Figure 6, we analyze the impact of beneficial noise on LLM output uncertainty across four powerful LLMs. Results indicate that when combined with beneficial noise (ISN or DN), LLMs generally exhibit lower uncertainty and increased confidence in their outputs. This suggests that LLMs pay more attention to provided golden context and respond with greater confidence (***H3***).

7 Conclusion

We define and categorize seven types of RAG noise into beneficial and harmful groups, exploring retrieval noise from linguistic and practical perspectives. To conduct this evaluation, we propose a systematic framework for generating various retrieval documents and establish a novel noise benchmark, NoiserBench. Our experiments reveal that beneficial noise can significantly enhance model performance through clearer reasoning paths, standardized answers, and increased confidence—acting much like Aladdin's Lamp. These findings may offer insights for leveraging beneficial noise mechanisms in future research.

Limitations

While our systematic analysis of RAG noises in real-world scenarios offers valuable insights, several limitations warrant consideration. First, our analysis of noise phenomena remains relatively preliminary. Future work will examine the underlying mechanisms by investigating parameter variations, particularly attention values, across each model layer. Additionally, the benchmark dataset constructed for this study comprises only a few thousand examples, which may be insufficient for comprehensive RAG methodology research. We plan to expand both the dataset diversity and scale to better serve the research community.

References

Aliaa Aloufi. 2021. Language and linguistic orthography. *English Language and Literature Studies*, 11(3).

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

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *Preprint*, arXiv:2303.12712.

Wallace L Chafe. 1971. Linguistics and human knowledge. *Monograph series on languages and linguistics*, (24):57.

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

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. https://lmsys.org/blog/2023-03-30-vicuna/.

Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, et al. 2024. The power of noise: Redefining retrieval for rag systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 719–729, New York, NY, USA. Association for Computing Machinery.

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.

Gary Feng and Li Yi. 2006. What if chinese had linguistic markers for counterfactual conditionals? language and thought revisited. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 28.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey. *Preprint*, arXiv:2312.10997.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.

Yves Grandvalet, Stéphane Canu, and Stéphane Boucheron. 1997. Noise injection: Theoretical prospects. *Neural Computation*, 9(5):1093–1108.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, et al. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *Preprint*, arXiv:2311.05232.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *Preprint*, arXiv:2112.09118.

Zhen Jia, Abdalghani Abujabal, Rishiraj Saha Roy, Jannik Strötgen, and Gerhard Weikum. 2018. Tempquestions: A benchmark for temporal question answering. In *Companion Proceedings of the The Web Conference 2018*, pages 1057–1062, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, et al. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, et al. 2024. Mixtral of experts. *Preprint*, arXiv:2401.04088.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *International Conference on Machine Learning*, pages 15696–15707. PMLR.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen tau Yih. 2020. Dense passage retrieval for open-domain question answering. *Preprint*, arXiv:2004.04906.

András Kertész and Csilla Rákosi. 2012. *Data and evidence in linguistics: A plausible argumentation model*. Cambridge University Press.

Samuel Kotz and Norman L. Johnson, editors. 1992. *Breakthroughs in Statistics: Methodology and Distribution*. Springer New York, New York, NY.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, et al. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *Preprint*, arXiv:1910.13461.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Meta, AI. 2024. Introducing meta llama 3: The most capable openly available llm to date. https://ai.meta.com/blog/meta-llama-3/.

Femi Olan, Uchitha Jayawickrama, Emmanuel Ogiemwonyi Arakpogun, Jana Suklan, and Shaofeng Liu. 2024. Fake news on social media: the impact on society. *Information Systems Frontiers*, 26(2):443–458.

OpenAI. 2023. Introducing chatgpt. https://openai.com/index/chatgpt/.

Preligens Lab. 2023. Textnoisr: Adding random noise to a dataset. https://github.com/preligens-lab/textnoisr.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A Smith, and Mike Lewis. 2022. Measuring and narrowing the compositionality gap in language models. *arXiv preprint arXiv:2210.03350*.

Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, et al. 2021. RocketQA: An optimized training approach to dense passage retrieval for open-domain question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5835–5847, Online. Association for Computational Linguistics.

Ishith Seth, Bryan Lim, Yi Xie, Jevan Cevik, Warren M Rozen, Richard J Ross, and Mathew Lee. 2023. Comparing the efficacy of large language models ChatGPT, BARD, and Bing AI in providing information on rhinoplasty: an observational study. *Aesthetic Surgery Journal Open Forum*, 5:ojad084.

C Shannon, Warren Weaver, and Ch Hockett. 1961. The mathematical theory of communication. *Urbana: University of Illinois*.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, et al. 2023. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, pages 31210–31227. PMLR.

Walter W Skeat. 1993. *The concise dictionary of English etymology*. Wordsworth Editions.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Robert Tumarkin and Robert F Whitelaw. 2001. News or noise? internet postings and stock prices. *Financial Analysts Journal*, 57(3):41–51.

Dilin Wang, Chengyue Gong, and Qiang Liu. 2019. Improving neural language modeling via adversarial training. In *International Conference on Machine Learning*, pages 6555–6565. PMLR.

Wenhui Wang, Hangbo Bao, Shaohan Huang, Li Dong, and Furu Wei. 2021. Minilmv2: Multi-head self-attention relation distillation for compressing pretrained transformers. *Preprint*, arXiv:2012.15828.

Yile Wang, Peng Li, Maosong Sun, and Yang Liu. 2023. Self-knowledge guided retrieval augmentation for large language models. *Preprint*, arXiv:2310.05002.

Zihao Wang, Anji Liu, Haowei Lin, Jiaqi Li, Xiaojian Ma, and Yitao Liang. 2024. Rat: Retrieval augmented thoughts elicit context-aware reasoning in long-horizon generation. *Preprint*, arXiv:2403.05313.

Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. Constructing datasets for multi-hop reading comprehension across documents. *Transactions of the Association for Computational Linguistics*, 6:287–302.

Jinyang Wu, Feihu Che, Xinxin Zheng, Shuai Zhang, Ruihan Jin, Shuai Nie, Pengpeng Shao, and Jianhua Tao. 2024. Can large language models understand uncommon meanings of common words? *Preprint*, arXiv:2405.05741.

Jinyang Wu, Zhiwei Ning, Yidong Ding, Ying Wang, Qinke Peng, and Laiyi Fu. 2023. Kgetcda: an efficient representation learning framework based on knowledge graph encoder from transformer for predicting circrna-disease associations. *Briefings in Bioinformatics*, 24(5):bbad292.

Chong Xiang, Tong Wu, Zexuan Zhong, David Wagner, Danqi Chen, and Prateek Mittal. 2024. Certifiably robust rag against retrieval corruption. *Preprint*, arXiv:2405.15556.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2024. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. *Preprint*, arXiv:2305.13300.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, et al. 2023. Baichuan 2: Open large-scale language models. *Preprint*, arXiv:2309.10305.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, et al. 2024. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

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

Junjie Ye, Nuo Xu, Yikun Wang, Jie Zhou, Qi Zhang, Tao Gui, and Xuanjing Huang. 2024. Llm-da: Data augmentation via large language models for few-shot named entity recognition. *Preprint*, arXiv:2402.14568.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. 2024. Making retrieval-augmented language models robust to irrelevant context. In *The Twelfth International Conference on Learning Representations*.

Penghao Zhao, Hailin Zhang, Qinhan Yu, Zhengren Wang, Yunteng Geng, Fangcheng Fu, Ling Yang, Wentao Zhang, Jie Jiang, and Bin Cui. 2024. Retrieval-augmented generation for ai-generated content: A survey. *Preprint*, arXiv:2402.19473.

Xinxin Zheng, Feihu Che, Jinyang Wu, Shuai Zhang, Shuai Nie, Kang Liu, and Jianhua Tao. 2024. Ks-llm: Knowledge selection of large language models with evidence document for question answering. *Preprint*, arXiv:2404.15660.

Kaijie Zhu, Jindong Wang, Jiaheng Zhou, Zichen Wang, Hao Chen, Yidong Wang, Linyi Yang, Wei Ye, et al. 2024. Promptrobust: Towards evaluating the robustness of large language models on adversarial prompts. *Preprint*, arXiv:2306.04528.

Appendix

Within this supplementary material, we elaborate on the following aspects:

- Appendix A: Models
- Appendix B: Implementation Details
- Appendix C: Results

A Models

We provide brief introductions to LLMs used in our experiments. For more details, please refer to the official websites or the corresponding Hugging Face Transformers repository.

- **Llama2** & **Llama3**: The Llama series model, developed by Meta AI's FAIR team, is a widely-used autoregressive language model. These models, particularly Llama3, achieve competitive performance compared to some state-of-the-art closed-source LLMs. We use the 13B model for Llama2, and the 8B and 70B models for Llama3.

- **Vicuna-v1.5**: The Vicuna model, derived from fine-tuning the LLaMA-2 base model by LMSYS, was developed using around 70K user-shared conversations obtained from ShareGPT.com through public APIs. We use the popular vicuna-13B here.

- **Qwen2**: Proposed by Alibaba Cloud, Qwen series are strong language models, which have been stably pretrained for up to 3 trillion tokens of multilingual data with a wide coverage of domains, languages (with a focus on Chinese and English), etc. Qwen2-7B-Instruct is utilized.

- **Mistral**: The Mistral series includes the Mistral-7B and Mixtral-8x7B models. The Mistral-7B is an autoregressive language model with 7 billion parameters, trained on a diverse corpus to ensure high performance in various tasks. The Mixtral-8x7B is a high-quality sparse mixture of expert models (SMoE) with open weights. This technique increases the number of parameters of a model while controlling cost and latency, as the model only uses a fraction of the total set of parameters per token.

- **Baichuan2**: Baichuan2 is the new generation of open-source language models launched by Baichuan Intelligence. It is trained on a high-quality corpus with 2.6 trillion tokens and has achieved the best performance in authoritative Chinese and English benchmarks of the same size. We use the 13B chat model.

Prompt 1: Counterfactual Evidence Generation

Instruction:
Given a question and its answer, please write a short piece of evidence within 50 words to support it. You can make up fake content and supporting evidence but it should be as realistic as possible. Ignore the correctness of the given answer.

Examples:
Question: What is the capital of France?
Answer: Lyon
Evidence: Lyon is the capital of France. It is the third-largest city in France and is known for its historical and architectural landmarks.

Question: where does aarp fall on the political spectrum?
Answer: Conservative-leaning
Evidence: AARP, the American Association of Retired Persons, has often been perceived as conservative-leaning due to its advocacy for policies that emphasize fiscal responsibility and traditional values.

Question: Who is the chief scientist of Google DeepMind?
Answer: Demis Hassabis
Evidence: Demis Hassabis is a British artificial intelligence researcher, neuroscientist, and entrepreneur. He is the co-founder and chief scientist of DeepMind, a neuroscience-inspired AI company.

Outputs:
Question: <your created question>
Answer: <corresponding answer to your question>
Evidence:

Figure 7: Example LLMs' input for counterfactual evidence generation. This prompt is composed of instruction, examples, and candidate counterfactual QA.

Table 5: Main libraries and the corresponding versions.

Package	Version
vllm	0.2.6
torch	2.1.2+cuda12.4
transformers	4.36.2

B Implementation Details

B.1 Compute Infrastructure

We execute the experiments using the following compute specifications.

- NVIDIA A100 80 GB GPU $\times$ 2
- 256 GB RAM

We use Python 3.10.0 and speed up inference using vllm[1], a fast and easy-to-use library. In Table 5, we list the main libraries along with their versions.

B.2 Dataset Construction

To construct our benchmark NoiserBench, we need to first gather candidate QA instances from multiple sources. In this paper, our source data is obtained from seven publicly available datasets, including single-hop NQ and RGB, explicit multi-hop HotpotQA, 2WikiMQA, Bamboogle, and implicit multi-hop StrategyQA and TempQA. Table 6 shows the full list of candidate instances, and in total, we use 26,855 instances.

Subsequently, we introduce various noisy documents using external tools. For counterfactual noise, we obtain relevant entities related to the golden answer from Google search[2] to construct counterfactual answers. For orthographic noise, we utilize the open-source textnoisr package[3], which enables the convenient introduction of noise to text datasets and precise control of the quality of results. Four types of "action" are implemented: insert, delete, substitute, and swap. For other types of noise, we utilize the 2018 English Wikipedia dump for document construction. We present the prompts in Figure 7-9.

[1] https://github.com/vllm-project/vllm

Prompt 2: Supportive Evidence Generation

Instruction:
Given a question and its answer, please write a short piece of evidence within 50 words to support it.

Examples:
Question: What is the capital of France?
Answer: Paris
Evidence: Paris is the capital of France. It is known for its iconic landmarks such as the Eiffel Tower and the Louvre Museum.

Question: who was governor of oregon when shanghai noon was released?
Answer: John Kitzhaber
Evidence: John Kitzhaber was serving as the governor of Oregon when 'Shanghai Noon' was released in 2000. He was in office from 1995 to 2003, focusing on healthcare reform and environmental issues.

Question: Who is the chief scientist of Google DeepMind?
Answer: Jeff Dean
Evidence: Jeff Dean is a computer scientist and Google Senior Fellow in the Research Group, where he leads the Google Brain project.

Outputs:
Question: <your created question>
Answer: <corresponding answer to your question>
Evidence:

Figure 8: Example LLMs' input for supportive evidence generation. This prompt is composed of instruction, examples, and candidate QA.

Prompt 3: Datatype Documents Construction

Instruction:
Given a question, gold answer, and gold evidence, insert code snippets (Python, C or other languages) or links as strings within the gold evidence to introduce noise and confusion, ensuring the gold answer portion remains uninterrupted.

Example:
Question: who directed the iconic animated short educational film ek anek aur ekta?
Answer: Vijaya Mulay
Evidence: Ek Anek Aur Ekta Directed by Vijaya Mulay Produced by Films Division of India Screenplay by Vijaya Mulay Music by Vasant Desai Edited by Waman B. Bhosle and Guru Dutt Shirali (Waman - Guru) Production company National Centre for Education Technology (NCERT) Release date Running time 7 : 07 Country India Language Hindi
Noisy Documents: \n<h1>Indian Films</h1>\nEk Anek Aur Ekta Directed by Vijaya Mulay Produced by Films Division of India Screenplay by Vijaya Mulay Music by Vasant Desai Edited by Waman B. Bhosle and Guru Dutt Shirali (Waman - Guru) (https://google.com/) Production company National Centre (https://Wikipedia.center) for Education Technology (NCERT) Release date Running time 7 : 07 Country India Language Hindi

Outputs:
Question: <your created question>
Answer: <corresponding answer to your question>
Evidence: <Supportive Evidence>
Noisy Documents:

Figure 9: Example LLMs' input for datatype noise construction. This prompt is composed of instruction, examples, candidate QA and corresponding evidence.

[2] We query Google search via SerpAPI: https://serpapi.com

[3] https://github.com/preligens-lab/textnoisr

Table 6: Statistics of source QA instances from a couple of knowledge-intensive datasets. 'E' and 'I' represent explicit and implicit, respectively.

Dataset	Category	Source	#Source pairs	#Samples	Example
NQ	Single-hop	Train set	2,889	500	Who won the 7 man elimination chamber match?
RGB	Single-hop	Test set	300	300	How many vehicles did Tesla deliver in 2021?
HotpotQA	Multi-hop (E)	Dev set	7,405	500	What election will take place on the same day as the United States Senate election in Texas?
2WikiMQA	Multi-hop (E)	Dev set	12,576	500	Where was the place of death of Isabella of Bourbon's father?
Bamboogle	Multi-hop (E)	All	125	125	Who was the first African American mayor of the most populous city in the United States?
StrategyQA	Multi-hop (I)	Train set	2,290	500	Can Arnold Schwarzenegger deadlift an adult Black rhinoceros?
TempQA	Multi-hop (I)	All	1,270	186	Who was the commander-in-chief of the colonial army during the revolutionary war?
PriorQA	Mix-hop	All	500	500	What were the primary strategies employed by the British army duringthe American Civil War?

B.3 Implementation Details

We utilize a CN for retrieval, where relevant Chinese documents are retrieved in response to the input query to enhance the prompt, rather than relying on a fixed set of N Chinese examples. This CN corpus is constructed using data from Wikidata and Google Search, intentionally designed to include misleading or outdated information in order to simulate real-world scenarios where context might be inaccurate or evolving.

During the experiment, we found that the optimal Top-k value for the contriver was 5, and the similarity threshold for all-MiniLM-L6-v2 was set to 0.3.

C Results

In this section, we provide supplementary results to further illustrate the role of RAG noise, especially beneficial noise. Our analysis primarily focuses on datatype noise, orthographic noise, and prior noise, as illegal sentence noise has been extensively discussed in the main text, and other forms of noise have been explored in previous studies. These additional results aim to provide a more comprehensive understanding of various noise types and their effects on the model's performance.

C.1 Results on Prior Noise

Table 7 presents results for RAG models affected by prior noise using our dataset, PriorQA. Questions in this dataset contain factual errors, such as "Which country hosted the 1980 FIFA World Cup?" (Actually, 1980 FIFA World Cup was not held). Accuracy is assessed by whether models correctly identify and respond with "The question is factually incorrect". We observe that all models except Llama2-13B perform well with direct prompts and benefit from retrieving background information due to extensive pre-training knowledge. However, models like Llama2-13B, which persist in searching based on incorrect priors, may retrieve false information and exhibit diminished performance. This underscores the need to detect prior errors in user queries before answering in future RAG system designs.

Table 7: The effects of prior noise on LLMs, which is measured by accuracy (%). 'Base' indicates the scenario with no retrieval. 'Misleading' refers to counterfactual content associated with prior noise. 'Background' denotes multiple retrieval results obtained after decomposing the query into its constituent entities.

Models	Base	Misleading	Background
Llama3-8B	93.40	47.80	90.00
Qwen2-7B	94.20	28.20	98.20
Mistral-7B	96.60	28.60	99.20
Llama2-13B	21.00	5.60	61.60
Vicuna-13B	91.00	25.80	99.20
Baichuan2-13B	90.00	45.20	96.40
Llama3-70B	99.00	78.40	99.80
Mixtral-8x7B	91.20	39.00	99.60
Average	**79.93**	**34.20**	**88.47**

C.2 Results Across Eight Models

As shown in Figure 10, we first present the average performance over seven datasets for datatype

Table 8: Impact of various noise types on accuracy (%) for eight representative LLMs on the RGB dataset. We assess performance across various retrieval scenarios: "Base" (no retrieval), "Golden Only" (only golden retrieval context), and "Golden & XXX" (golden context + specific retrieval noises, including Counterfactual, Supportive, Orthographic, Semantic, Datatype, Illegal Sentence Noise).

Scenario	**Small**						**Large**		**Average**
	L2-13B	L3-8B	Q2-7B	M-7B	V-13B	B2-13B	L3-70B	M-8x7B	
Base	17.00	47.00	31.33	27.00	35.33	27.67	60.00	43.00	36.04
Golden Only	29.33	80.00	76.33	80.33	80.33	78.00	76.00	77.33	72.20
Golden & CN	14.00	36.33	26.00	19.33	19.33	15.00	42.33	31.00	25.42
Golden & SuN	26.00	80.00	74.00	72.33	61.67	65.33	73.67	76.67	66.21
Golden & ON	14.33	75.00	74.00	72.67	77.67	69.00	77.00	72.67	66.54
Golden & SeN	18.00	81.33	77.67	56.67	52.00	59.00	76.33	77.33	62.30
Golden & DN	40.00	81.67	84.33	85.33	85.67	81.67	85.00	81.00	78.08
Golden & ISN	72.33	83.00	80.00	81.00	82.33	79.67	79.67	79.00	79.63

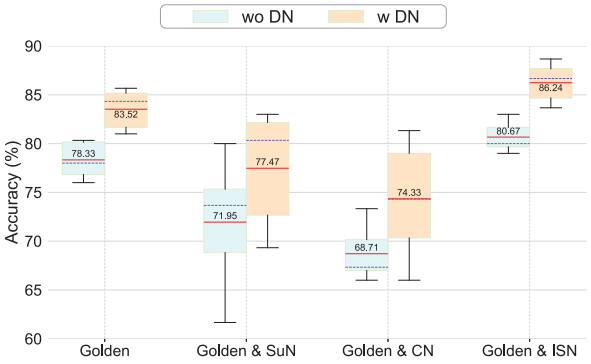

Figure 10: Impact of datatype noise (DN) on the average accuracy of eight representative LLMs on RGB. 'Golden', 'SuN', 'CN', and 'ISN' represent golden context only, golden context with supportive, counterfactual, and illegal sentence noise, respectively. The mean is marked by a red solid line and the median by a purple dashed line.

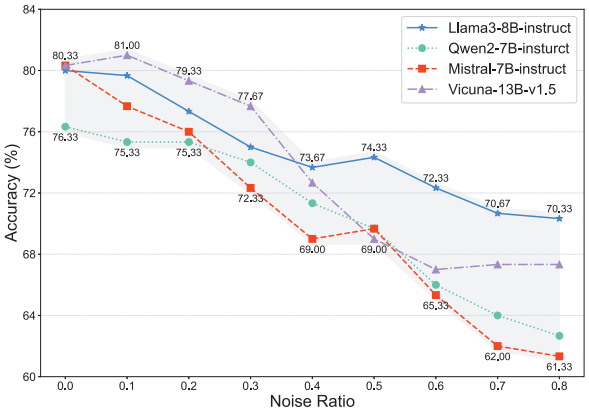

Figure 11: The experimental results of noise robustness measured by accuracy (%), under different orthographic noise ratios. Performance is benchmarked across state-of-the-art open-source models, such as Llama3-8B-instruct, for noise ratios ranging from 0 to 0.8. The maximum and minimum accuracy for all models at each noise ratio is annotated, with a shaded region representing ±0.5% threshold to illustrate the overall trend in model performance better as the noise ratio increases.

noise to demonstrate that beneficial noise improves performance across various LLMs with different model architectures and scales. We apply proportional scaling to counterfactual data to make a clearer illustration within one figure while maintaining consistent conclusions. The results indicate that datatype noise significantly enhances model performance in all scenarios, with the most substantial improvement under harmful noise.

Additionally, we provide detailed results for eight models on the RGB dataset, which is based on recent news corpora and thus better reflects the impact of noise. As shown in Table 8, all models exhibit improved performance with the addition of beneficial noises such as illegal sentence noise and datatype noise. The effect of semantic noise varies among models, with performance gains observed in some cases. In contrast, counterfactual, supportive, and orthographic noises consistently degrade model performance, highlighting their harmful nature. Therefore, we can confidently classify illegal sentence noise and datatype noise as beneficial, and counterfactual, supportive, and orthographic noises as harmful. The classification of semantic noise as beneficial or harmful depends on the models, as it can be beneficial in some cases and harmful in others. In this paper, we provisionally categorize it as beneficial due to its generally positive impact (4 models) and its alignment with conclusions proposed in prior research.

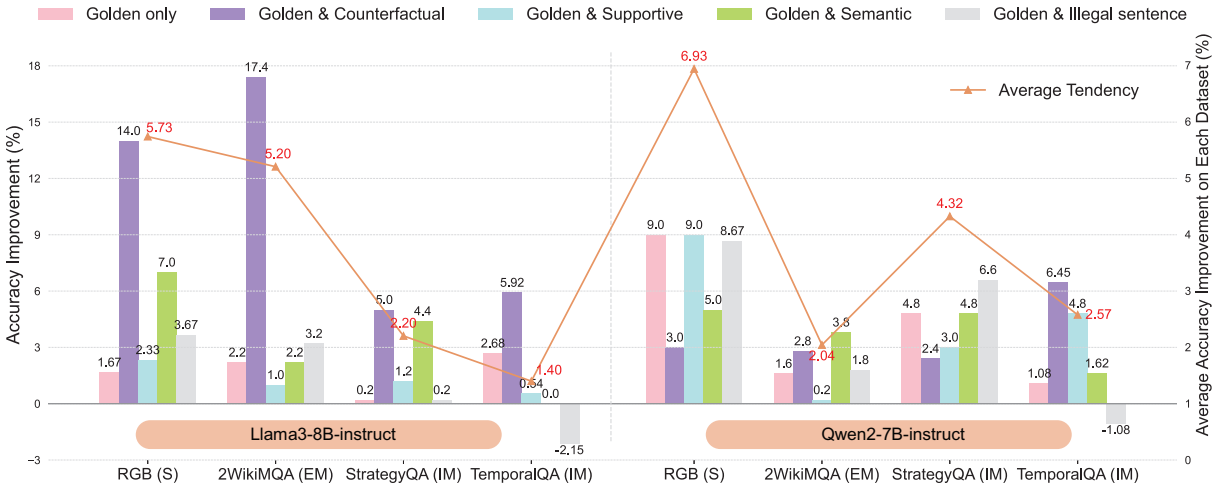

Figure 12: Results for the impact of datatype noise on the Llama3-8B-instruct and Qwen2-7B-instruct models when exposed to five typical noise categories across four datasets, including both single-hop (S) and multi-hop (explicit: EM, implicit: IM) reasoning tasks. The bar charts show performance differences upon introducing datatype noise. The line graphs illustrate the average accuracy improvement across noise types per dataset.

Table 9: Additional results for different illegal sentence noise (ISN) ratios on RGB. L2-13B, L3-8B, Q2-7B, M-7B, V-13B, B2-13B, L3-70B and M-8x7B represents Llama2-13B, Llama3-8B-Instruct, Qwen2-Instruct, Mistral-7B-Instruct-v0.2, Vicuna-13B-v1.5, Baichuan2-13B-chat, Llama3-70B-Instruct, Mixtral-8x7B-Instruct.

Scenario	Small						Large		Average
	L2-13B	L3-8B	Q2-7B	M-7B	V-13B	B2-13B	L3-70B	M-8x7B	
0	29.33	80.00	76.33	80.33	80.33	78.00	76.00	77.33	72.21
+ ISN	72.33	83.00	80.00	81.00	82.33	79.67	79.67	79.67	**79.71** +7.50
0.2	18.67	77.33	75.33	76.00	79.33	73.33	76.67	73.67	68.79
+ ISN	73.67	82.67	80.33	76.67	80.00	72.33	80.33	73.67	**77.46** +8.67
0.4	12.33	73.67	71.33	69.00	72.67	68.00	76.33	65.67	63.63
+ ISN	70.67	77.00	73.00	71.00	73.33	68.33	80.00	66.67	**72.50** +8.87
0.6	8.67	72.33	66.00	65.33	67.00	63.67	82.00	64.33	61.17
+ ISN	69.33	72.00	66.67	64.67	70.00	66.33	79.33	63.67	**69.00** +7.33
0.8	8.00	70.33	62.67	61.33	68.67	63.67	78.00	62.33	59.38
+ ISN	68.33	70.67	64.67	63.33	69.00	66.33	78.33	63.67	**68.04** +8.66

C.3 Results Across the Number of Quires

We leverage keywords extracted from queries (ranging from 1 to 4 keywords per query) for content retrieval. Experimental results on the 2WikiMQA dataset demonstrate that the core findings regarding the impact of beneficial and harmful noise remain consistent across queries of varying complexity.

Table 10: Accuracy in different scenarios and with different numbers of queries.

Scenario	Number of queries			
	1	2	3	4
G only	79.80	84.20	84.60	84.60
G&CN	21.20	23.60	23.60	23.80
G&ISN	87.40	89.40	89.20	89.40

C.4 Performance Under Other Noise Disturbances

To illustrate the impact of beneficial noise under other noise disturbances, we analyze the effect of datatype noise (DN) in five scenarios: no noise (i.e., Golden only), harmful noise (i.e., Golden & counterfactual noise, Golden & supportive noise), and beneficial noise (i.e., Golden & illegal sentence noise, Golden & Semantic noise). Figure 12 shows the model's accuracy gains after introducing DN

Table 11: Additional control experiments by varying the repetition of answer-containing text chunks.

	512	748	1024	2048	Avg
Golden only	79.80	83.00	83.20	83.40	82.35
Golden & CN	21.20	23.60	23.80	23.80	23.10
Golden & SuN	81.00	82.00	82.60	82.40	82.00
Golden & ON	78.60	79.80	80.00	80.00	79.60
Golden & SeN	87.20	89.20	89.00	89.20	$88.65_{+6.30}$
Golden & DN	82.00	85.60	85.80	86.20	$84.90_{+2.55}$
Golden & ISN	87.40	89.80	90.00	89.80	$89.25_{+6.90}$

in these scenarios. We find that DN generally enhances performance across all datasets, particularly when combined with harmful noise like counterfactual noise, with average accuracy improvements exceeding 10 percentage points. This consistent enhancement underscores beneficial noise's potential significance for future RAG research.

C.5 Noise Robustness of RAG Models under Different Noise Ratios

We provide the results of four representative LLMs under different orthographic noise ratios. Specifically, for insert, delete, and substitute actions, the noise ratio ranges from 0.0 to 0.9, while for swapping, it ranges from 0.0 to a maximum of 0.5. As shown in Figure 11, the maximum and minimum accuracy for all models at each noise ratio is annotated, with a shaded region representing ±0.5% threshold to better illustrate the overall trend in model performance as the noise ratio increases. We observe that increasing noise rates pose a challenge for RAG in LLMs, particularly when the ratio exceeds 0.3. Therefore, we use a default ratio of 0.3 in our main results to objectively assess the impact of harmful noise.

C.6 The effects of Beneficial Noise under Different Noise Ratios

To demonstrate the positive effects at different harmful noise ratios, we present comprehensive results for illegal sentence noise disturbances with ratios ranging from 0 to 0.8. As shown in Table 9, we see that the introduction of illegal sentence noise (beneficial noise) consistently enhances model performance, thereby illustrating the generalization of beneficial noise.

C.7 Additional Control Experiments

We conducted additional control experiments by varying the repetition of answer-containing text chunks. Results on the 2WikiMQA dataset using Llama3-8B are presented. While these factors did have some impact on performance, our core finding—that RAG noise can be categorized into beneficial and harmful types—remained consistent.

Table 12: Evaluation results (accuracy (%) for mathematical reasoning using GPT-3.5-turbo as the base model. The four conditions are zero-shot without noise, two-shot without noise, and perturbations to numeric and operator elements in 2-shot examples. △ denotes the accuracy improvement (%) with noise compared to no noise.

Scenario	GSM8K	GSMHard	Average (△)
0-shot	50.40	40.20	45.30
2-shot-no-noise	55.40	47.80	51.60
2-shot-num	65.40	50.60	58.00 (+6.40)
2-shot-operator	62.20	53.20	57.70 (+6.10)

C.8 The Impact of RAG Noise on Other Tasks like Mathematical Reasoning

Given that previous discussions focused on QA tasks, it remains unclear whether the beneficial noise affects other tasks. To address this, we conduct experiments on mathematical reasoning, which requires higher cognitive and reasoning abilities. Following prior research, we apply the PAL methodology to evaluate mathematical reasoning results. This approach involves using LLMs to parse natural language problems, generate intermediary programmatic solutions, and execute these solutions via a Python interpreter.

As shown in Table 12, introducing numeric or operator perturbations to retrieved examples significantly improves model performance (by 6.40% and 6.10%, respectively). We hypothesize that this mechanism resembles adversarial training (Wang et al., 2019). Specifically, these perturbations likely help the model implicitly learn to identify and address potential errors or ambiguities, thereby enhancing its robustness. As a result, LLMs are better equipped to reason accurately amidst unclear or noisy test examples due to this implicit training.

C.9 Case Study

Table 13-16 presents the reasoning process of Llama3-8B-instruct. Without beneficial noise, the model ignores correct information and exhibits logical flaws under counterfactual noise influence. However, upon introducing beneficial noise, the model exhibits heightened attention to the golden context and successfully distinguishes between correct and incorrect information

Table 13: Example #1 of LLM outputs on the single-hop dataset without and with beneficial noise (BN). The **blue** and **green** colors denote the correct and incorrect responses, respectively. 'GT' denotes ground truth. Upon introducing BN, LLMs exhibit clearer reasoning processes, more standardized response formats, and increased focus on golden context.

Retrieved Context wo BN	[1] Paul R. Ehrlich, author of "The Population Bomb," was born on June 14, 1931. Ehrlich, a renowned biologist, published his influential book on population control in 1968, which sparked . . . **[Counterfactual]** [2] The author of The Population Bomb, Paul Ehrlich, was born on May 29, 1932. He is an American biologist and educator, known for his warnings about the consequences of overpopulation . . . **[Golden Context]**
Question	When was the author of The Population Bomb born? 'A': June 14, 1931 'B': Uncertain 'C': May 29, 1932 (Ground Truth) 'D': April 22, 1934
Answer wo BN	**A** ###\n\n### Explanation:\n The correct answer is 'A': June 14, 1931. According to the information provided, Paul R. Ehrlich, the author of "The Population Bomb" was born on June 14, 1931. The other options are incorrect, as they provide different birth dates for the author. ✖
Answer w BN	**C**\n### Explanation: According to information [2], Paul Ehrlich, the author of The Population Bomb, was born on May 29, 1932. Therefore, the correct answer is 'C'. Information [1] provides a different birthdate, but it is not the correct answer. The other options are not mentioned in the given information. ✔

Table 14: Example #2 of LLM outputs on the single-hop dataset without and with beneficial noise (BN). The **blue** and **green** colors denote the correct and incorrect responses, respectively. 'GT' denotes ground truth. Upon introducing BN, LLMs exhibit clearer reasoning processes, more standardized response formats, and increased focus on golden context.

Retrieved Context wo BN	[1] Northrop Corporation built the fastest air-breathing manned aircraft, the Northrop Grumman B-2 Spirit. Known for its stealth technology and high-speed capabilities, it represents a pinnacle in modern aviation technology. **[Counterfactual]** [2] The Lockheed Corporation built the fastest air-breathing manned aircraft, the Lockheed SR-71 Blackbird. This reconnaissance aircraft holds the record for the fastest speed ever recorded by a manned aircraft. **[Golden Context]** . . .
Question	Who built the fastest air-breathing manned aircraft? 'A': Uncertain 'B': Boeing Corporation 'C': Lockheed Corporation (GT) 'D': Northrop Corporation
Answer wo BN	**D**\n \n### Explanation:\n The information provided states that Northrop Corporation built the Northrop Grumman B-2 Spirit, which is known for its stealth technology and high-speed capabilities. However, it does not explicitly state that this aircraft is the fastest air-breathing manned aircraft. ✖
Answer w BN	**C**\n### Explanation: The information provided states that the Lockheed Corporation built the Lockheed SR-71 Blackbird, which is the fastest air-breathing manned aircraft. Therefore, the answer is 'C'. ✔

Table 15: Example #3 of LLM outputs on the multi-hop dataset without and with beneficial noise (BN). The **blue** and **green** colors denote the correct and incorrect responses, respectively. Upon introducing BN, LLMs exhibit clearer reasoning processes, more standardized response formats, and increased focus on golden context.

Retrieved Context wo BN	[1] Robert Harley, Earl of Oxford and Earl Mortimer, was the paternal grandfather of Edward Harley, 5th Earl of Oxford and Earl Mortimer. He was a prominent statesman and served as the Lord High Treasurer of Great Britain. **[Counterfactual]** [2] John Harley (29 September 1728 – 7 January 1788) was a British bishop. Harley was the second son of Edward Harley, 3rd Earl of Oxford and Earl Mortimer. He was Archdeacon of Shropshire from 1760 to 1769 and then Archdeacon of Hereford from 1769 to 1787. **[Golden Context]** ...
Question	Who is the paternal grandfather of Edward Harley, 5Th Earl Of Oxford And Earl Mortimer? 'A': Edward Harley, 3rd Earl of Oxford and Earl Mortimer (GT) 'B': Uncertain 'C': Uncertain 'D': Thomas Harley, 4th Earl of Oxford and Earl Mortimer
Answer wo BN	**C** \n\n The paternal grandfather of Edward Harley, 5th Earl of Oxford and Earl Mortimer, is Robert Harley, Earl of Oxford and Earl Mortimer. Therefore, the correct answer is 'C'. ✖
Answer w BN	**A**\n\n### Explanation: nThe information provided states that John Harley (1728-1788) was the second son of Edward Harley, 3rd Earl of Oxford and Earl Mortimer. Therefore, Edward Harley, 3rd Earl of Oxford and Earl Mortimer is the paternal grandfather of Edward Harley, 5th Earl of Oxford and Earl Mortimer. ✔

Table 16: Example #4 of LLM outputs on the multi-hop dataset without and with beneficial noise (BN). The **blue** and **green** colors denote the correct and incorrect responses, respectively. Upon introducing BN, LLMs exhibit clearer reasoning processes, more standardized response formats, and increased focus on golden context.

Retrieved Context wo BN	[1] With an estimated base length of 230 meters, 200 men end to end would only cover about 46% of the Great Pyramid of Giza's base, highlighting the massive scale and size of the ancient structure. **[Counterfactual]** [2] The base of the Great Pyramid of Giza is 756 feet long. The average height of a man is 5 foot 9. **[Golden Context]** ...
Question	Can 200 men end to end cover Great Pyramid of Giza's base? 'A': Uncertain 'B': False (GT) 'C': True
Answer wo BN	**C** ✖
Answer w BN	**B**\n\n### Information:\n The base of the Great Pyramid of Giza is 756 feet long. The average height of a man is 5 foot 9. (int) The number of men required to cover the base end to end is 756 / 5.75. ✔