Worse than Zero-shot? A Fact-Checking Dataset for Evaluating the Robustness of RAG Against Misleading Retrievals

Linda Zeng[*]
26lindaz@students.harker.org
The Harker School
San Jose, California, USA

Rithwik Gupta[*]
rithwikca2020@gmail.com
Irvington High School
Fremont, California, USA

Divij Motwani
divijmotwani@gmail.com
Palo Alto High School
Palo Alto, California, USA

Diji Yang[†]
dyang39@ucsc.edu
University of California Santa Cruz
Santa Cruz, California, USA

Yi Zhang[†]
yiz@ucsc.edu
University of California Santa Cruz
Santa Cruz, California, USA

ABSTRACT

Retrieval-augmented generation (RAG) has shown impressive capabilities in mitigating hallucinations in large language models (LLMs). However, LLMs struggle to handle misleading retrievals and often fail to maintain their own reasoning when exposed to conflicting or selectively-framed evidence, making them vulnerable to real-world misinformation.

In such real-world retrieval scenarios, misleading and conflicting information is rampant, particularly in the political domain, where evidence is often selectively framed, incomplete, or polarized. However, existing RAG benchmarks largely assume a clean retrieval setting, where models succeed by accurately retrieving and generating answers from gold-standard documents. This assumption fails to align with real-world conditions, leading to an overestimation of RAG system performance.

To bridge this gap, we introduce RAGUARD, a fact-checking dataset designed to evaluate the robustness of RAG systems against misleading retrievals. Unlike prior benchmarks that rely on synthetic noise, our dataset constructs its retrieval corpus from Reddit discussions, capturing naturally occurring misinformation. It categorizes retrieved evidence into three types: *supporting*, *misleading*, and *irrelevant*, providing a realistic and challenging testbed for assessing how well RAG systems navigate different retrieval information.

Our benchmark experiments reveal that when exposed to misleading retrievals, all tested LLM-powered RAG systems perform worse than their zero-shot baselines (i.e., no retrieval at all), highlighting their susceptibility to noisy environments. To the best of our knowledge, RAGUARD is the first benchmark to systematically assess RAG robustness against misleading evidence. We expect this benchmark will drive future research toward improving RAG

systems beyond idealized datasets, making them more reliable for real-world applications.[1]

CCS CONCEPTS

• **Information systems** → **Question answering**; **Language models**; • **Computing methodologies** → Natural language processing.

KEYWORDS

Retrieval-Augmented Generation (RAG) benchmark, Fact-checking dataset, Noisy retrieval corpus, Misleading retrievals

1 INTRODUCTION

Retrieval-augmented generation (RAG) systems have shown significant promise in mitigating LLM hallucination and enhancing trustworthiness. By combining the generative capabilities of large language models (LLMs) with the retrieval power of external corpora, RAG aims to ground responses in relevant, contextually appropriate information, thereby improving factual consistency and output credibility [12, 16, 24]. However, while existing RAG approaches primarily focus on optimizing retrieval relevance and maximizing the amount of information in retrieved-context [6, 20, 43], a critical challenge remains largely unaddressed: how to handle cases where retrieved content is misleading or irrelevant. This issue is particularly concerning when misinformation, adversarial perturbations, or biased sources influence the retrieval process, potentially degrading the reliability of LLM outputs. Addressing this robustness gap is essential for ensuring the trustworthiness of RAG systems, especially in high-stakes applications such as fact-checking [33] and legal or medical domains [15, 42].

Prior work has mitigated noisy retrievals by prompting models to justify relevance, aggregating sources, or using debate-based selection [36, 38, 40]. However, most approaches align retrieved content with LLMs' prior knowledge rather than addressing real-world contradictions [21, 35]. Furthermore, current datasets overly rely

---

---

[1]The dataset is available at https://huggingface.co/datasets/UCSC-IRKM/RAGuard.
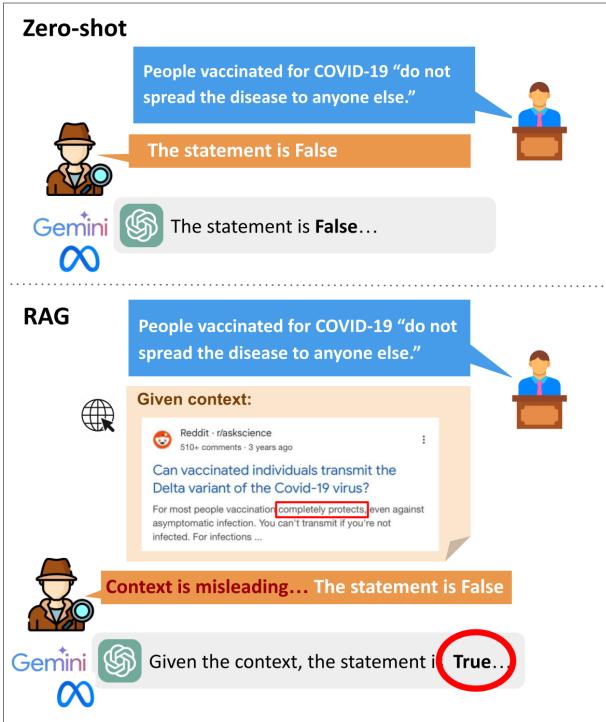
Figure 1: Example of a false claim initially classified correctly but later misclassified as true due to misleading retrieved-context, alongside the ideal human judgment.

on curating reliable documents, limiting robustness testing against misinformation [19, 23, 30, 44]. While some introduce counterfactuals or retrieval noise [8, 25], they rely on artificial perturbations or costly human annotation. This highlights the need for an evaluation framework that challenges RAG systems with real-world contradictions, exposing their limitations and improving resilience in complex retrieval scenarios.

Fact-checking plays a crucial role in combating misinformation, yet most existing datasets assume the availability of gold-standard evidence that aligns with the verdict [2, 4, 17, 22, 34, 45]. In reality, retrieved information often presents conflicting perspectives, making automated verification more challenging. The political domain is particularly rich in such complexities, as controversial claims generate both supporting and opposing narratives from diverse sources [27, 29, 32, 34]. To develop fact-checking systems capable of handling real-world misinformation, it is essential to move beyond idealized settings and expose models to the conflicting and misleading evidence that humans work with in the real-world.

To bridge this gap, we introduce RAGuard, a benchmark dataset based on political discourse claims and their verifications from PolitiFact incorporating real-world misinformation. Given the prevalence of polarizing and deceptive information in political discourse, we develop an automated pipeline that retrieves relevant yet potentially misleading documents from Reddit via Google Search. Reddit, with its diverse and often controversial user-generated content, serves as a realistic source for challenging retrieval scenarios.

We introduce a novel LLM-guided approach to annotate retrieved documents by simulating a fact-checking exam. This method labels documents as *supporting, misleading,* or *irrelevant* based on their impact on the LLM's decision, providing a scalable benchmark to evaluate RAG systems in real-world, noisy retrieval scenarios. Each data point in our dataset consists of a claim, a fact-checking verdict, and multiple labeled associated documents. This structure enables a rigorous evaluation of the ability of RAG systems to navigate situations with both noisy and supporting information, reflecting real-world conditions where accurate retrieval cannot be guaranteed. Our benchmark supports verifying robustness on documents solely labeled as misleading or on the full dataset to systematize generalization capabilities in complex scenarios.

We evaluate widely used LLMs and RAG systems, testing their ability to predict the correct fact-checking verdict across three task configurations: Zero-Context Prediction (given claims with no retrieved documents), Standard RAG (given claims with retrieved documents), and Oracle Retrieval (given claims with their associated documents). Our results reveal that current LLMs are highly vulnerable and lack robustness in real-world scenarios. Performance drops significantly across all configurations when using the RAGuard knowledge base. Notably, incorporating associated documents as context leads to an even steeper decline compared to dynamically retrieved documents, demonstrating how our dataset effectively assigns impactful misinformation to claims. This further exposes the limitations of LLMs in handling misleading content. Qualitative analysis shows that RAG systems are particularly susceptible to overtly misleading information, falling short of human reasoning. Figure 1 highlights the motivation behind our dataset and the susceptibility of current RAG systems to misleading retrievals.

In summary, our work advocates for a shift in focus from developing idealized RAG settings to those that better simulate real-world noisy information. We provide a benchmark to evaluate the robustness of RAG systems against misleading retrievals, addressing the gap in naturally-occurring misinformation RAG datasets. Our baseline results reveal the current shortcomings of RAG systems, showing performance worse than zero-shot. We expect our dataset to contribute to the development of more reliable and resilient RAG systems in the future.

2 DATASET

We introduce RAGuard, a benchmark for evaluating the robustness of RAG systems in political fact-checking. RAGuard simulates noisy real-world retrieval settings, where systems must navigate supporting, misleading, and irrelevant evidence. The dataset comprises 2,648 political claims, corresponding fact-checking verdicts, and 16,331 associated documents labeled by their agreement with the verdicts.

2.1 Definitions

The main task in RAGuard is retrieval-augmented fact-checking, where claims are verified as *true* or *false* based on retrieved documents that may support, mislead, or be irrelevant to the claim. Prior works employ varying terminology to describe the presence of such noise in retrieved contexts or retrieval corpora [8, 11, 25, 39]. To
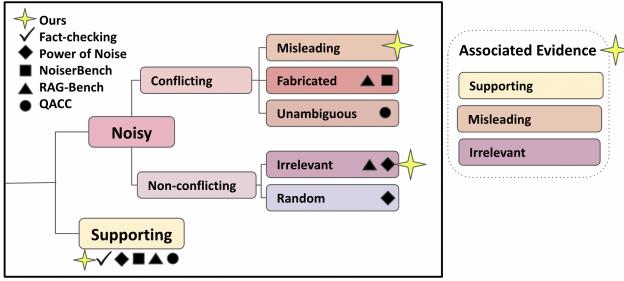
Figure 2: Taxonomy of terminology to classify different types of evidence, labeled with prior works' contributions (*left*), and our dataset's composition (*right*).

Dataset	Evidence Retrieval	Conflicting Evidence	Real World	Domain	Claims
Fact-Checking					
FEVER [34]	✓	✗	✗	General	185K
FEVEROUS [2]	✓	✗	✗	General	87K
Liar [37]	✗	✗	✓	Political	12.8K
Mocheg [45]	✓	✗	✓	Political	15.6K
Snopes [17]	✓	✗	✓	Political	6.4K
PUBHEALTH [22]	✓	✗	✓	Health	11.8K
MultiFC [4]	✓	✗	✓	Political	43.8K
Noisy Contexts					
Power of Noise [8]	✓	✗	✓	General	10K
RAAT [11]	✓	✓	✗	General	7.8K
NoiserBench [39]	✓	✓	✗	General	4K
QACC [25]	✓	✓	✓	General	1.5K
RAGUARD	✓	✓	✓	Political	2.6K

Table 1: Comparison of RAGUARD with related fact-checking and RAG datasets. The columns indicate whether the dataset requires automatic evidence retrieval, contains conflicting evidence documents, and consists of naturally occurring real-world claims and evidence, as well as their domain and size.

establish consistency, we define a structured taxonomy and align existing definitions (See Figure 2).

Typical RAG datasets, including all prior fact-checking datasets to our knowledge, exclusively contain non-noisy *supporting* documents as associated evidence, leading to overly optimistic performance [8]. Instead of relying solely on answer-containing documents, our dataset adopts a broader notion of supporting evidence. Specifically, we consider a document to be *supporting* if it provides information that enables an LLM to infer the correct answer, even if it does not explicitly state the ground-truth output. This reflects real-world fact-checking, where human verifiers rely on contextual information rather than single authoritative documents.

We categorize different types of noisy evidence based on whether the information directly conflicts with aspects of the correct prediction. As in prior work [8, 11], we include non-conflicting documents in RAGUARD, such as irrelevant texts that may hurt performance. However, our primary focus is conflicting documents, which include misleading, fabricated, and unambiguous evidence. Previous datasets primarily include conflicting evidence as fabricated or unambiguous documents, oversimplifying real-world complexity and ambiguity (see Section 2.2 for further discussion) [11, 25, 39]. Notably, no prior work has introduced *misleading* documents.

In RAGUARD, misleading documents distort facts through selective framing, omission, or biased presentation, leading the system toward incorrect predictions while still containing partial truths. Unlike fabricated evidence, which is explicitly engineered to contradict the correct prediction (i.e., adversarial perturbations), misleading evidence subtly misguides the model rather than directly opposing it. Additionally, while prior work such as QACC [25] introduces unambiguous evidence—a term we adopt to ensure consistency with past research—which includes some naturally conflicting evidence but only for a limited set of unambiguous questions, we focus on more natural yet scalable conflicting evidence.

For reference, we provide a list of all defined terms. Each term defines a type of document or piece of evidence.

(1) *Associated:* any document linked to a claim, regardless of label
(2) *Supporting:* aids the system in producing a correct prediction through containing the correct answer explicitly or providing contextual support

(3) *Noisy:* challenges or disrupt system performance, thereby enhancing robustness
(4) *Conflicting:* contradicts either the correct answer or some aspect of the prediction
(5) *Misleading:* introduces factual distortions through selective framing, omission, or biased presentation; may contain partial truths
(6) *Fabricated:* synthetically constructed to include factual errors (e.g., adversarial perturbations)
(7) *Unambiguous:* naturally conflicting evidence but only for a limited set of unambiguous questions (special case of [25])
(8) *Non-Conflicting:* does not directly contradict the correct answer but still introduces noise by distracting the model
(9) *Irrelevant:* does not contain specific enough information to determine the correct prediction, despite being topically or semantically related to the query
(10) *Random:* unrelated; often introduced through random selection or artificial generation

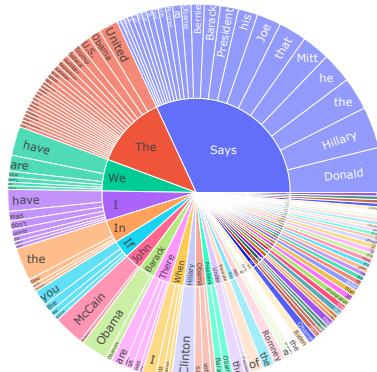## 2.2 Comparison with Existing Datasets

Table 1 depicts a comparison of RAGUARD with other fact-checking and RAG datasets.

Fact-Checking Datasets. All existing fact-checking datasets that include evidence retrieval contain only *supporting* documents. While FEVEROUS [2] labels some documents as *refute*, this terminology is misleading—these documents actually support the falsehood of the claim rather than providing conflicting evidence. Therefore, while FEVEROUS categorizes evidence into *support* for true claims and *refute* for false claims, it does not include documents that actively contradict the claim's verdict. Additionally, both FEVER [34] and FEVEROUS [2] rely on rewritten Wikipedia statements rather than naturally occurring claims, as noted by [4].
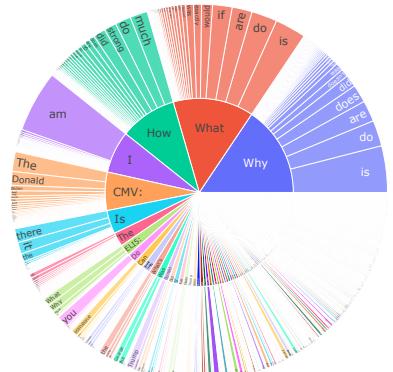
Liar [37] and Mocheg [45] are the most similar to our dataset since they also source claims from Politifact. However, Liar [37] does not support evidence retrieval as it lacks evidence documents. Mocheg [45] includes only documents cited by Politifact fact-checkers,

Statistic	Number
Total Claims	**2,648**
True	1,333 (50.3%)
False	1,315 (49.7%)
Avg. Claim Length (words)	17.6
Total Documents	**16,331**
Supporting	2,685 (16.4%)
Misleading	1,812 (11.1%)
Irrelevant	11,834 (72.5%)
Avg. Document Length (words)	161
Avg. Documents Per Claim	6.2
Claims with Supporting Docs	955 (36.1%)
Claims with Misleading Docs	788 (29.8%)

(a) Main statistics of RAGᴜᴀʀᴅ.

(b) Word distribution for claims.

(c) Word distribution for documents.

Figure 3: Key statistics and word distributions of RAGᴜᴀʀᴅ.

which support the verdict rather than introducing conflicting or misleading evidence. Other datasets [4, 17, 22] primarily use journalist-written explanations from fact-checking websites, which are structured to justify the verdict rather than reflect the complexity of real-world misinformation.

In contrast, RAGᴜᴀʀᴅ explicitly incorporates conflicting evidence, making it more representative of real-world misinformation challenges. Unlike curated fact-checking content, our dataset sources evidence from Reddit discussions, which naturally contain misleading information through diverse viewpoints. This increases the difficulty of RAG by better aligning ot with real-world misinformation.

Datasets with Noisy Contexts. Prior datasets that include *noisy* evidence primarily build on open-domain question answering (QA) datasets [8, 11, 25, 39]. As these datasets differ in how they define and introduce noise, we use our definition framework in Figure 2 to better distinguish each dataset.

Power of Noise [8] classifies evidence into gold, relevant, distracting, and random categories. Gold and relevant documents serve as supporting evidence, with *gold* documents being preexisting gold-standard documents and *relevant* documents being newly retrieved documents that explicitly contain the correct answer. Non-conflicting evidence includes distracting documents, which are simply non-gold retrievals and therefore irrelevant. Additionally, Power of Noise is the only dataset that introduces *random* evidence, which is entirely unrelated to the query. Notably, it does not include any *conflicting* evidence.

RAG-Bench [11] classifies evidence into golden context, irrelevant retrieval noise, relevant retrieval noise, and counterfactual retrieval noise. It constructs *supporting* evidence using gold-standard documents from non-noisy QA datasets. It introduces *fabricated* evidence by modifying documents to contain incorrect answers. While this results in factually incorrect *conflicting* evidence, it does not capture *misleading* evidence, which may contain partial truths but manipulates the information through selective framing or omission. Both relevant and irrelevant retrieval noise are considered *irrelevant*, as their retrieval via semantic search implies a degree

of semantic relatedness to the query while not directly conflicting with the task content. Nonetheless, they can still distract RAG systems.

NoiserBench [39] introduces a wide range of types of noise, including inserting counterfactual noise. While this introduces *conflicting* evidence, all noise is artificially constructed, limiting its reflection of real-world misleading information. Like RAG-Bench, it focuses on *fabricated* evidence rather than capturing more complex distortions such as selective framing or omission.

QACC [25] employs human annotators to label retrieved documents as conflicting or non-conflicting with answers from AmbigQA. Rather than artificially injecting errors, it includes *conflicting* evidence that directly contradicts the correct answer. However, its reliance on human annotation limits scalability, and its approach does not fully capture real-world ambiguity, as it focuses on clear-cut conflicts (i.e., questions labeled as "unambiguous" from AmbigQA) rather than more nuanced misleading evidence.

Unlike these datasets, our work focuses on the political domain, which presents distinct challenges. Political misinformation has tangible consequences, influencing public opinion, policy decisions, and elections [37]. Misleading evidence in political discourse is often more nuanced, relying on selective framing rather than outright falsehoods. Furthermore, political fact-checking requires domain-specific reasoning, as claims are frequently shaped by ideological bias and rhetorical strategies.

2.3 Dataset Structure

RAGᴜᴀʀᴅ consists of 2,648 political claims made by U.S. presidential candidates (2000–2024), each labeled as either *true* or *false*, and a knowledge base comprising 16,331 documents. The dataset's key statistics are presented in Table 3a. Each claim is linked to a set of associated documents, categorized as *supporting*, *misleading*, or *irrelevant*, with an average of 6.2 documents per claim. Notably, the dataset contains more *supporting* documents than *misleading* ones, reflecting that political discussions online are more often aligned with factual information. However, not every claim has both misleading and supporting documents, highlighting the imbalanced nature of political discourse, where certain narratives dominate while others lack counterpoints. The dataset is provided in two
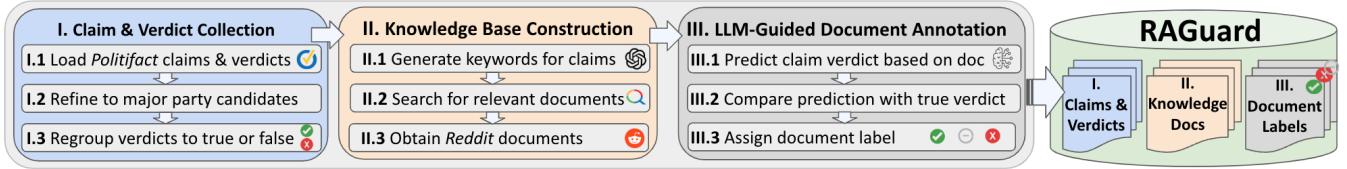
Figure 4: RAGᴜᴀʀᴅ dataset construction, consisting of three stages to obtain claims and verdicts; associated documents; and labels for the each document's relationship to the claim and verdict.

comma-separated value (CSV) files, with an example depicted in Tables 2 and 3.

RAGᴜᴀʀᴅ contains a diverse set of claims and documents. Figures 3b and 3c visualize the most frequent opening words in claims and documents. Claims tend to reference well-known political figures and events, whereas documents frequently begin with questions, reflecting the uncertainty and variability inherent in real-world online discussions. For clarity, words occurring fewer than three times are omitted from the visualization.

Claims.csv	
ID	1517
Claim	Insulin for Medicare beneficiaries dec...
Verdict	True
Associated Docs	[9400, 9402, 9405, ...]
Document Labels	[irrelevant, supporting, misleading, ...]

Table 2: Example of data structure of claims.

Documents.csv	
ID	9405
Title	My dad is spending $700/mo on insulin...
Full Text	...
Document Label	Misleading
Associated Claim	1517

Table 3: Example of data structure of documents.

2.4 Supported Tasks

To benchmark the performance of current RAG systems in real-world fact-checking scenarios, we define a series of tasks using RAGᴜᴀʀᴅ. Each task evaluates a different aspect of RAG system robustness against misleading or conflicting contextual data.

Zero-Context Prediction. This task assesses a RAG system's ability to fact-check claims without external contextual information. The goal is to evaluate the intrinsic knowledge encoded in each model during pre-training and its effectiveness in verifying claims. This serves as a baseline to measure the impact of external retrieval on performance.

Standard RAG. This task simulates a real-time RAG system retrieving documents from the entire dataset corpus. The retrieved documents may include supporting, misleading, or irrelevant information to the claim, introducing retrieval noise. In extreme cases of poor retrieval, models may receive documents unrelated to the claim, mimicking real-world retrieval limitations.

Oracle Retrieval. This task provides RAG systems with the ssociated documents for each claim, isolating the impact of the associated documents labeled in our dataset. Unlike the previous task, where retrieval noise is unpredictable, this setting ensures systems receive only the supporting, misleading, and irrelevant documents associated with each claim. This setup evaluates how well models can filter out deceptive content when retrieval errors are controlled. We include two specific evaluations for this task. In the first, each model receives an associated document for a claim, regardless of the categorization as *supporting*, *misleading*, or *irrelevant*, to assess its ability to generalize to complex scenarios where documents have potential to support, mislead, or be irrelevant. In the second, we isolate only instances where the associated document is labeled as *misleading*, which conflicts with the claim's ground truth fact-checking verdict, testing susceptibility to misleading information.

3 DATASET CONSTRUCTION

We construct RAGᴜᴀʀᴅ in three stages, depicted in Figure 4. First, we collect political claims and fact-checking labels from PolitiFact, a reputable source for verified political fact-checking information. Next, we construct a knowledge base by retrieving relevant Reddit documents via a search engine, leveraging its diverse, real-time content to reflect real-world discourse. Finally, we introduce a novel, scalable LLM-guided approach to classify documents as misleading, supporting, or irrelevant by simulating the LLM taking an exam. The following sections outline the details of each stage.

3.1 Claim and Verdict Collection

To collect claims and fact-checking verdicts for RAGᴜᴀʀᴅ, we scrape PolitiFact,[2] a reputable platform where expert journalists assess the truthfulness of a wide range of political claims. We focus on claims made by major U.S. presidential candidates from 2000 to 2024 to ensure the inclusion of widely discussed statements that have been frequently fact-checked and debated, generating substantial online discourse and reflecting politically significant information. To facilitate document retrieval, we condense PolitiFact's six-point truth scale (*true, mostly true, half true, mostly false, false, pants on fire*) into binary labels—*true* and *false*, as it is challenging for a document to specifically mislead a *half true* verdict, undermining our core contribution.

3.2 Knowledge Base Construction

To construct the RAGᴜᴀʀᴅ knowledge base, we employ a multi-step retrieval process that balances coverage, diversity, and realism. First,

---

[2]https://www.politifact.com/

	Gemini 1.5 Flash	GPT-4o Mini	Claude 3.5 Sonnet	Llama 3	Mistral
Task 1: Zero-Context Prediction	**61.06**	**67.33**	**74.51**	**62.50**	**63.97**
Task 2: Standard RAG					
RAG-1	56.68 ↓ -4.38% / -7.2%	64.80 ↓ -2.53% / -3.8%	70.09 ↓ -4.42% / -5.9%	59.40 ↓ -3.10% / -5.0%	59.14 ↓ -4.83% / -7.5%
RAG-5	57.59 ↓ -3.47% / -5.7%	65.90 ↓ -1.43% / -2.1%	68.58 ↓ -5.93% / -8.0%	61.37 ↓ -1.13% / -1.8%	58.91 ↓ -5.06% / -7.9%
Task 3: Oracle Retrieval					
All Documents	52.38 ↓ -8.68% / -14.2%	53.22 ↓ -14.11% / -20.9%	51.17 ↓ -23.34% / -31.3%	61.09 ↓ -1.41% / -2.3%	51.61 ↓ -12.36% / -19.3%
Misleading-Only	30.57 ↓ -30.49% / -49.9%	45.97 ↓ -21.36% / -31.7%	37.05 ↓ -37.46% / -50.3%	36.81 ↓ -25.69% / -41.1%	28.22 ↓ -35.75% / -55.9%

Table 4: Performance of various LLM backbones in RAG setup on three tasks, reported in Accuracy (%). The red numbers indicate the absolute/relative accuracy drop compared to Zero-Context Prediction (Task 1) under each setting.

GPT-4 extracts the keywords from each claim. This keyword expansion ensures a broader and more nuanced search space, increasing the likelihood of retrieving both corroborating and contradicting information. Next, we perform a keyword-based Google Search to retrieve up to ten relevant Reddit posts per claim. Google's ranking ensures contextual relevance, while Reddit's user-generated content introduces diverse perspectives, from speculative theories to well-supported arguments. Unlike curated fact-checking datasets, Reddit captures real-world discourse, including misinformation and conflicting viewpoints. This retrieval pipeline creates a realistic testbed for fact-checking, combining GPT-4-assisted keyword expansion with search engine retrieval to mirror the complexities of real-world misinformation challenges.

3.3 LLM-Guided Document Annotation

Our work defines a document's role in a RAG system based on its influence on the LLM's decision-making. Unlike prior studies that introduce counterfactual evidence or rely on human annotators, our approach directly evaluates whether a document aids or misleads the LLM in real time. We achieve this by simulating a fact-checking scenario during annotation, treating the LLM as an exam taker.

To generate labels, we simulate the RAG fact-checking process at inference time. Given a claim (Section 3.1) and a retrieved document (Section 3.2), GPT-4 classifies the claim as *true* or *false* based on the document's content. If the classification aligns with the ground-truth, the document is labeled *supporting*; if it contradicts the gold label, it is *misleading*; and if it does not contribute to verification, it is *irrelevant*. By basing labels on the LLM's actual behavior, our approach ensures document annotations reflect their real impact on fact-checking, providing a scalable, empirically grounded alternative to human annotation.

4 BASELINES

4.1 Experimental Setup

Evaluation. We frame fact-checking as a binary classification task where the model must generate a response that aligns with one of the predefined options. Accuracy, calculated with the gold label serving as the reference, is used to evaluate performance. If a model generates an out-of-scope response that does not match any of the given options, it is treated as an incorrect prediction.

Implementation Details. We evaluate RAG systems using both open-source and closed-source LLMs to assess their capabilities for real-world applications. For closed-source benchmarks, we test Google's Gemini 1.5 Flash [14], OpenAI's GPT-4o Mini [28], and Anthropic's Claude 3.5 Sonnet [3] via their respective APIs. For open-source benchmarks, we evaluate Meta's LLama3 8B Instruct [9] and Mistral's Mistral 7B Instruct [18] by running local inference.

In all settings, two-shot examples—one *true* and one *false* claim from RAGUARD training data—are provided in the context. In the Standard RAG setting, we employ OpenAI's text-embedding-ada-002 for semantic search using the original claim as the query and provide the top one and five retrieved documents as context to the LLM. In the Oracle Retrieval setting, our system only takes one document at a time to find the impact of each document to the result. In our prompt, we specify that contextual documents may not be relevant or correct.

4.2 Results

Table 4 displays baseline results on RAGUARD for three tasks using two open and three closed-source LLMs.

Zero-Context Prediction. In the zero-context prediction, RAG systems operate without context documents from the RAGUARD knowledge base. We use this as our baseline (a.k.a, zero-shot baseline). All systems achieve the highest accuracy scores, which is counterintuitive, considering this setting does not benefit from retrieval.

Standard RAG. Performance decreases for all models when integrating retrieval, with all scores falling below the zero-shot baseline. The decline is consistent across both RAG-1 and RAG-5 settings, though the magnitude varies. GPT-4 remains the most robust, exhibiting only a minor drop, while other models, particularly Mistral and Gemini, experience more pronounced declines. Increasing retrieval (RAG-1 to RAG-5) does not consistently improve performance and sometimes worsens it. This suggests that retrieval introduces both useful and misleading information, and when retrieval quality is not optimal, the additional context can confuse rather than help. These findings challenge the assumption that retrieval always benefits downstream performance and reinforce prior research [8, 31, 41, 46] on the risks of noisy retrieval in high-stakes tasks.

Oracle Retrieval. The results in the Oracle Retrieval setting reveal a striking trend: incorporating associated documents from

RAGUARD leads to a significant drop in performance compared to the zero-context baseline. This suggests that models are highly sensitive to misleading or irrelevant information. This trend holds across all models except Llama3, indicating that randomly retrieved *irrelevant* documents are less harmful than *misleading* content. This finding underscores the challenge posed by RAGUARD, which systematically tests model robustness misleading information.

In this task's *All Documents* setting, where models receive all associated documents for a claim, performance generally declines, suggesting that models struggle to reconcile conflicting evidence. The impact is even more severe in the *Misleading-Only* setting, where models are provided only with misleading documents that contradict the claim's ground truth. Most models falling to around 30% accuracy despite the binary nature of the task. This confirms that models are highly susceptible to misleading information and struggle to distinguish factual content from misinformation, highlighting LLM limitations in handling misleading evidence.
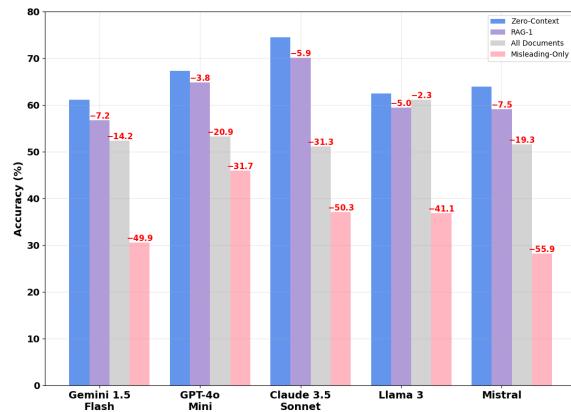

Figure 5: Performance decreases from the Zero-Context baseline to Task 2 and 3 when using RAGuard across various models. Results are measured in Accuracy and include relative percent decreases.
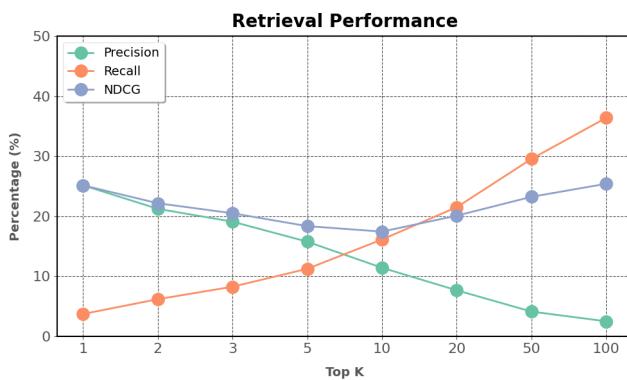

Figure 6: Retrieval Accuracy, Recall, and NDCG at Different Top K Levels

5 DISCUSSION

Comparison of Model Robustness. Figure 5 displays the relative performance decreases across different models. Notably, Claude 3.5 Sonnet, which achieved the highest accuracy in the zero-context baseline, experienced the greatest decreases when exposed to noisy retrieval. In the All Documents Oracle Retrieval setting, Claude suffered one of the steepest declines, and its performance dropped even more drastically in the Misleading-Only condition. This suggests that while Claude performs well in ideal conditions, it is particularly susceptible to misleading evidence, struggling to filter out incorrect information when retrieval introduces contradictory or noisy context. This may also be due to its high baseline performance.

Conversely, GPT-4o Mini demonstrated the highest robustness against misleading evidence. Its relative performance drop in the Misleading-Only setting was 31.7%, significantly lower than the approximate 50% declines observed for Gemini, Claude, and Mistral. This suggests that GPT-4o Mini is better at handling misleading content and maintaining accuracy in noisy retrieval conditions, highlighting differences in model sensitivity to retrieval-induced noise.

Effect of RAG on Accuracy. Figure 5 challenges the common assumption that RAG consistently enhances model performance. When retrieval introduces misleading or irrelevant evidence, it actively degrades accuracy, often leading to worse performance than zero-shot baselines.

This is particularly evident in the Misleading-Only condition from Task 3, where each model is intentionally provided with documents contradicting the claim's ground truth. Across all models, accuracy drops significantly, with an average performance decrease of 45.8%, highlighting the substantial risks posed by retrieval-induced misinformation. These findings emphasize that retrieval, when not carefully controlled, can be detrimental rather than beneficial, further reinforcing the importance of robust filtering and sufficiency evaluation in RAG systems.

Retrieval Performance. Retrieval performance is a standard metric in RAG benchmarks, but our dataset focuses on how models handle misleading or conflicting evidence. High retrieval accuracy alone does not ensure reliable answers due to misleading information in the corpus. Nonetheless, to To provide a full view of system behavior, we report both conventional retrieval metrics and a tailored measurement called Misleading Retrieval Recall.

Figure 6 shows Retrieval Precision, Recall, and Normalized Discounted Cumulative Gain (NDCG) for Task 2 (Standard RAG). Recall naturally rises with K, while precision decreases. NDCG follows a non-monotonic trend, dipping around $K = 10$ before recovering due to relevant items being unevenly distributed across ranked positions, causing reordering as K changes.

We also report Misleading Retrieval Recall—the fraction of claims retrieving at least one misleading document. Task 1 (Zero-Context) scores 0%, while Task 3 (Oracle Retrieval) is 100%. In Task 2, RAG-1 scores 21.3%, increasing to 44.8% for RAG-5, showing a higher risk of retrieving misleading content when retrieving more documents. As seen in Table 4, this correlates with lower overall accuracy.

Qualitative Example. Figure 7 presents example system predictions on RAGUARD, illustrating the impact of misleading documents.

Claim 1: After shedding jobs for more than 10 years, our manufacturers have added about 500,000 jobs over the past three. **(True)**

1a) Model predictions without context
Human:True.
Gemini:True **GPT-4**:True **Sonnet**:True **Llama**:False **Mistral**:True

1b) With overtly misleading context
Context: Bloomberg: **California** Added Only 5,400 Private-Sector Jobs Since 2022.
Human:True
Gemini:False **GPT-4**:False **Sonnet**:False **Llama**: False **Mistral**: False

1c) With partially true misleading context
Context: What is up with these official job reports indicating hundreds of thousands of jobs **added** every month, but then hearing about **layoffs** 24/7, and nobody can get interviews or positions after months being **unemployed**?
Human:True.
Gemini:False **GPT-4**:True **Sonnet**:False **Llama**:False **Mistral**:True

1d) With challenging misleading context
Context: What are your thoughts on The federal government (Bureau of Labor Statistics) announcing today that there were 818,000 **fewer** jobs created through March 2024 than previously reported and that It's the largest **downward** revision in 15 years?
Human:False.
Gemini:False **GPT-4**:False **Sonnet**:False **Llama**:False **Mistral**:False

Claim 2: Hillary Clinton says she wants to 'raise taxes on the middle class.' **(False)**

2a) Model predictions without context.
Human:False. **GPT-4**:False

2b) With non-associated context.
Context: **Harris** maintains Biden's pledge not to raise taxes on the middle class.
Human:False. **GPT-4**:False

2c) With 5 non-associated contextual documents.
Context: Doc 1. **Harris** maintains Biden's pledge not to raise taxes on middle class and lower class… Doc 3. Chris Christie says **Bernie Sanders**'s plan is 'to raise your taxes… Doc 4. **Biden** to call in State of the Union for business tax hikes, middle class tax cuts… Doc 5. Joe **Biden** to propose big tax rises for billion…
Human:False. **GPT-4**:False

2d) With misleading context.
Context: Which tax plan is better for middle-class voters, Clinton's or Trump's plan? I compared Hillary Clinton and Donald Trump's tax plan to see how it would affect me as a middle class tax payer. Here's the breakdown…Donald Trump's tax plan would put an **extra $300.00 a paycheck** in my pocket every paycheck.
Human:False. **GPT-4**:True

Claim 3: Joe Biden brought Republicans and Democrats together to pass the 1994 crime bill, putting 100,000 cops on the streets and starting an eight-year drop in crime across the country. **(True)**

3a) Model predictions without context.
Human:True. **GPT-4**:False

3b) With non-associated context.
Context:
Statement from President Joe Biden on **Record Decrease** in Violent Crime in 2024
Human:True. **GPT-4**:True

3c) With 4 non-associated documents & 1 misleading document.
Context: Doc 1. Statement from President Joe Biden on Record **Decrease** Doc 2. We are experiencing the largest-ever year-over-year **decline** in homicides. Doc 3. Violent crime is **falling** just as rapidly. Doc 4. September 13, 1994. President Clinton signs into law a controversial $30 billion anti-crime bill, accused of causing **mass incarceration**. Doc 5. …
Human:True. **GPT-4**:False

3d) With misleading context.
Context:
September 13, 1994. President Clinton signs into law a controversial $30 billion anti-crime bill, accused of causing **mass incarceration**.
Human:True. **GPT-4**:False

Figure 7: Example predictions on RAGUARD, compared to the expected human response. Note that each column compares different prediction scenarios based on varying retrieved contexts for the same claim rather than a multi-turn process. *Left:* Each system's classification of a true claim with three progressively misleading documents. *Middle:* GPT-4-based system's classification of a false claim with one noisy non-associated document, many noisy non-associated documents, and a misleading document. *Right:* GPT-4-based system's classification of a true claim with a supporting non-associated document, one misleading document along with other supporting non-associated documents, and a misleading document.

The left example highlights how misleading documents negatively affect the classification of a true claim. While misleading documents generally degrade system performance compared to zero-shot predictions, their specific influence varies based on their complexity. We distinguish three categories of misleading documents:

(1) Overtly Misleading Document: This category includes documents that are evidently misleading to humans but still lead to incorrect predictions by all RAG systems. For example, in Figure 7, the document falsely comparing California's job growth to the national average misleads all systems (1b), despite their correct zero-shot predictions (1a). This suggests a form of selective bias, where the systems prioritize the provided information simply because it is included in the prompt, even though the instructions explicitly caution against assuming its correctness.

(2) Partially True Misleading Document: These documents contain partial truths, making it necessary to apply reasoning to recognize their misleading nature. For example, as shown in Figure 7, one document criticizes unemployment but also states that "official job reports are reporting jobs added" (1c). While this statement supports the claim that 500,000 jobs were added, the document's overall tone suggests rising unemployment. However, this suggestion is more of an opinion than a fact. Some LLMs, such as GPT-4 and Mistral, were able to reason through this contradiction and classify the claim correctly.

(3) Challenging Misleading Document: These documents present significant challenges, even for human annotators. For example, a claim referencing job growth in the 2000s is incorrectly classified because the RAG system retrieves data from 2024, which accurately reports lower job creation (1d). The

temporal misalignment in retrieved documents presents a fundamental challenge in this dataset and task.

The middle example demonstrates GPT-4's ability to filter out noise from retrieved documents that are not associated with the claim but could be considered misleading documents in our dataset (e.g., documents using the same phrasing but referring to different individuals, such as "Harris" instead of "Clinton" in example 2b). Even when five irrelevant documents are retrieved (2c), GPT-4 remains robust. However, when presented with a misleading document from the dataset (2d), GPT-4 fails, reinforcing the dataset's effectiveness in challenging model performance beyond conventional RAG noise. This further explains the lower accuracy observed in the Oracle Retrieval setting in our baseline experiments.

The right example shows how GPT-4 tends to assign disproportionate weight to misleading documents, allowing them to override even non-associated supporting evidence. In the example, a non-associated document that contains supporting information (3b) enables GPT-4 to correct its initially incorrect zero-shot prediction (3a). However, when a misleading document is retrieved alongside other non-associated supporting documents (3c), the system incorrectly classifies the claim, similar to its behavior when only the misleading document is retrieved (3d). This demonstrates that misleading documents can have a stronger influence on the model's classification, regardless of the presence of supporting evidence, highlighting a significant vulnerability in RAG systems.

These examples highlight three key findings: (1) LLMs remain highly susceptible to misleading documents, even when their content is transparently incorrect, (2) misleading documents retrieved from the dataset exert a stronger influence than non-associated documents retrieved erroneously, and (3) when misleading documents are present, they can significantly outweigh supporting

evidence, leading to incorrect predictions. These findings emphasize the strength and uniqueness of our dataset in evaluating and challenging RAG-based model performance.

6 RELATED WORK

Retrieval-Augmented Generation with Noisy Contexts. Retrieval-Augmented Language Models (RALMs) have demonstrated strong performance across various NLP tasks [16, 24]. However, their effectiveness is constrained by the retriever's ability to find supporting information. In real-world applications, retrieval often introduces irrelevant or misleading content, which can significantly degrade model performance [8, 31, 41, 46]. Prior work has identified two primary effects of such noise.

The first is the impact of irrelevant documents on RAG performance [7]. To make RAG systems more robust to this issue, researchers have explored several strategies, including prompting the language model to generate a rationale connecting retrieved documents to the query [38], employing multi-agent debate systems to identify the most relevant information [36], and aggregating multiple documents to produce a more reliable final response [40].

The second challenge arises when retrieved documents conflict with an LLM's internal knowledge [21]. To address this, AstuteRAG introduced an iterative system that consolidates internal and external knowledge, reducing inconsistencies in generated responses [35]. However, this study focuses on disagreements between documents in existing datasets, which do not intentionally mislead RAG systems and do not represent conflicting information in the real world.

Researchers have increasingly attempted to improve RAG robustness through developing datasets that expose the model to conflicting contexts [8, 25] and retrieval noises [39] that may later be used for adversarial training [11]. However, existing datasets fail to fully capture the complexities of real-world misinformation.

In contrast, RAGUARD is the first to capture *misleading* context that reflects real-world ambiguities, polarized opinions, and partial truths. Unlike prior datasets that primarily focus on synthetic conflicts or document-model disagreements, RAGUARD captures naturally occurring misinformation, making it a more realistic benchmark for evaluating RAG robustness.

The Limitations of Open-Book QA Datasets. RAG is studied primarily through open-book question answering (QA), where models answer questions based on retrieved knowledge [13]. However, most Open-Book QA datasets carefully curate their documents, avoiding noisy information [5, 10, 19, 23, 26, 30, 44, 47]. This leads to strong performance in controlled settings, but poor generalization in real-world scenarios, where conflicting documents often degrade performance [8].

Some datasets attempt to address this issue by synthetically introducing counterfactual information [11, 39]. However, synthetic noise may not fully capture the complexities of real-world misinformation. Others rely on human annotators to identify conflicting documents [25], but this approach is costly and difficult to scale. [8] classifies retrieved documents as distracting based on their equivalence to a gold-standard document, but this approach may not reflect the existence of truly deceptive or contradictory information.

In contrast to these approaches, we leverage real-world misinformation by using political fact-checking data, which contains misleading and conflicting information. This allows us to construct a dataset that better reflects the challenges RAG systems face.

Fact-Checking and RAG. Fact-checking is a well-studied task with datasets sourced from platforms like Twitter [27], Wikipedia [34], and PolitiFact [1, 29, 32, 37, 45]. Given that fact-checking often relies on external evidence, it aligns well with RAG, where retrieval can support the verification or negation of a claim. However, existing fact-checking datasets typically use gold-standard evidence from the same source as the verdict [4, 17, 22, 34, 45], meaning there is no exposure to noisy or contradictory evidence. Despite the inherently polarizing nature of online and political discourse, the retrieved evidence in these datasets rarely contradicts the final verdict. Our work introduces noisy and conflicting information into fact-checking datasets, ensuring that retrieved evidence is not always aligned with the verdict.

7 CONCLUSION

In this paper, we introduce RAGUARD, a challenging and diverse fact-checking dataset designed to assess the robustness of RAG systems against misleading retrievals. RAGUARD comprises 2,648 claims—1,333 true and 1,315 false—paired with 16,331 documents, averaging 6.2 documents per claim. These documents are labeled using a novel LLM-guided approach that simulates an exam-like evaluation, analyzing how the model processes and interprets retrieved evidence at inference time to determine whether the documents support, mislead, or are irrelevant to the claim. Unlike prior RAG benchmarks that rely on synthetically noisy data or curated gold-standard documents, RAGUARD utilizes real-world evidence, even in cases where no gold-standard documents exist. By incorporating naturally occurring misleading data from Reddit discussions alongside verified evidence and claims from PolitiFact, it mirrors the complexities of real-world misinformation, which is necessary for more robust systems.

Our findings show that the performance of current RAG systems deteriorates significantly when exposed to misleading evidence, challenging the assumption that retrieval always enhances model accuracy. These results highlight the need for more resilient fact-checking pipelines. Future research should focus on enhancing retrieval robustness through methods such as adversarial retrieval training, which exposes models to misleading evidence during training to improve resilience, and uncertainty-aware retrieval, which prioritizes evidence credibility over mere relevance. Additionally, fact-verification mechanisms that incorporate multi-step reasoning and cross-document consistency checks can mitigate the impact of misleading sources, while confidence calibration techniques may further refine the model's ability to discern factual inconsistencies.

By providing a challenging yet realistic benchmark, RAGUARD encourages the development of more sophisticated retrieval-based fact-checking methodologies. We hope this dataset will facilitate progress in designing retrieval pipelines that are not only effective but also resistant to misinformation, ultimately contributing to more reliable and trustworthy AI systems.

REFERENCES

[1] Tariq Alhindi, Savvas Petridis, and Smaranda Muresan. 2018. Where is Your Evidence: Improving Fact-checking by Justification Modeling. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal (Eds.). Association for Computational Linguistics, Brussels, Belgium, 85–90. https://doi.org/10.18653/v1/W18-5513

[2] Rami Aly, Zhijiang Guo, Michael Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. FEVER-OUS: Fact Extraction and VERification Over Unstructured and Structured information. arXiv:2106.05707 [cs.CL] https://arxiv.org/abs/2106.05707

[3] Anthropic. 2024. Claude 3.5 sonnet. https://www.anthropic.com/news/claude-3-5-sonnet

[4] Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. 2019. MultiFC: A Real-World Multi-Domain Dataset for Evidence-Based Fact Checking of Claims. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, Hong Kong, China, 4685–4697. https://doi.org/10.18653/v1/D19-1475

[5] Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic Parsing on Freebase from Question-Answer Pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, David Yarowsky, Timothy Baldwin, Anna Korhonen, Karen Livescu, and Steven Bethard (Eds.). Association for Computational Linguistics, Seattle, Washington, USA, 1533–1544. https://aclanthology.org/D13-1160

[6] Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. 2022. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*. PMLR, 2206–2240.

[7] Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2023. Benchmarking Large Language Models in Retrieval-Augmented Generation. arXiv:2309.01431 [cs.CL] https://arxiv.org/abs/2309.01431

[8] Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. 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 (SIGIR 2024)*. ACM, 719–729. https://doi.org/10.1145/3626772.3657834

[9] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* (2024).

[10] Matthew Dunn, Levent Sagun, Mike Higgins, V Ugur Guney, Volkan Cirik, and Kyunghyun Cho. 2017. Searchqa: A new q&a dataset augmented with context from a search engine. *arXiv preprint arXiv:1704.05179* (2017).

[11] 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. arXiv:2405.20978 [cs.AI] https://arxiv.org/abs/2405.20978

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

[13] 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. arXiv:2312.10997 [cs.CL] https://arxiv.org/abs/2312.10997

[14] Gemini. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530* (2024).

[15] Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, et al. 2024. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in Neural Information Processing Systems* 36 (2024).

[16] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval Augmented Language Model Pre-Training. In *Proceedings of the 37th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 119)*, Hal Daumé III and Aarti Singh (Eds.). PMLR, 3929–3938. https://proceedings.mlr.press/v119/guu20a.html

[17] Andreas Hanselowski, Christian Stab, Claudia Schulz, Zile Li, and Iryna Gurevych. 2019. A Richly Annotated Corpus for Different Tasks in Automated Fact-Checking. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, Mohit Bansal and Aline Villavicencio (Eds.). Association for Computational Linguistics, Hong Kong, China, 493–503. https://doi.org/10.18653/v1/K19-1046

[18] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. arXiv:2310.06825 [cs.CL] https://arxiv.org/abs/2310.06825

[19] Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. arXiv:1705.03551 [cs.CL] https://arxiv.org/abs/1705.03551

[20] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 6769–6781. https://doi.org/10.18653/v1/2020.emnlp-main.550

[21] Evgenii Kortukov, Alexander Rubinstein, Elisa Nguyen, and Seong Joon Oh. 2024. Studying Large Language Model Behaviors Under Context-Memory Conflicts With Real Documents. arXiv:2404.16032 [cs.LG] https://arxiv.org/abs/2404.16032

[22] Neema Kotonya and Francesca Toni. 2020. Explainable Automated Fact-Checking for Public Health Claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 7740–7754. https://doi.org/10.18653/v1/2020.emnlp-main.623

[23] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural Questions: A Benchmark for Question Answering Research. *Transactions of the Association for Computational Linguistics* 7 (2019), 452–466. https://doi.org/10.1162/tacl_a_00276

[24] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. arXiv:2005.11401 [cs.CL] https://arxiv.org/abs/2005.11401

[25] Siyi Liu, Qiang Ning, Kishaloy Halder, Wei Xiao, Zheng Qi, Phu Mon Htut, Yi Zhang, Neha Anna John, Bonan Min, Yassine Benajiba, and Dan Roth. 2024. Open Domain Question Answering with Conflicting Contexts. arXiv:2410.12311 [cs.CL] https://arxiv.org/abs/2410.12311

[26] Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering. arXiv:2209.09513 [cs.CL] https://arxiv.org/abs/2209.09513

[27] Dan S Nielsen and Ryan McConville. 2022. Mumin: A large-scale multilingual multimodal fact-checked misinformation social network dataset. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*. 3141–3153.

[28] OpenAI. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276* (2024).

[29] Nele Põldvere, Md. Zia Uddin, and Aleena Thomas. 2023. The PolitiFact-Oslo Corpus: A New Dataset for Fake News Analysis and Detection. *Inf.* 14 (2023), 627. https://api.semanticscholar.org/CorpusID:265420523

[30] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. arXiv:1606.05250 [cs.CL] https://arxiv.org/abs/1606.05250

[31] Artsiom Sauchuk, James Thorne, Alon Halevy, Nicola Tonellotto, and Fabrizio Silvestri. 2022. On the Role of Relevance in Natural Language Processing Tasks. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Madrid, Spain) *(SIGIR '22)*. Association for Computing Machinery, New York, NY, USA, 1785–1789. https://doi.org/10.1145/3477495.3532034

[32] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2019. FakeNewsNet: A Data Repository with News Content, Social Context and Spatialtemporal Information for Studying Fake News on Social Media. arXiv:1809.01286 [cs.SI] https://arxiv.org/abs/1809.01286

[33] James Thorne and Andreas Vlachos. 2018. Automated Fact Checking: Task Formulations, Methods and Future Directions. In *Proceedings of the 27th International Conference on Computational Linguistics*, Emily M. Bender, Leon Derczynski, and Pierre Isabelle (Eds.). Association for Computational Linguistics, Santa Fe, New Mexico, USA, 3346–3359. https://aclanthology.org/C18-1283/

[34] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a Large-scale Dataset for Fact Extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, Marilyn Walker, Heng Ji, and Amanda Stent (Eds.). Association for Computational Linguistics, New Orleans, Louisiana, 809–819. https://doi.org/10.18653/v1/N18-1074

[35] Fei Wang, Xingchen Wan, Ruoxi Sun, Jiefeng Chen, and Sercan Ö. Arık. 2024. Astute RAG: Overcoming Imperfect Retrieval Augmentation and Knowledge Conflicts for Large Language Models. arXiv:2410.07176 [cs.CL] https://arxiv.org/abs/2410.07176

[36] Haotian Wang, Xiyuan Du, Weijiang Yu, Qianglong Chen, Kun Zhu, Zheng Chu, Lian Yan, and Yi Guan. 2024. Learning to Break: Knowledge-Enhanced Reasoning in Multi-Agent Debate System. arXiv:2312.04854 [cs.CL] https://arxiv.org/abs/2312.04854

[37] William Yang Wang. 2017. "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Regina Barzilay and Min-Yen Kan (Eds.). Association for Computational Linguistics, Vancouver, Canada, 422–426. https://doi.org/10.18653/v1/P17-2067

[38] Zhepei Wei, Wei-Lin Chen, and Yu Meng. 2024. InstructRAG: Instructing Retrieval-Augmented Generation via Self-Synthesized Rationales. arXiv:2406.13629 [cs.CL] https://arxiv.org/abs/2406.13629

[39] Jinyang Wu, Feihu Che, Chuyuan Zhang, Jianhua Tao, Shuai Zhang, and Pengpeng Shao. 2024. Pandora's Box or Aladdin's Lamp: A Comprehensive Analysis Revealing the Role of RAG Noise in Large Language Models. arXiv:2408.13533 [cs.CL] https://arxiv.org/abs/2408.13533

[40] Chong Xiang, Tong Wu, Zexuan Zhong, David Wagner, Danqi Chen, and Prateek Mittal. 2024. Certifiably Robust RAG against Retrieval Corruption. arXiv:2405.15556 [cs.LG] https://arxiv.org/abs/2405.15556

[41] 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. arXiv:2305.13300 [cs.CL] https://arxiv.org/abs/2305.13300

[42] Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking Retrieval-Augmented Generation for Medicine. arXiv:2402.13178 [cs.CL] https://arxiv.org/abs/2402.13178

[43] Diji Yang, Jinmeng Rao, Kezhen Chen, Xiaoyuan Guo, Yawen Zhang, Jie Yang, and Yi Zhang. 2024. Im-rag: Multi-round retrieval-augmented generation through learning inner monologues. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 730–740.

[44] 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. arXiv:1809.09600 [cs.CL] https://arxiv.org/abs/1809.09600

[45] Barry Menglong Yao, Aditya Shah, Lichao Sun, Jin-Hee Cho, and Lifu Huang. 2023. End-to-End Multimodal Fact-Checking and Explanation Generation: A Challenging Dataset and Models. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23)*. ACM, 2733–2743. https://doi.org/10.1145/3539618.3591879

[46] Xunjian Yin, Baizhou Huang, and Xiaojun Wan. 2023. ALCUNA: Large Language Models Meet New Knowledge. arXiv:2310.14820 [cs.CL] https://arxiv.org/abs/2310.14820

[47] Wenhao Yu, Meng Jiang, Peter Clark, and Ashish Sabharwal. 2023. IfQA: A Dataset for Open-domain Question Answering under Counterfactual Presuppositions. arXiv:2305.14010 [cs.CL] https://arxiv.org/abs/2305.14010