

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

Jinyang Wu<sup>1</sup>, Feihu Che<sup>1</sup>, Chuyuan Zhang<sup>1</sup>, Jianhua Tao<sup>1,2,\*</sup>, Shuai Zhang<sup>1</sup>, Pengpeng Shao<sup>1</sup>

<sup>1</sup>Department of Automation <sup>2</sup>Beijing National Research Center for Information Science and Technology  
Tsinghua University, Beijing, China

{wu-jy23,cyzhang24}@mails.tinghua.edu.cn, {qkr, zhang\_shuai, ppshao, jhtao}@tsinghua.edu.cn

## 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 more robust, adaptable RAG solutions and mitigating hallucinations across diverse retrieval scenarios.

## Introduction

Large language models (LLMs) (OpenAI 2023; Meta, AI 2024) have demonstrated remarkable proficiency across various tasks (Bubeck et al. 2023). Despite these impressive capabilities, LLMs face challenges such as reliance on outdated knowledge and hallucination (Huang et al. 2023; Kandpal et al. 2023).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 (Gao et al. 2024). RAG enhances LLM performance by augmenting inputs with additional information retrieved from external sources during inference.

However, the internet is filled with various non-standard noises, including AI-generated fake news, outdated content, spelling errors, and data contamination, which may potentially influence model performance (Shi et al. 2023; Xie et al. 2024).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.

Recently, several studies (Chen et al. 2024; Xiang et al. 2024) have attempted to extend RAG systems to complex

**### Question:**  
what team did david beckham play for before la galaxy?

A: Manchester United      B: AC Milan  
C: Uncertain      **D: Real Madrid (Ground Truth)**

---

**### Retrieved texts:**  
[1] Before joining LA Galaxy, David Beckham was a prominent player for **Manchester United**. He played at Old Trafford from 1992 until 2003 ... (Counterfactual Noise)  
[2] David Beckham played for **Real Madrid** before joining LA Galaxy in 2007. During his tenure at Real Madrid from 2003 to 2007, Beckham was ... (Golden Context)

---

**Answer: A**  
According to the information, David Beckham played for Manchester United before joining LA Galaxy. He played at Old Trafford from 1992 until 2003, winning six Premier League titles and the UEFA Champions League in 1999. Therefore, the correct answer is 'A'.

**+ Beneficial Noise**

---

**Answer: D**  
According to the given information, David Beckham played for **Real Madrid** before joining LA Galaxy in 2007. Therefore, the correct answer is 'D': Real Madrid. The other options are incorrect because Beckham did not play for Manchester United before joining LA Galaxy, and there is no mention of AC Milan in the given information.

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’.

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 LLM. 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

\*Corresponding author

noise allows the LLM to avoid the harmful effects of counterfactual noise, concentrate on the golden context, and produce accurate responses. Thus, there’s an urgent need to re-define and describe noise 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 increased 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. The main contributions are:

- We define seven types of noise and categorize them into two groups: beneficial and harmful. This is the first comprehensive study to define and assess RAG noises from both linguistic and practical perspectives.
- We propose a novel framework for constructing diverse retrieval documents and create the noise RAG benchmark (NoiserBench), which effectively simulates the impact of real-world noise on RAG models.
- Evaluated on eight datasets and representative 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.

## 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 work primarily focuses 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. RetRobust (Yoran et al. 2024) introduces an

automated data generation method to fine-tune the generator to utilize retrieved passages against noise effectively. RobustRAG (Xiang et al. 2024) proposes a defense framework that enhances RAG model robustness against retrieval corruption attacks through an isolate-then-aggregate strategy, achieving certifiable robustness via secure text aggregation techniques. 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 present a systematic analysis of RAG noise and reveal its role.

### Noise Injection in LLMs

Noise injection (Grandvalet, Canu, and Boucheron 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 LLM’s 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.

## A Taxonomy of RAG Noise

As shown in Figure 2, we categorize RAG noise into seven types from a linguistic perspective. 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 the *Experiments* section.

**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, Weaver, and Hockett 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 types of data: 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), which poses significant 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).

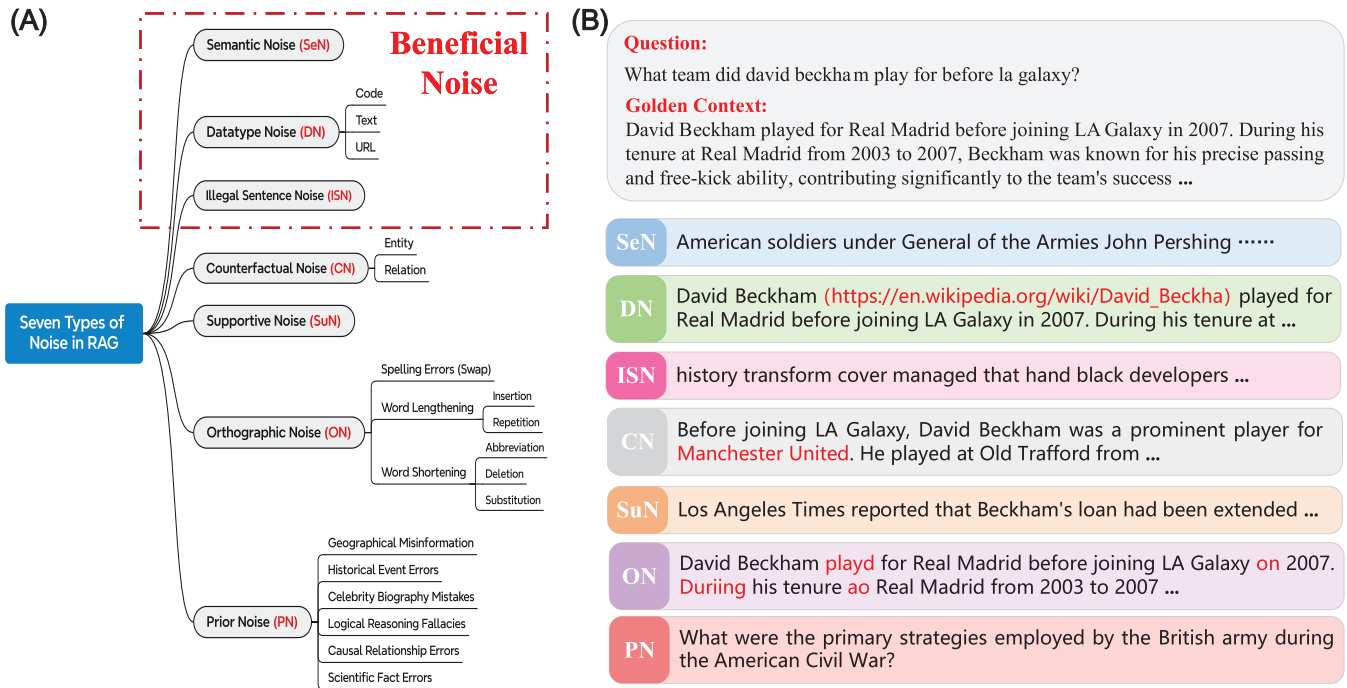


Figure 2: **(A)** Seven types of noise in RAG comprehensively reflect real-world scenarios. **(B)** This detailed illustration of diverse RAG noise intuitively showcases various types. Note that significant noise injections are highlighted in red.

**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.

## Noise RAG Benchmark Construction

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

### 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 six 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; Cucunasu 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 the answer that he played for Real Madrid before joining LA Galaxy. Therefore, we use the natural language inference model bart-large-mnli-407M (Lewis et al. 2019) to ensure evidence properly entails the answer. Note that, we only keep those examples with an entailment probability  $\geq 0.8$ .

**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

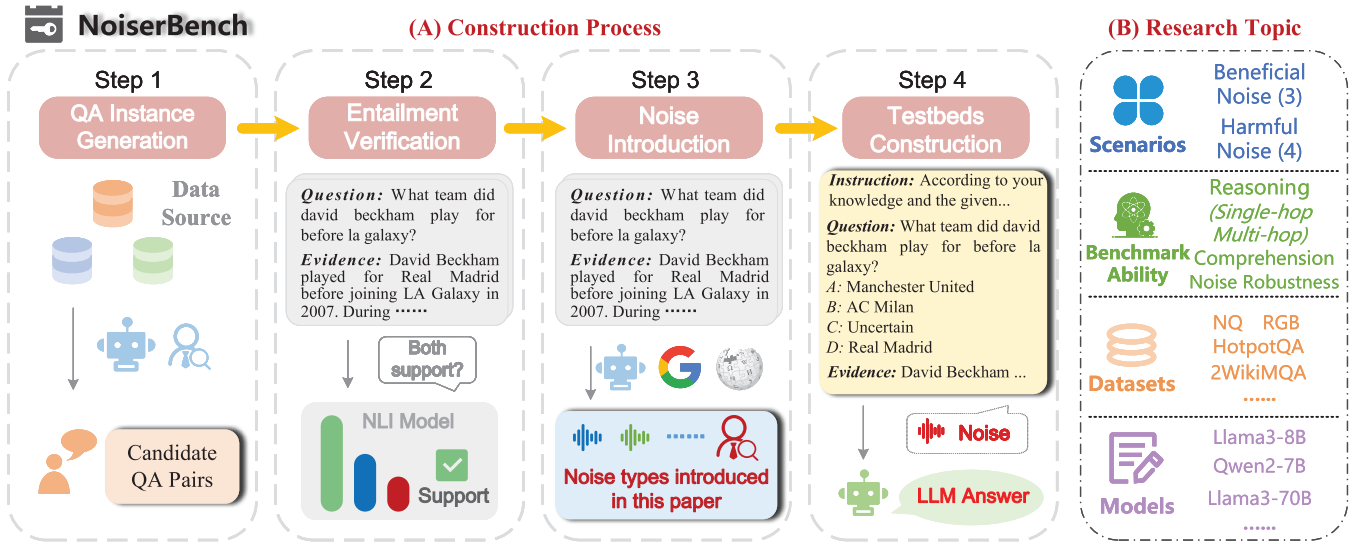


Figure 3: The overall framework for simulating the impact of real-world noise on RAG models. Initially, we generate and obtain QA instances, utilizing ChatGPT to filter out ambiguous examples (*Step 1*). Then, we perform entailment verification using NLI models to maintain evidence quality (*Step 2*). 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.

**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 4: Example LLM input for counterfactual evidence generation. The context of the prompt is composed of instruction, examples, and candidate counterfactual QA.

then employed to construct corresponding supportive evidence, followed by entailment verification. We present the prompts in Figure 4. For Supportive and semantic noise, we utilize the 2018 English Wikipedia dump (Karpukhin et al. 2020) as source documents, with off-the-shelf ContrieverMS 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 textnoir package (Preligens Lab 2023), which enables convenient noise introduction. Four types of “action” are implemented: insert, delete, substitute, and swap. In summary, 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 LLM 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 LLM sensitivity to option order (Wu et al. 2024).

Finally, eight datasets are obtained for NoiserBench. 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 dataset contains fewer than 500.

## Evaluation Metrics

This benchmark aims to reveal the role that RAG noise plays on LLMs. We use accuracy as the primary metric and report the weighted average accuracy across datasets, calculated by aggregating accuracy for each dataset.



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.

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

## Experiments

### Experiment Setup

**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, Stenetorp, and Riedel 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 employ 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.

**Baseline Models** We evaluate 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. Detailed descriptions of each model are provided in the official

websites or the corresponding Huggingface repository<sup>1</sup>.

**Implementation Details** 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<sup>2</sup>, a fast and easy-to-use library.

### Main Results

Firstly, we discuss the role of diverse RAG noises. While prior work has studied the harmful effects of RAG noise, we focus on beneficial noise. Specifically, after revealing the role of noises, we evaluate the effectiveness of beneficial noise across multiple dimensions, including model architectures, scales, and RAG system designs. Then, we investigate whether beneficial noise improves performance amidst other noise types and verify its effectiveness statistically.

**The Role of Diverse RAG Noises** Table 1 illustrates the impact of diverse noise types (the first six) on two state-of-the-art 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

<sup>1</sup><https://huggingface.co/models>

<sup>2</sup><https://github.com/vllm-project/vllm>

Table 2: Effects of prior noise 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

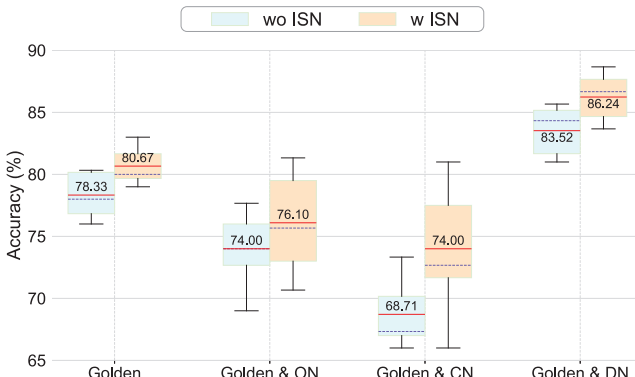


Figure 5: Impact of illegal sentence noise (ISN) on the average accuracy of eight representative LLMs on RGB. ‘Golden’, ‘ON’, ‘CN’, and ‘DN’ represent golden context only, golden context with orthographic, counterfactual, and datatype noise, respectively. The mean is marked by a red solid line and the median by a purple dashed line.

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.

For prior noise, we assess eight LLMs on our dataset, PriorQA. 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

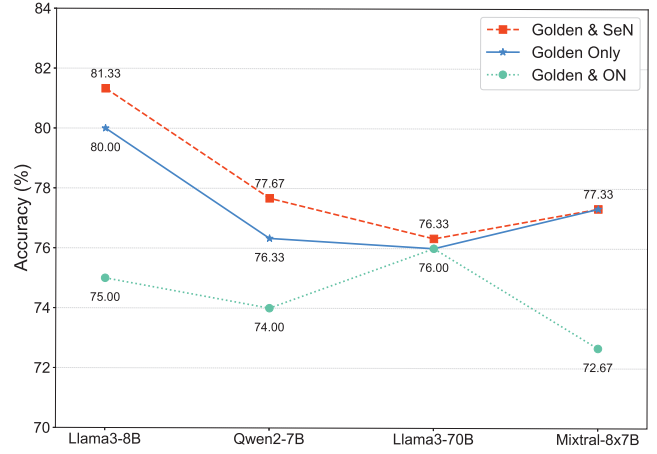


Figure 6: Impact of three noise types on accuracy (%) on RGB. We assess performance across various retrieval scenarios: “Golden Only” (only golden retrieval context), “Golden & ON” (golden context + orthographic noise), and “Golden & SeN” (golden context + semantic noise).

factually incorrect”. As shown in Table 2, results show an average accuracy of 79.93% across eight LLMs when handling prior noise. However, when models fail to identify prior errors and continue retrieval, performance drops significantly to 34.20%. This underscores the need to detect prior errors in user queries before answering.

### Beneficial Noise Enhances Performance Across Models

We consider both model architectures (Figure 5) and RAG system designs (Table 3) to demonstrate the positive effects of beneficial noise across various models. We present results for illegal sentence noise here. Additionally, since prior research has highlighted the positive effect of semantic noise (Cuconasu et al. 2024), our subsequent discussion will focus on two types: datatype noise and illegal sentence noise.

**(1) Results across various architectures and scales** As shown in Figure 5, we evaluate the impact of illegal sentence noise (ISN) on eight LLMs (different architectures and scales) by calculating 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 model performance in all scenarios, with the most substantial improvement under harmful noise. To better illustrate the impacts of certain noise types, which may not be immediately apparent in tabular form, we plot their performance across multiple models using line graphs (Figure 6) under three conditions: golden only, golden & orthographic noise, and golden & semantic noise. These visualizations clearly demonstrate the negative effect of orthographic noise and the slight performance boost provided by semantic noise.

**(2) Noise effects on specialized RAG models** As illustrated in Table 3, introducing illegal sentence noise to the specialized RAG model Self-RAG (Asai et al. 2024) consis-

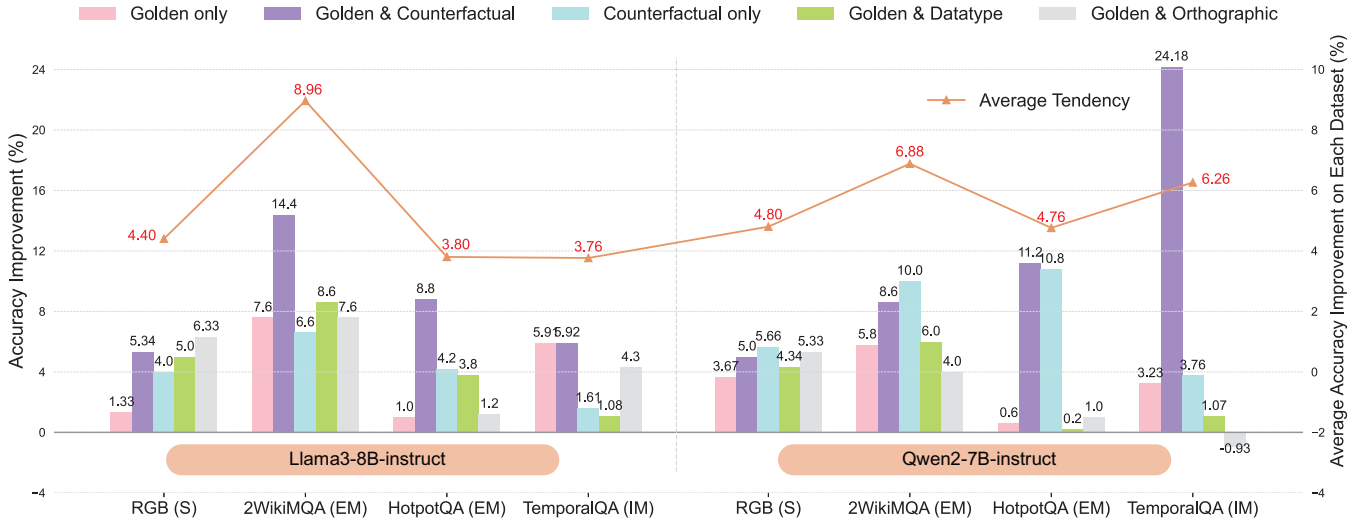


Figure 7: Results for the impact of illegal sentence noise on the Llama3-8B-instruct and Qwen2-7B-instruct models when exposed to five typical noise categories across four datasets, including both single-hop (S) and multi-hop (explicit: EM, implicit: IM) reasoning tasks. The bar charts show performance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

Table 3: Effects of beneficial noise on Self-RAG (13B). We assess performance through 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

tently enhances model performance across various datasets (NQ, RGB, and StrategyQA) and scenarios (without noise, with harmful and beneficial noise). This further validates the positive effects of beneficial noise.

In conclusion, based on our comprehensive analysis, we can classify illegal sentence noise, datatype noise, and semantic noise as beneficial, while counterfactual, supportive, and orthographic noises are categorized as harmful.

**Beneficial Noise Remains Effective Under Other Noise Disturbances** To illustrate the impact of beneficial noise under other noise disturbances, we analyze the effect of illegal sentence noise (ISN) in five 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 7 shows the model’s accuracy gains after introducing ISN in these scenarios. We find that ISN generally enhances performance across all datasets, particularly when combined with harmful noise like counterfactual noise, with average accuracy improvements exceeding 10 percentage points. The consistent positive effects of ISN in various real-world scenarios underscore its potential significance for future RAG research.

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

**Beneficial Noise Is Statistically Significant** 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 statistical 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.

### Exploring the Mechanisms Behind Beneficial Noise

We investigate why beneficial noise positively impacts RAG systems. We propose 3 hypotheses that beneficial noise contributes to:

- **H1: Clearer and more explicit reasoning process**
- **H2: More standardized response formats**
- **H3: Increased confidence with golden context**

We confirm them through case study and statistical analysis.

Table 5: Examples of LLM outputs 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.

<b>Retrieved Context wo BN</b>	[1] Paul R. Ehrlich, author of "The Population Bomb," was born on <b>June 14, 1931</b> . Ehrlich, a renowned biologist, published his influential book on population control in 1968, which sparked . . . <b>[Counterfactual]</b> [2] The author of The Population Bomb, Paul Ehrlich, was born on <b>May 29, 1932</b> . He is an American biologist and educator, known for his warnings about the consequences of overpopulation . . . <b>[Golden Context]</b> ...
<b>Question</b>	When was the author of The Population Bomb born? 'A': June 14, 1931 'B': Uncertain 'C': <b>May 29, 1932 (Ground Truth)</b> 'D': April 22, 1934
<b>Answer wo BN</b>	<b>A</b> ###\n\n### Explanation:\n\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 <b>June 14, 1931</b> . The other options are incorrect, as they provide different birth dates for the author. <b>✗</b>
<b>Answer w BN</b>	<b>C</b> \n### Explanation: According to information [2], Paul Ehrlich, the author of The Population Bomb, was born on <b>May 29, 1932</b> . 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. <b>✓</b>

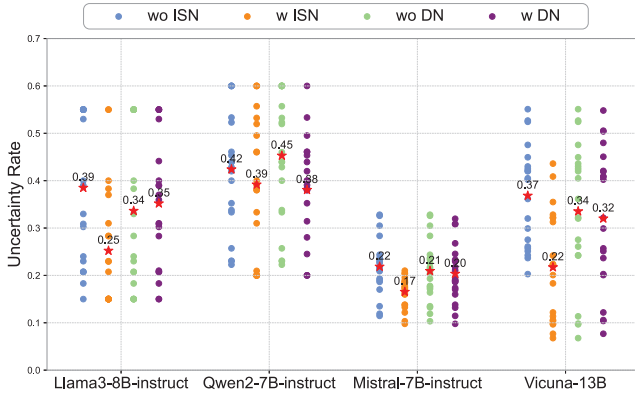


Figure 8: The effects of beneficial noise on LLM output uncertainty (anti-confidence). 'ISN' and 'DN' denote Illegal Sentence Noise and Datatype Noise, respectively. The red star ★ represents the mean uncertainty rate ( $\mu$ ).

**Case Study** Table 5 presents the complete reasoning and generation process of Llama3-8B-instruct on the multi-hop dataset Bamboogle. When exposed to harmful noise without any beneficial noise, the model ignores correct information and exhibits logical flaws under the influence of 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 LLM's 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 Analysis** 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**).

Second, as shown in Figure 8, we analyze the impact of beneficial noise on LLM output uncertainty across four powerful LLMs. The 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 the provided golden context and respond with greater confidence (**H3**).

## Conclusion

In this paper, we provide clear definitions for seven types of RAG noise and categorize them into two groups: beneficial and harmful noise. This is the first comprehensive study to explore retrieval noise from both 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. Evaluated on eight representative LLMs, extensive experimental results reveal the role that different noise plays in RAG systems. The most surprising finding is that beneficial noise can act like the power of Aladdin's Lamp and enhance model performance by leading to clearer reasoning paths, more standardized answers, and increased confidence. We anticipate that future research will propose methods to fully leverage beneficial mechanisms of noise while avoiding the negative effects of harmful noise.



## References

- Aloufi, A. 2021. Language and Linguistic Orthography. *English Language and Literature Studies*, 11(3).
- Asai, A.; Wu, Z.; Wang, Y.; Sil, A.; and Hajishirzi, H. 2024. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. In *The Twelfth International Conference on Learning Representations*.
- Bubeck, S.; Chandrasekaran, V.; Eldan, R.; Gehrke, J.; Horvitz, E.; Kamar, E.; Lee, P.; Lee, Y. T.; Li, Y.; Lundberg, S.; et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. arXiv:2303.12712.
- Chafe, W. L. 1971. Linguistics and human knowledge. *Monograph series on languages and linguistics*, (24): 57.
- Chen, J.; Lin, H.; Han, X.; and Sun, L. 2024. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 17754–17762. AAAI Press.
- Chiang, W.-L.; Li, Z.; Lin, Z.; Sheng, Y.; Wu, Z.; Zhang, H.; Zheng, L.; Zhuang, S.; Zhuang, Y.; Gonzalez, J. E.; Stolica, I.; and Xing, E. P. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%\* ChatGPT Quality. <https://lmsys.org/blog/2023-03-30-vicuna/>.
- Cuconasu, F.; Trappolini, G.; Siciliano, F.; Filice, S.; Campagnano, C.; Maarek, Y.; 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*, 719–729. New York, NY, USA: Association for Computing Machinery.
- Fang, F.; Bai, Y.; Ni, S.; Yang, M.; Chen, X.; and Xu, R. 2024. Enhancing Noise Robustness of Retrieval-Augmented Language Models with Adaptive Adversarial Training. arXiv:2405.20978.
- Feng, G.; and Yi, L. 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.
- Gao, Y.; Xiong, Y.; Gao, X.; Jia, K.; Pan, J.; Bi, Y.; Dai, Y.; Sun, J.; Wang, M.; and Wang, H. 2024. Retrieval-Augmented Generation for Large Language Models: A Survey. arXiv:2312.10997.
- Geva, M.; Khashabi, D.; Segal, E.; Khot, T.; Roth, D.; and Berant, J. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9: 346–361.
- Grandvalet, Y.; Canu, S.; and Boucheron, S. 1997. Noise injection: Theoretical prospects. *Neural Computation*, 9(5): 1093–1108.
- Huang, L.; Yu, W.; Ma, W.; Zhong, W.; Feng, Z.; et al. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. arXiv:2311.05232.
- Izacard, G.; Caron, M.; Hosseini, L.; Riedel, S.; Bojanowski, P.; Joulin, A.; and Grave, E. 2022. Unsupervised Dense Information Retrieval with Contrastive Learning. arXiv:2112.09118.
- Jia, Z.; Abujabal, A.; Saha Roy, R.; Strötgen, J.; and Weikum, G. 2018. Tempquestions: A benchmark for temporal question answering. In *Companion Proceedings of the The Web Conference 2018*, 1057–1062. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee.
- Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; et al. 2023. Mistral 7B. arXiv:2310.06825.
- Jiang, A. Q.; Sablayrolles, A.; Roux, A.; Mensch, A.; et al. 2024. Mixtral of Experts. arXiv:2401.04088.
- Kandpal, N.; Deng, H.; Roberts, A.; Wallace, E.; and Raffel, C. 2023. Large language models struggle to learn long-tail knowledge. In *International Conference on Machine Learning*, 15696–15707. PMLR.
- Karpukhin, V.; Oğuz, B.; Min, S.; Lewis, P.; Wu, L.; Edunov, S.; Chen, D.; and tau Yih, W. 2020. Dense Passage Retrieval for Open-Domain Question Answering. arXiv:2004.04906.
- Kertész, A.; and Rákosi, C. 2012. *Data and evidence in linguistics: A plausible argumentation model*. Cambridge University Press.
- Kotz, S.; and Johnson, N. L., eds. 1992. *Breakthroughs in Statistics: Methodology and Distribution*. New York, NY: Springer New York.
- Kwiatkowski, T.; Palomaki, J.; Redfield, O.; Collins, M.; Parikh, A.; Alberti, C.; Epstein, D.; Polosukhin, I.; Devlin, J.; Lee, K.; et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7: 453–466.
- Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; et al. 2019. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. arXiv:1910.13461.
- Meta, AI. 2024. Introducing meta llama 3: The most capable openly available llm to date. <https://ai.meta.com/blog/meta-llama-3/>.
- Olan, F.; Jayawickrama, U.; Arakpogun, E. O.; Suklan, J.; and Liu, S. 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. Textnoir: Adding random noise to a dataset. <https://github.com/preligens-lab/textnoir>.
- Press, O.; Zhang, M.; Min, S.; Schmidt, L.; Smith, N. A.; and Lewis, M. 2022. Measuring and narrowing the compositionality gap in language models. *arXiv preprint arXiv:2210.03350*.
- Qu, Y.; Ding, Y.; Liu, J.; Liu, K.; Ren, R.; 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*, 5835–5847. Online: Association for Computational Linguistics.
- Seth, I.; Lim, B.; Xie, Y.; Cevik, J.; Rozen, W. M.; Ross, R. J.; and Lee, M. 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.
- Shannon, C.; Weaver, W.; and Hockett, C. 1961. The mathematical theory of communication. *Urbana: University of Illinois*.
- Shi, F.; Chen, X.; Misra, K.; Scales, N.; Dohan, D.; Chi, E. H.; Schärli, N.; et al. 2023. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, 31210–31227. PMLR.
- Skeat, W. W. 1993. *The concise dictionary of English etymology*. Wordsworth Editions.
- Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; et al. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv:2307.09288.
- Tumarkin, R.; and Whitelaw, R. F. 2001. News or noise? Internet postings and stock prices. *Financial Analysts Journal*, 57(3): 41–51.
- Wang, W.; Bao, H.; Huang, S.; Dong, L.; and Wei, F. 2021. MiniLMv2: Multi-Head Self-Attention Relation Distillation for Compressing Pretrained Transformers. arXiv:2012.15828.
- Wang, Y.; Li, P.; Sun, M.; and Liu, Y. 2023. Self-Knowledge Guided Retrieval Augmentation for Large Language Models. arXiv:2310.05002.
- Wang, Z.; Liu, A.; Lin, H.; Li, J.; Ma, X.; and Liang, Y. 2024. RAT: Retrieval Augmented Thoughts Elicit Context-Aware Reasoning in Long-Horizon Generation. arXiv:2403.05313.
- Welbl, J.; Stenetorp, P.; and Riedel, S. 2018. Constructing datasets for multi-hop reading comprehension across documents. *Transactions of the Association for Computational Linguistics*, 6: 287–302.
- Wu, J.; Che, F.; Zheng, X.; Zhang, S.; Jin, R.; Nie, S.; Shao, P.; and Tao, J. 2024. Can large language models understand uncommon meanings of common words? arXiv:2405.05741.
- Wu, J.; Ning, Z.; Ding, Y.; Wang, Y.; Peng, Q.; and Fu, L. 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.
- Xiang, C.; Wu, T.; Zhong, Z.; Wagner, D.; Chen, D.; and Mittal, P. 2024. Certifiably Robust RAG against Retrieval Corruption. arXiv:2405.15556.
- Xie, J.; Zhang, K.; Chen, J.; Lou, R.; and Su, Y. 2024. Adaptive Chameleon or Stubborn Sloth: Revealing the Behavior of Large Language Models in Knowledge Conflicts. arXiv:2305.13300.
- Yang, A.; Xiao, B.; Wang, B.; Zhang, B.; Bian, C.; Yin, C.; Lv, C.; Pan, D.; et al. 2023. Baichuan 2: Open Large-scale Language Models. arXiv:2309.10305.
- Yang, A.; Yang, B.; Hui, B.; Zheng, B.; Yu, B.; Zhou, C.; et al. 2024. Qwen2 Technical Report. arXiv:2407.10671.
- Yang, Z.; Qi, P.; Zhang, S.; Bengio, Y.; Cohen, W. W.; Salakhutdinov, R.; and Manning, C. D. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. arXiv:1809.09600.
- Ye, J.; Xu, N.; Wang, Y.; Zhou, J.; Zhang, Q.; Gui, T.; and Huang, X. 2024. LLM-DA: Data Augmentation via Large Language Models for Few-Shot Named Entity Recognition. arXiv:2402.14568.
- Yoran, O.; Wolfson, T.; Ram, O.; and Berant, J. 2024. Making Retrieval-Augmented Language Models Robust to Irrelevant Context. In *The Twelfth International Conference on Learning Representations*.
- Zhao, P.; Zhang, H.; Yu, Q.; Wang, Z.; Geng, Y.; Fu, F.; Yang, L.; Zhang, W.; Jiang, J.; and Cui, B. 2024. Retrieval-Augmented Generation for AI-Generated Content: A Survey. arXiv:2402.19473.
- Zheng, X.; Che, F.; Wu, J.; Zhang, S.; Nie, S.; Liu, K.; and Tao, J. 2024. KS-LLM: Knowledge Selection of Large Language Models with Evidence Document for Question Answering. arXiv:2404.15660.
- Zhu, K.; Wang, J.; Zhou, J.; Wang, Z.; Chen, H.; Wang, Y.; Yang, L.; Ye, W.; et al. 2024. PromptRobust: Towards Evaluating the Robustness of Large Language Models on Adversarial Prompts. arXiv:2306.04528.