ELOQ: Resources for Enhancing LLM Detection of Out-of-Scope Questions

Zhiyuan Peng*
Santa Clara University
Santa Clara, CA, USA
zpeng@scu.edu

Jinming Nian*
Santa Clara University
Santa Clara, CA, USA
jnian@scu.edu

Alexandre Evfimievski†
Adobe Inc.
San Jose, CA, USA
aevfim@outlook.com

Yi Fang
Santa Clara University
Santa Clara, CA, USA
yfang@scu.edu

Abstract

Retrieval-augmented generation (RAG) has become integral to large language models (LLMs), particularly for conversational AI systems where user questions may reference knowledge beyond the LLMs' training cutoff. However, many natural user questions lack well-defined answers, either due to limited domain knowledge or because the retrieval system returns documents that are relevant in appearance but uninformative in content. In such cases, LLMs often produce hallucinated answers without flagging them. While recent work has largely focused on questions with false premises, we study out-of-scope questions, where the retrieved document appears semantically similar to the question but lacks the necessary information to answer it. In this paper, we propose a guided hallucination-based approach ELOQ [1], for automatically generating a diverse set of out-of-scope questions from post-cutoff documents, followed by human verification to ensure quality. We use this dataset to evaluate several LLMs on their ability to detect out-of-scope questions and generate appropriate responses. Finally, we introduce an improved detection method that enhances the reliability of LLM-based question-answering systems in handling out-of-scope questions.

CCS Concepts

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

Keywords

Large Language Models, Question Answering, Out-of-Scope Question, Retrieval Augmented Generation

ACM Reference Format:
Zhiyuan Peng, Jinming Nian, Alexandre Evfimievski, and Yi Fang. 2025. ELOQ: Resources for Enhancing LLM Detection of Out-of-Scope Questions. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25), July 13–18, 2025, Padua, Italy*. ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3726302.3730333

1 Introduction

Retrieval-augmented generation (RAG) has become a standard approach to building context-grounded conversational AI agents [5, 6, 15]. Given an inquiry, a RAG system retrieves a relevant document from a curated knowledge base and presents it to a large language model (LLM), which generates a response. RAG performance is evaluated on dimensions such as response accuracy, faithfulness, completeness, answer-context relevance, and style [4, 18, 27, 38]. Errors may stem from a document-to-question mismatch, hallucinated response, or misunderstanding of the document or the question.

Many user questions have no direct answer: some rely on a false premise, others are ambiguous, and some are out-of-scope for the knowledge base. Prior studies indicate that about 25% of natural questions contain false assumptions [35] and over 50% are ambiguous [20]. Additionally, out-of-scope errors have been identified as the leading cause of enterprise AI assistants generating responses that appear convincing but are incorrect [19]. A well-designed RAG system should implement an exception handling mechanism that detects confusing questions and responds by clarifying the confusion, retrieving additional documents, or seeking assistance.

Enabling "deconfusion" in a RAG system involves three challenges: (1) creating a diverse dataset of confusing questions, (2) training a classifier to detect confusion, and (3) developing a response generator for each confusion type. As shown in Table 1, prior studies create confusing questions with different categories, such as multi-answer [20], false premise [7, 35, 36, 39], and cunning texts [16]. For instance, "How many sacks does clay matthews have in his career?" is a multi-answer question as "clay matthews" may refer to "Clay Matthews Jr." or "Clay Matthews III", "Where in Liverpool did John Lennon die?" is a false premise question as John Lennon died in New York city and "What should I do if I forget which ATM machine I deposited my money in?" is a cunning text. In contrast to these questions, out-of-scope questions look relevant to the document, but there is no answer within the document and

*Both authors contributed equally to this research.
†The author contributed to this work before he joined Adobe Inc.
[1] https://github.com/zhiyuanpeng/ELOQ.git

SIGIR '25, Padua, Italy

ACM ISBN 979-8-4007-1592-1/2025/07
https://doi.org/10.1145/3726302.3730333

Dataset	Source	Category	Cutoff	Annotation Method
SQuAD 2.0 [25]	Wiki	Unanswerable	✗	Human
SQuAD2-CR [13]	Wiki	Unanswerable	✗	Human
BoolQ$_{3L}$ [28]	Wiki	Unanswerable	✗	Algoritm+Human
CAsT-answerability [12]	Wiki	Unanswerable	✗	Algoritm+Human
AMBIGQA [20]	Wiki	Multi-answers	✗	Human
CREPE [35]	Reddit	False Premise	✗	Human
FalseQA [7]	–	False Premise	✗	Human
KG-FPQ [39]	Wiki	False Premise	✗	LLM
FAITH [36]	Wiki	False Premise	✗	LLM
FLUB [16]	Baidu Tieba	Cunning Texts	✗	Human
ELOQ	News	Out-of-scope	✓	LLM+Human

Table 1: ELOQ vs Existing Datasets. "Cutoff" represents whether the data is beyond the LLMs' knowledge or not.

thus can mislead LLMs into generating hallucinated answers that seem correct but are actually incorrect, representing the most dangerous type of error [19]. For instance, "How does Justin Jefferson's exceptional catching ability influence the Vikings' decision to select him in the first round of the 2020 NFL Draft despite his lack of speed?" is an out-of-scope question as it asks about "Vikings' decision to select Jefferson in the first round of the 2020 NFL Draft" which was never discussed in the paired news (Appendix D). Our work focuses on mitigating this type of confusion.

Existing studies [12, 13, 25, 28] on unanswerable questions, of which out-of-scope questions are a subset, have typically created the dataset either through manual annotation by trained human annotators [7, 13] or via simple automatic generation algorithms, as seen in BoolQ$_{3L}$ [28] and CAsT-answerability [12]. These methods often lack a balance between efficiency and quality, as human-annotated questions are of high quality but low efficiency, and automatically generated questions are of high efficiency but too simple to be distinguished to some degree. Recently, LLMs have demonstrated strong generation and following instruction abilities, and thus KG-FPQ [39] and FAITH [36] utilize LLMs to generate high-quality false premise questions, mitigating the balance of efficiency and quality issues. However, all these datasets are usually constrained to knowledge before the LLMs' training cutoff date, which limits their effectiveness in evaluating real-world RAG systems that must retrieve up-to-date information likely beyond the model's training data to answer user questions.

To address these challenges above, we propose a framework for automatically generating out-of-scope questions and evaluate its effectiveness in training classifiers to detect out-of-scope situations. While questions in existing datasets such as MS MARCO [23] may serve as out-of-scope questions for the corresponding hard negative passages, they fail beyond the LLM's knowledge cutoff. ELOQ fills this gap by providing a diverse benchmark for out-of-scope questions, leveraging LLM-assisted generation to reduce human annotation effort. Our contributions are as follows:

- We propose a framework for generating synthetic out-of-scope questions from a given corpus.
- We demonstrate the utility of our synthetic data in training out-of-scope question detectors.
- We conduct a comparative evaluation of multiple LLMs as RAG agents to measure their out-of-scope question detection accuracy.

2 Related Work

2.1 Benchmark Datasets

Recent work on LLM responses to confusing questions primarily focuses on stand-alone questions answered using the LLM's pre-training knowledge, without a supporting context document. These questions may be naturally collected from online sources [17, 35], written by human annotators [7], or synthetically generated [36, 39]. Earlier research on ambiguous open-domain questions introduced AmbigQA [21], a dataset of naturally ambiguous questions. More relevant benchmarks are unanswerable questions [12, 13, 25, 28] of which out-of-scope questions are a subset. SQuAD 2.0 [25] is created by asking crowdworkers to write unanswerable questions similar to the answerable question given a paragraph. SQuAD2-CR [13] tags SQuAD 2.0 [25] with answerable reasons. CAsT-answerability [12] extends CAsT-snippets [40] by including five randomly selected non-relevant passages to each question. BoolQ$_{3L}$ [28] matches each selected "Yes" or "No" question of BoolQ [2] to a passage within BoolQ that has the greatest overlap with the questions in terms of nouns and verbs. These unanswerable datasets are sourced from Wikipedia, which is typically included in LLM's training data. Our dataset is derived from news articles, allowing us to easily select content published after LLM's knowledge cutoff dates. Table 1 presents a comparison of these benchmark datasets.

2.2 Question Generation

Zhu et al. [39] and Yuan et al. [36] describe a method to synthetically generate stand-alone false premise questions by selecting a set of factual triples of (subject, relation, object) using Wikipedia data, replacing the object with a similar but incorrect one, and then filling in a template to generate a false premise question. Content-grounded (non-confusing) synthetic data generation has been studied in [14, 30, 34]. Given the grounding passages, these methods generate data entries using few-shot prompting with task-specific examples, followed by filtering to ensure data quality, faithfulness, and diversity.

2.3 Mitigation

Mitigation is commonly achieved through in-context learning [17, 39], chain-of-thought (CoT) [31] and fine-tuning. Experiments show that fine-tuning outperforms both in-context learning and CoT reasoning [7]. CoT reasoning leads to inconsistent improvements, whereas in-context learning improves performance as the number of examples increases [17]. Yuan et al. [36] identifies approximately 1% of attention heads as responsible for confusion caused by false-premise questions and demonstrates a 20% performance gain by constraining these heads. Kim et al. [10] retrieves a Wikipedia page for each question and responds with the identified unverifiable premises if the page does not entail them. Sulem et al. [28] trains a classification model based on BERT to identify unanswerable questions.

3 Methods

We developed the ELOQ data generator to evaluate and mitigate LLM confusion when responding to out-of-scope questions for a given document. We assume that the user has limited domain

Algorithm 1 The guided hallucination method

```
1: Call $\text{LLM}_q$ to extract from $d$ a list of $n$ claims $c_1, \ldots, c_n$; batching claims reduces $\text{LLM}_q$-calls (Appendix B.1)
2: Partition $\{c_1, \ldots, c_n\}$ into disjoint subsets $S_1, \ldots, S_k$, e.g. $S_j = \{c_i : i \bmod 3 = j - 1\}$ for $j = 1, 2, 3$
3: for several rounds (e.g. 3), and in each round, for all $S_j$ in partition $(S_1, \ldots, S_k)$ do
4:     In the claims list $c_1, \ldots, c_n$ replace all claims $c_i$ where $i \in S_j$ by text "(missing)"
5:     Provide $\text{LLM}_q$ with this modified list (but not $d$) and prompt it to recover the missing claims (Appendix B.2)
6:     In the claims list $c_1, \ldots, c_n$, replace the missing claims with the claims recovered in the above step.
7: end for
8: Call $\text{LLM}_q$ to remove all claims supported by $d$ or by the original claims, leaving only out-of-scope claims (Appendix B.11)
9: Call $\text{LLM}_q$ to generate one short question per each novel claim, focusing on one key element of it (Appendix B.10)
10: Call $\text{LLM}_q$ to filter out any questions answerable in the context of $d$ (Appendix B.3)
```

knowledge, resulting in asking questions that appear relevant but are actually out of scope and cannot be answered based on the document's content. We also assume that, due to the limitations of the retriever itself, relevant documents without an answer to the questions are possibly retrieved and ranked at the top. The best answer to this kind of question is to refuse to answer it, rather than providing a fabricated response. In this step, given a standard prompt p, an out-of-scope question q, and a relevant document d, the LLM_r generates a response r. Ideally, r should clarify the confusing part of the question – a process we refer to as defusion (or de-confusion) – rather than attempting to answer the question directly and risking hallucination.

3.1 Data Collection

We focus on the scenario where the domain knowledge is maintained separately from the LLM in a document database. Hence, we prefer documents that are novel to the evaluated LLMs, such as news articles published after all pretraining cutoff dates, ensuring that LLMs cannot reproduce claims or answer questions correctly by pure parametric knowledge. Instead, they must generate responses grounded in their understanding of the news content. Therefore, using the Newscatcher [1], we collected 200 news articles published between Jan 1, 2024, and September 30, 2024, for each topic[2]. We require each document to be concise enough to fit within the LLM prompt, along with instructions and examples, but sufficiently long to make at least 4 to 10 separate claims (i.e., a few paragraphs in length). Thus, we select news articles with more than 150 words and, for each, sequentially extract sentences from the beginning until the word count exceeds 300. Finally, we collected 2,000 news and sampled a subset for human annotation. We represent the human-annotated data as Gold and the remaining data as Silver. Table 2 presents the statistics of the dataset. Our ELOQ dataset [3] is publicly available.

Attribute	Gold	Silver
# of documents	45	1,955
# of in-scope questions	103	8,949
# of out-of-scope questions	113	10,105
average words per document	280	276
average words per in-scope question	16	16
average words per out-of-scope question	16	16

Table 2: Statistics of ELOQ. Gold is sampled from crawled 2,000 news and annotated by annotators (Section 3.4).

3.2 Question Generation

We introduce our guided hallucination method, designed to generate out-of-scope questions from a given document. The process consists of three main steps: claim extraction, hallucination injection, and question generation.

3.2.1 *Claim Extraction.*

The first step involves extracting claims from the document d to serve as the basis for generating out-of-scope questions. Since our goal is to generate questions that appear related to d but are actually unanswerable using its content, we start by obtaining a structured set of claims from the document. To achieve this, we use a LLM denoted as LLM_q, to extract a set of n claims $c_1, c_2, \ldots, c_n$ from d. These claims represent key factual statements or assertions made in the document. We extract the claims in batches to minimize the number of LLM calls and improve computational efficiency. Once the claims are extracted, we partition them into disjoint subsets $S_1, S_2, \ldots, S_k$. A simple partitioning scheme is used, such as:

$$S_j = \{c_i \mid i \mod 3 = j - 1\}, \quad \text{for } j = 1, 2, 3.$$

This ensures that claims are distributed evenly among the subsets, facilitating an efficient iterative process in the next step. Partitioning the claims allows us to selectively manipulate certain claims while preserving the overall document structure.

3.2.2 *Hallucination Injection.*

The second step is to generate hallucinated claims that are similar to the factual claims generated in Section 3.2.1 but contain altered or missing information. The process follows an iterative masking and recovery approach over multiple rounds (e.g. 3). For each round, we follow steps 3 to 6 in Algorithm 1.

In practice, we find that even after repeating the hallucination injection process multiple times, some claims remain too general and can still be supported by the document. Therefore, we design a prompt-based filter B.11 to discard these in-scope claims that are explicitly supported by d. This ensures that only truly out-of-scope claims remain in our dataset.

3.2.3 *Question Generation.*

The final step converts the hallucinated claims into out-of-scope questions while ensuring they cannot be answered using the document. Specifically, for each hallucinated claim, LLM_q generates a concise and precise question focusing on a key aspect of the claim. The question is designed to appear related to the document while being unanswerable based on its content. Once the questions are generated, we further filter them by ensuring they are truly unanswerable. This is achieved by prompting LLM_q

[2] sport, business, science, food, politics, travel, entertainment, music, news, tech

[3] https://huggingface.co/datasets/zhiyuanpeng/ELOQ

Algorithm 2 Evaluation steps

```
1: Call LLM$_r$ on $(d, q)$ with RAG prompt and get its response $r$ (Appendix B.6, B.7, B.8)
2: Call LLM$_r$ $m$ times on $(d, q)$ to check and explain if $q$ is out-of-scope given $d$ (Appendix B.3)
3: Aggregate $m$ predictions by majority vote
4: if $q$ is generated as out-of-scope then
5:     Call LLM$_q$ $m$ times on $(d, q, r)$ to check and explain if $r$ defused the confusion (Appendix B.4)
6:     Aggregate $m$ predictions by majority vote
7: end if
```

to verify whether the generated questions can be answered using d. Any question that remains answerable in the context of d is discarded.

In addition to generating out-of-scope questions, we also prompt LLM$_q$ to generate in-scope questions for comparison. The complete question generation process is outlined in Algorithm 1, and the associated prompts are in Appendix B.

3.3 Evaluation

We first assessed the quality of ELOQ by sampling a subset of questions (ELOQ-Gold) and manually verifying whether they were truly out-of-scope (Section 3.4). Next, we evaluated how well various LLM$_r$ models could defuse the out-of-scope questions in ELOQ. Given the large number of out-of-scope questions, manual verification of LLM$_r$ responses is not feasible. Therefore, we proposed AutoDefuseEval using GPT-4o-mini to automatically detect whether a response successfully defuses the question. To validate AutoDefuseEval, we compared its performance with human annotations (ELOQ-Gold) on a sampled dataset (Section 3.4). Formally, we define our tasks as follows:

Question generation: What is the quality of out-of-scope and in-scope questions generated by Algorithm 1? (Section 3.4)

Defusion detection: How accurately can AutoDefuseEval detect whether an LLM$_r$ response to a out-of-scope question defuses the confusion? (Section 3.4)

Out-of-scope response: How often does an LLM$_r$ successfully defuse an out-of-scope question using different prompting methods? (Section 4.1)

Out-of-scope detection: How accurately can LLMs identify which context-grounded questions are out-of-scope and require special handling? (Section 4.2)

Algorithm 2 has the steps we run for evaluating the above tasks. We run steps 1, 2, and 3 for "out-of-scope detection" and 1, 4, 5, 6, and 7 for "out-of-scope response". Following the self-consistency method [29], we perform multiple LLM calls in steps 2 and 5 of Algorithm 2, taking the majority vote to determine the final label.

3.4 Human Annotation

Following the methodology from [3], we sampled 216 questions, evenly divided into three groups, ensuring that each question was annotated by two different annotators. As shown in Table 3, Cohen's Kappa values exceeded 0.75 for confusion labels (where "Yes" indicates a question is out-of-scope and "No" indicates otherwise), and surpassed 0.81 for defusion labels (where "Yes" indicates the LLM$_r$'s response defuses the question, and "No" indicates it does not), signifying substantial to near-perfect inter-annotator agreement and annotation consistency. Two additional annotators manually resolved all disagreements to establish ground truth labels. All annotators were computer science graduate students recruited via email and incentivized with complimentary meals. They were trained by first reviewing the annotation guidelines and then labeling a small dataset previously annotated by the authors. The annotation guidelines are released along with the data and code.

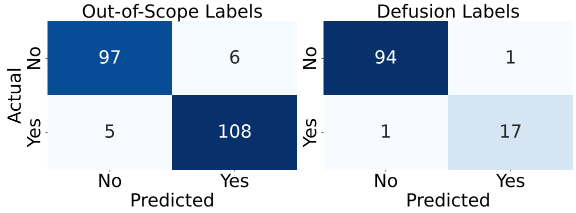

Figure 1: Confusion matrix of out-of-scope and defusion on ELOQ-Gold.

We used two metrics to assess our method: "Annotator Acc" and "Group's Agree Acc." "Annotator Acc" treats each annotator's labels as ground truth, computing the accuracy of the generated labels. "Group's Agree Acc" uses the instances where both annotators within the same group agreed, treating this consensus as the ground truth. This method yielded higher accuracy than "Annotator Acc," leading us to resolve disagreements with two additional annotators, ultimately establishing the final labeled data as ground truth.

As reflected in "Ground Truth Acc," in Table 3, our method achieved 94.91% accuracy in generating out-of-scope and in-scope questions, while our proposed AutoDefusionEval method achieved 98.23% accuracy in evaluating defusion. As shown in Figure 1, the confusion matrix on the left aligns closely with the ground truth, with minimal errors (5 false negatives and 6 false positives). The defusion matrix on the right reflects a similarly high accuracy, with only one error for false negatives and false positives.

Model Name	Quantization	Knowledge Cutoff Date
Llama 3.2 3B Instruct Turbo	FP8	2023/12
Llama 3.1 8B Instruct Turbo	FP8	2023/12
Llama 3.1 70B Instruct Turbo	FP8	2023/12
Llama 3.3 70B Instruct Turbo	FP8	2023/12
Mistral (7B) Instruct v0.3	FP16	2024/5
gpt-3.5-turbo	—	2021/9

Table 4: Different LLM$_r$ being evaluated. In ELOQ-Silver, 66 news articles are published before Mistral 7B v0.3's knowledge cutoff date of 5/22/2024.

4 Experimental Results

4.1 Out-of-Scope Response

In this section, we examine LLMs' ability to defuse out-of-scope questions. The selected LLMs are listed in Table 4. Most of these

Group	Annotator	Cohen's Kappa		Annotator Acc		Group's Agree Acc		Ground Truth Acc	
		Out-of-Scope	Defusion	Out-of-Scope	Defusion	Out-of-Scope	Defusion	Out-of-Scope	Defusion
1	A1	0.8053	0.8917	88.89	97.37	92.31	96.97	94.91	98.23
	A2			87.50	94.59				
2	A3	0.7508	0.8187	93.06	97.14	98.41	96.77		
	A4			91.67	90.00				
3	A5	0.8615	0.8333	91.67	93.75	98.51	100.00		
	A6			98.61	100.00				

Table 3: Agreement scores and the performance of our method on the ELOQ-Gold. "Annotator" represents the annotator's name. "Out-of-Scope" represents whether the question is out-of-scope or not. "Defusion" refers to whether an LLM's response indicates that the question has no suitable answer based on the document.

Prompt	Model	business	entm	food	music	news	politics	science	sport	tech	travel	Avg	Std Dev
Basic	GPT-3.5	18.06	10.96	12.32	14.27	13.59	15.95	10.15	18.37*	15.59	9.63	13.89	2.97
	Llama 3.2 3B	63.59	61.55	61.09	62.07	63.49	64.01	53.35	71.71*	63.69	56.51	62.11	4.60
	Mistral 7B v0.3	68.40	67.63	60.88	67.74	67.96	66.86	56.07	76.23*	66.87	61.06	65.97	5.20
	Llama 3.1 8B	71.25	69.42	67.56	72.04	67.16	**71.89**	59.73	77.90*	70.15	63.49	69.06	4.74
	Llama 3.1 70B	**71.54**	70.22	67.56	71.07	67.66	70.75	**61.51**	79.76*	70.67	68.04	69.88	4.33
	Llama 3.3 70B	71.25	**72.71**	**68.17**	**73.51**	**69.84**	69.90	60.04	**82.42***	**73.23**	**68.78**	**70.98**	5.29
Two-shot	GPT-3.5	65.65	65.24	57.60	65.88	62.20	66.48	52.41	71.71*	59.90	56.83	62.39	5.43
	Llama 3.2 3B	80.77	79.28	76.08	79.47	76.79	78.25	72.28	84.09*	79.90	71.64	77.85	3.61
	Mistral 7B v0.3	80.86	77.69	74.54	78.79	79.17	79.11	67.47	86.84*	78.15	71.43	77.40	5.02
	Llama 3.1 8B	83.32	80.98	80.18	84.16	78.77	**82.53**	72.70	87.92*	82.77	78.41	81.17	3.87
	Llama 3.1 70B	85.18	**83.37**	**80.60**	**85.83**	**79.76**	82.24	**75.84**	**90.57***	83.08	**79.79**	**82.62**	3.84
	Llama 3.3 70B	**85.48**	81.27	80.39	84.26	79.27	82.34	**75.84**	88.90*	**83.49**	78.62	81.99	3.57
Zero-shot-CoT	GPT-3.5	63.00	61.95	55.95	64.03	62.40	62.58	53.03	71.71*	59.69	56.30	61.07	4.96
	Llama 3.2 3B	83.91	78.09	77.21	79.96	78.77	81.77	73.85	85.46*	80.41	75.77	79.52	3.39
	Mistral 7B v0.3	77.72	78.39	70.94	78.59	79.07	77.02	67.26	84.77*	76.41	72.06	76.22	4.69
	Llama 3.1 8B	79.39	78.09	74.85	78.40	78.67	79.20	70.29	83.89*	77.64	74.50	77.49	3.43
	Llama 3.1 70B	90.28	87.85	87.27	88.56	86.51	89.08	83.68	91.55*	86.97	85.19	87.69	2.21
	Llama 3.3 70B	**93.52**	**91.24**	**90.55**	**92.57**	**90.77**	**92.12**	**89.12**	**95.09***	**93.85**	**90.90**	**91.97**	1.71

Table 5: Evaluation of LLMs' accuracy in defusing out-of-scope questions from the ELOQ-Silver dataset across diverse news topics. For each LLM, the underscored value and starred (*) value are the minimum and maximum values, respectively, across all the topics within the same prompting method. Bold values are the maximum values for each topic across all the LLMs within the same prompting method. "entm" is the abbreviation of "entertainment".

models have a knowledge cutoff date of December 2023, which predates the publication of the news articles in ELOQ. As a result, they must generate responses based solely on their understanding of the provided news content. As shown in Table 5, we evaluated LLMs on their ability to defuse out-of-scope questions with three prompting methods (Appendix B.6, B.7, B.8). As shown in Table 5, a) Two-shot prompt boosts all evaluated LLMs' accuracy substantially, especially GPT-3.5, likely due to the examples clarifying how to respond to out-of-scope questions. In contrast, Zero-shot-CoT [11] merely invites the LLM to "reason step by step" beats Two-shot on three out of six LLMs. b) Larger models (70B) benefit more from Zero-shot-CoT than from Two-shot prompting, achieving accuracy gains of 5.07% to 9.98%, whereas smaller models (3B to 8B) show only marginal changes, ranging from -3.68% to 1.67%. This difference likely stems from larger models' ability to better utilize their reasoning capabilities and extensive knowledge; c) Most LLMs perform worst on the "science" topic and best on "sport". This is expected, as scientific content tends to contain more implicit premises assumed by the author rather than explicitly stated facts, thus requiring more domain knowledge. In contrast, sports content is simpler and unambiguous, with facts often clearly stated in the document.

4.2 Out-of-Scope Detection

The primary challenge in preventing an LLM from hallucinating or generating incorrect answers to out-of-scope questions is detecting such scenarios effectively. Once we can accurately identify these cases, we can handle them easily by using a prompt and in-context examples specifically designed for out-of-scope questions. We use two methods for out-of-scope detection. First is a prompt-based method where we simply prompt the LLM to do such detection (Appendix B.3), denoted as "Direct Generate" in Figure 2. Second

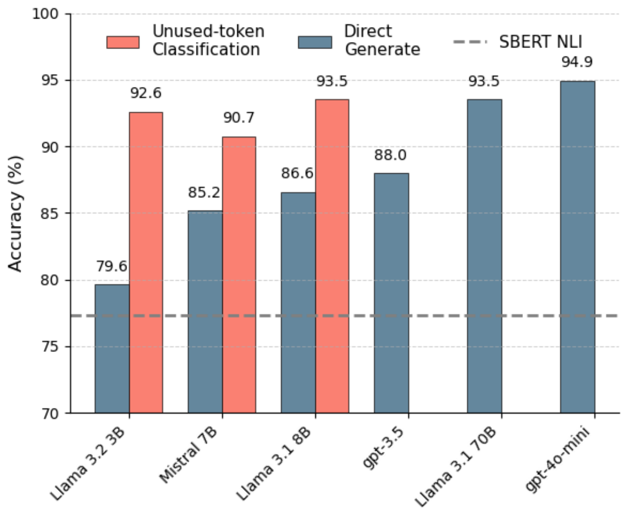

Figure 2: Evaluation of out-of-scope detection on ELOQ-Gold.

is a classification method where we train a binary classifier on ELOQ-Silver using the embeddings of an appended unused-token as features, denoted as "Unused-token Classification" in Figure 2.

4.2.1 Unused-token Classification. The out-of-scope detection problem can be framed simply as a binary classification task, where we determine if a given question q is out-of-scope or not for a given document d. The binary label is $y \in \{0, 1\}$ where $y = 1$ indicates the question is relevant to the document and $y = 0$ indicates the question is out-of-scope. The objective is to learn a classification function: $f : (q, d) \mapsto p(y = 1 \mid q, d)$, which estimates the likelihood of the question-document pair to be in-scope. We would like to test whether the LLM's internal features contain enough information to classify a question-document pair effectively to simulate the effect that the LLM "realizes" the out-of-scope scenario before generating a response. To achieve this, we use the sequence representation of the entire input denoted as $x = \text{concat}(I, q, d,)$ where I is the instruction prompt provided to the LLM (e.g. "Can this document answer the question?"). We extract the sequence representation by appending a reserved unused-token to the end of the input sequence and extracting its representation from the final hidden state of the last transformer layer. The rationale behind using unused-token is that, since LLM developers reserve these tokens for potential future use in specialized training tasks, they are untrained and uninfluenced by any prior data. Thus, the unused-token does not have any predefined meanings and can be viewed as a blank buffer, aggregating interactions among all the input tokens and capturing an uninfluenced representation of the sequence. We also tried using the <eos> token, which yielded similar results.

The classifier is a two-layer MLP that takes the unused-token representation as input and is optimized against the binary label y. This approach is commonly known as "probing" [33], where the LLM weights are frozen, and the classifier can be viewed as an external probe to assess the information contained in the model's internal representations. The performance reported in Figure 2 was tested on ELOQ-Gold, the human-labeled data. For more details about the training, please refer to Appendix C. We also provide a "Sentence BERT NLI" baseline where we adopt NLI style training using the off-the-shelf [4] model from HuggingFace. Specifically, we follow [26] to extract document and question `[CLS]` embeddings e_d and e_q to use $[e_d; e_q; |e_d - e_q|]$ as the input features.

4.2.2 Results Discussion. For the "Direct Generation" approach, we can observe a clear positive correlation between model size and accuracy, aligning with the general trend of increased capability with larger models. However, due to the proprietary nature of OpenAI's models and computational constraints, we are only able to obtain hidden states to train the binary classifier for three LLMs: Llama 3.2 3B, Mistral 7B, and Llama 3.1 8B. Notably, the classifiers consistently outperform the "Direct Generation" method, indicating that the LLM's internal representations encode the necessary information for out-of-scope detection, yet the model may fail when directly prompted. Our dataset, ELOQ, demonstrates its usefulness by enabling even smaller models, such as Llama 3.2 3B-based classifier, to achieve accuracy comparable to significantly larger models that rely on direct generation, including Llama 3.1 70B and GPT-4o-mini. Moreover, as shown in Figure 2, all models and methods outperform the "SBERT NLI" baseline, suggesting that out-of-scope detection is non-trivial and requires a deeper and more complex understanding of linguistic relationships that larger models capture during pretraining and fine-tuning. Lastly, in our exploratory experiments, we trained a classifier on GPT-2, but its accuracy is no better than random guessing, further highlighting the importance of model capacity in leveraging internal representations for classification.

Model	Recall@1	Recall@5	Recall@10	MRR
BM25 [24]	0.4637	0.6880	0.7487	0.5653
SBERT [26]	0.4685	0.7058	0.7858	0.5773
BGE [32]	0.5326	0.7604	0.8267	0.6352
Stella [37]	0.4875	0.7376	0.8128	0.5994
Linq [9]	0.5603	0.7890	0.8492	0.6618

Table 6: Retrieval performance of out-of-scope questions on ELOQ

4.3 Semantic Relevance

To demonstrate that the generated out-of-scope questions are indeed semantically similar to the document from which they are generated, we conduct retrieval experiments using all out-of-scope questions against our document corpus (2,000 documents). We employ the following retrieval models: BM25 [5], SBERT [6], `bge-large-en-v1.5` [7], `stella_en_1.5B_v5`[8], and `Linq-Embed-Mistral`[9].

Among these models, BM25 and SBERT are well-known retrievers that serve as established baselines, `bge-large-en-v1.5` was

[4] https://huggingface.co/sentence-transformers/all-mpnet-base-v2
[5] https://github.com/castorini/pyserini
[6] https://huggingface.co/sentence-transformers/all-mpnet-base-v2
[7] https://huggingface.co/BAAI/bge-large-en-v1.5
[8] https://huggingface.co/NovaSearch/stella_en_1.5B_v5
[9] https://huggingface.co/Linq-AI-Research/Linq-Embed-Mistral

Module	Accuracy
Base (gpt-4o-mini)	87.61
+ examples	97.35
+ self-consistency (m=3)	98.23
+ self-consistency (m=9)	98.23

Table 7: Evaluation of different prompts for "defusion detection" on ELOQ-Gold.

accepted in SIGIR 2024 resource track as a high-performance embedding model, `stella_en_1.5B_v5` ranks #1 on MTEB's English v1 dataset [22], and `Linq-Embed-Mistral` leads multiple benchmarks, including BEIR [8] and MTEB's English v2 dataset at the time of writing. As shown in Table 6, the Recall@1 values across all models suggest that for approximately half of the out-of-scope questions, the top-retrieved document is the same one from which the question was generated. When we relax this condition to Recall@10, around 80% of out-of-scope questions retrieve their associated document within the top 10 results. This demonstrates a strong semantic connection between out-of-scope questions and their source documents, despite the fact that the document lacks the necessary information to answer the question.

4.4 Ablation Study

We examined the impact of various prompts on "defusion detection" task. As seen in Table 7, incorporating examples increased accuracy from 87.61% to 97.35%, and applying self-consistency with $m = 3$ further boosted performance to 98.23%. However, further increasing m did not lead to additional improvements. Since ELOQ-Gold is a smaller dataset, using $m = 9$ did not yield gains, but this doesn't rule out its potential benefit on the larger ELOQ-Silver dataset. Therefore, we adopted $m = 9$ in our experiments to balance accuracy and expenses (Appendix A).

5 Conclusion and Future Work

We introduced ELOQ, a benchmark for evaluating and enhancing LLM detection of out-of-scope questions for a given document. Human annotation validates that ELOQ achieves 94.91% accuracy in generating out-of-scope and in-scope questions, while our AutoDefusionEval achieves 98.23% accuracy in detecting effective defusion responses. We further demonstrated ELOQ's utility by training out-of-scope detectors, and the results show that a small model (Llama 3.1 8B) can achieve comparable results to the larger one (Llama 3.1 70B) without fine-tuning the LLM. For future work, we plan to apply for OpenAI funding to utilize more powerful models to improve both data quality and scale. Additionally, we aim to explore post-training techniques to fine-tune LLMs on ELOQ, ensuring they align better with human preferences by identifying the confusing aspects of a question rather than generating hallucinated answers.

Acknowledgments

We would like to express our sincere gratitude to Xiaoxiao Shang, Ethan Lin, Wei Mo, Zhan Shi, Xinze Ye, and Haisong Wang for their invaluable contributions to data labeling.

Appendix

A Expenses

Generating the ELOQ dataset using GPT-4o-mini costs approximately $80. For LLMs not provided by OpenAI, we used the API service from https://www.together.ai/. Excluding the dataset generation, all other experiments cost around $100.

B Prompt

B.1 Extract Claims

You will be provided with a document delimited by triple quotes. Read the document and follow user's instructions.
Read the document and list {num_fact} most important facts it contains. Each fact should be stated in a clear, standalone sentence with sufficient context to be understood independently, avoiding undefined pronouns. Ensure that each fact is directly derived from the document and does not include any information not mentioned within it.

Document:
"""{document}"""

{num_fact} most important facts:

B.2 Recover Missing Claims

Read the document below with a list of {num_fact} facts it contains. Note that some of the facts are missing represented by (missing). Your task is to guess the missing facts could have said and complete the missing facts. Each fact should be stated in a clear, standalone sentence with sufficient context to be understood independently, avoiding undefined pronouns. Please complete the missing facts and return all the {num_fact} facts in original order. You must return all the {num_fact} facts.

Document:
"""{list_of_fact}"""

Completed list of facts:

B.3 Out-of-scope Judgement

You will be provided with a document delimited by triple quotes. Read the document and follow user's instructions.
Read the document and the question. Check if the question contains any confusing parts: mentions of people, locations, organizations, or other entities that are not present in the document

Document:
"""{document}"""

Question:
{question}

Does the question contain any confusing parts: mentions of people, locations, organizations, or other entities that are not present in the document? Please start by explaining your reasoning. End your response with 'The answer is: Yes or No.':

B.4 Response Defusion Judgement

You will be provided with a document delimited by triple quotes. Read the document and follow user's instructions. Read the document and answer the question based on the document.
Document:
"""{example_document}"""

Question:
{example_questions}

Answer:
{example_LLM_responses}

{Defusion Instruction (Appendix B.5)}
{example_defuse_reasoning}

Read the document and answer the question based on the document.

Document:
"""{document}"""

Question:
{question}

Answer:
{LLM_response}

{Defusion Instruction (Appendix B.5)}

B.5 Defusion Instruction

The question presented is confusing and unanswerable based on the document provided. Read the response carefully. If the response attempts to answer the question, like answering 'Yes' to agree with the question or 'No' to disagree with the question, you should immediately conclude with 'The answer is: No.' even if it later acknowledges the question's confusion. Otherwise, determine if the response clarifies that the question cannot be answered or if the response indicates that the information needed to answer the question is not available or not mentioned in the document. Begin with your reasoning and finally conclude with 'The answer is: Yes.' or 'The answer is: No.'. Note that if the response attempts to answer the question at any point, like answering 'Yes' to agree with the question or 'No' to disagree with the question, you should immediately conclude with 'The answer is: No.' even if it later acknowledges the question's confusion:

B.6 RAG Basic

You will be provided with a document delimited by triple quotes. Read the document and follow user's instructions.
Read the document and answer the question based on the document.

Document:
"""{document}"""

Question:
{question}

Answer:

B.7 RAG Two-shot

You will be provided with a document delimited by triple quotes. Read the document and follow user's instructions.
Read the document and answer the question based on the document.

Document:
"""{example_document}"""

Question:
{two_example_questions}

Answer:
{two_example_responses}

Read the document and answer the question based on the document.

Document:
"""{document}"""

Question:
{question}

Answer:

B.8 RAG Zero-shot-CoT

You will be provided with a document delimited by triple quotes. Read the document and follow user's instructions.
Read the document and reason step by step to answer the question based on the document. If the question cannot be answered using the document, state explicitly that the question cannot be answered.

Document:
"""{document}"""

Question:
{question}

Answer:

B.9 In-scope Question Generation

You will be provided with a document delimited by triple quotes. Read the document and follow user's instructions. Read the document attentively and compile a numbered list of the top {num_q} questions that the document directly answers. Ensure each question is clear, accurate, and devoid of confusion, false assumptions, undefined pronouns, or misinformation. Avoid referencing people, locations, organizations, or other entities not explicitly mentioned in the document. Construct each question to be thought-provoking, containing between 13 to 18 words, and sufficiently detailed to avoid being overly straightforward.

Document:
"""{document}"""

Questions:

B.10 Out-of-scope Question Generation

You will be provided with a document delimited by triple quotes. Read the document and follow user's instructions.
Read the document and review the list of hallucinated facts. For each hallucinated fact, craft a single, specific and concise question containing 13 to 18 words that incorporate the key element of the fact, ensuring the question is intentionally confusing. The question should not be answerable using any information present in the document. The question should not combine multiple queries and each question should address only one specific aspect. If a question cannot be formulated for a particular hallucinated fact, you may omit it.

Document:
"""{document}"""

hallucinated facts:
{hallucinated_facts}

Questions:

B.11 Remove Claims

You will be provided with a document delimited by triple quotes. Read the document and follow user's instructions.
Read the document below with a list of {num_true_fact} ground-truth facts it contains and a list of {num_false_fact} hallucinated facts that are not supported by the document. Your task is to remove any hallucinated facts that can be supported by either the document or the {num_true_fact} ground-truth facts. Please only return the remaining hallucinated facts, along with their original order numbers.

Document:
"""{example_document}"""

{num_true_fact} ground-truth facts:
{true_facts}

{num_false_fact} hallucinated facts:
{false_facts}

Remaining hallucinated facts:
{remained_facts}

Read the document below with a list of {num_true_fact} ground-truth facts it contains and a list of {num_false_fact} hallucinated facts that are not supported by the document. Your task is to remove any hallucinated facts that can be supported by either the document or the {num_true_fact} ground-truth facts. Please only return the remaining hallucinated facts, along with their original order numbers.

Document:
"""{document}"""

{num_true_fact} ground-truth facts:
{true_facts}

{num_false_fact} hallucinated facts:
{hallucinated_facts}

Remaining hallucinated facts:

C Embedding Classifier Implementation Details

We train the classifier on ELOQ-Silver, which is split as: 80% for training, 10% for evaluation, and 10% for testing. After training, we test the classifier on ELOQ-Gold data.

For the Llama family, we use `<|reserved_special_token_0|>` as the unused-token. This token is not trained and is reserved for future use cases such that the users do not need to resize the vocabulary. For Mistral-7B-v0.3, we use `[control_555]` as the unused-token because Mistral's official documentation indicated that they reserve 768 control tokens for future use, and none of them are used during training. 555 is a random choice.

The unused tokens ' hidden state is used as input features for a 2-layer MLP, which is optimized with a binary cross-entropy (BCE) loss. A dropout of 0.1 is applied after the first layer. We use Adam optimizer with 1e-4 learning rate. We train each classifier for 10 epochs with batch size 8. The best validation checkpoint is saved and its performance is reported on ELOQ-Gold.

D Example News

Justin Jefferson, CeeDee Lamb, NFL Injury Statuses and Fantasy Impact for Week 3 Stephen Maturen/Getty Images

Two of the best wide receivers in the NFL are expected to be on the field for their respective Week 3 games. Justin Jefferson and CeeDee Lamb each received positive prognosis about their injury issues, which is something that can't be said for the rest of the stars across the NFL. Christian McCaffrey is already on injured reserve, Deebo Samuel is out for a weeks and a slew of other running backs and wide receivers are dealing with ailments that could keep them out of Week 3. Below is a look at all of the significant injuries that could affect fantasy football matchups across Week 3. Justin Jefferson Off Injury Report

Justin Jefferson was taken off the Minnesota Vikings injury report on Friday. Jefferson's status was up in the air because of a quad injury, but he practiced well enough this week that the injury is not a concern. The superstar wide out will be needed for Minnesota's home clash with the Houston Texans, which has the potential to be a high-scoring affair. Jefferson is always the primary target in Minnesota when healthy, but he should have more targets in Week 3 because Jordan Addison and T.J. Hockenson are still out. Jefferson earned seven targets from Sam Darnold in Week 2. Only running back Aaron Jones had more than four targets against the New York Giants. Houston's defense may give Jefferson some fits, led by cornerback Derek Stingley Jr. but it has allowed 330 receiving yards to opposing wide outs on just 20 catches through two weeks. Jefferson is an automatic start whenever he's healthy, and his star power may be needed more on certain fantasy rosters in Week 3 depending on how many injuries affect a single roster.

References

[1] Artem Bugara, Maksym Sugonyaka, and Becket Trotter. 2020. WNewscatcher. https://github.com/kotartemiy/newscatcher

[2] Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the Surprising Difficulty of Natural Yes/No Questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, 2924–2936. doi:10.18653/V1/N19-1300

[3] William G. Cochran. 1977. *Sampling Techniques, 3rd Edition.* John Wiley.

[4] Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. 2024. RAGAS: Automated Evaluation of Retrieval Augmented Generation. In *EACL'24: Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations.* St. Julians, Malta, 150–158. https://arxiv.org/abs/2309.15217

[5] 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 https://arxiv.org/abs/2312.10997

[6] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Retrieval Augmented Language Model Pre-Training. In *ICML'20: Proceedings of the 37th International Conference on Machine Learning.* Vienna, Austria, 3929–3938. https://arxiv.org/abs/2002.08909

[7] Shengding Hu, Yifan Luo, Huadong Wang, Xingyi Cheng, Zhiyuan Liu, and Maosong Sun. 2023. Won't Get Fooled Again: Answering Questions with False Premises. In *ACL'23: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).* Toronto, Canada, 5626–5643. https://arxiv.org/abs/2307.02394

[8] Ehsan Kamalloo, Nandan Thakur, Carlos Lassance, Xueguang Ma, Jheng-Hong Yang, and Jimmy Lin. 2023. Resources for Brewing BEIR: Reproducible Reference Models and an Official Leaderboard. arXiv:2306.07471 [cs.IR]

[9] Junseong Kim, Seolhwa Lee, Jihoon Kwon, Sangmo Gu, Yejin Kim, Minkyung Cho, Jy yong Sohn, and Chanyeol Choi. 2024. Linq-Embed-Mistral:Elevating Text Retrieval with Improved GPT Data Through Task-Specific Control and Quality Refinement. Linq AI Research Blog. https://getlinq.com/blog/linq-embed-mistral/

[10] Najoung Kim, Ellie Pavlick, Burcu Karagol Ayan, and Deepak Ramachandran. 2021. Which Linguist Invented the Lightbulb? Presupposition Verification for Question-Answering. In *ACL'21: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics.* Bangkok, Thailand, 3932–3945. https://aclanthology.org/2021.acl-long.304.pdf

[11] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large Language Models are Zero-Shot Reasoners. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (Eds.). http://papers.nips.cc/paper_files/paper/2022/hash/8bb0d291acd4acf06ef112099c16f326-Abstract-Conference.html

[12] Weronika Lajewska and Krisztian Balog. 2024. Towards Reliable and Factual Response Generation: Detecting Unanswerable Questions in Information-Seeking Conversations. In *Advances in Information Retrieval - 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24-28, 2024, Proceedings, Part III (Lecture Notes in Computer Science, Vol. 14610)*, Nazli Goharian, Nicola Tonellotto, Yulan He, Aldo Lipani, Graham McDonald, Craig Macdonald, and Iadh Ounis (Eds.). Springer, 336–344. doi:10.1007/978-3-031-56063-7_25

[13] Gyeongbok Lee, Seung-won Hwang, and Hyunsouk Cho. 2020. SQuAD2-CR: Semi-supervised Annotation for Cause and Rationales for Unanswerability in SQuAD 2.0. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association, 5425–5432. https://aclanthology.org/2020.lrec-1.667/

[14] Young-Suk Lee, Md Arafat Sultan, Yousef El-Kurdi, Tahira Naseem, Asim Munawar, Radu Florian, Salim Roukos, and Ramón Fernandez Astudillo. 2023. Ensemble-Instruct: Instruction Tuning Data Generation with a Heterogeneous Mixture of LMs. In *Findings of the Association for Computational Linguistics: the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP'23).* Singapore, 12561–12571. https://arxiv.org/abs/2310.13961

[15] 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. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *NeurIPS'20: Proceedings of the 34th Conference on Neural Information Processing Systems.* Vancouver, Canada, 9459–9474. https://arxiv.org/abs/2005.11401

[16] Yinghui Li, Qingyu Zhou, Yuanzhen Luo, Shirong Ma, Yangning Li, Hai-Tao Zheng, Xuming Hu, and Philip S Yu. 2024. When llms meet cunning questions: A fallacy understanding benchmark for large language models. *arXiv preprint arXiv:2402.11100* (2024).

[17] Yinghui Li, Qingyu Zhou, Yuanzhen Luo, Shirong Ma, Yangning Li, Hai-Tao Zheng, Xuming Hu, and Philip S. Yu. 2024. When LLMs Meet Cunning Texts: A Fallacy Understanding Benchmark for Large Language Models. arXiv:2402.11100 https://arxiv.org/abs/2402.11100

[18] Yuxuan Liu, Tianchi Yang, Shaohan Huang, Zihan Zhang, Haizhen Huang, Furu Wei, Weiwei Deng, Feng Sun, and Qi Zhang. 2024. Calibrating LLM-Based Evaluator. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024).* Torino, Italia, 2638–2656. https://arxiv.org/abs/2309.13308

[19] Akash Maharaj, Kun Qian, Uttaran Bhattacharya, Sally Fang, Horia Galatanu, Manas Garg, Rachel Hanessian, Nishant Kapoor, Ken Russell, Shivakumar Vaithyanathan, and Yunyao Li. 2024. Evaluation and Continual Improvement for an Enterprise AI Assistant. In *DaSH 2024: Proceedings of the Fifth Workshop on Data Science with Human-in-the-Loop.* Association for Computational Linguistics, Mexico City, Mexico, 17–24. https://arxiv.org/abs/2407.12003

[20] Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. AmbigQA: Answering Ambiguous Open-domain Questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, 5783–5797. doi:10.18653/V1/2020.EMNLP-MAIN.466

[21] Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. AmbigQA: Answering Ambiguous Open-domain Questions. In *EMNLP'20: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing.* 5783–5797. https://arxiv.org/abs/2004.10645

[22] Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2022. MTEB: Massive Text Embedding Benchmark. *arXiv preprint arXiv:2210.07316* (2022). doi:10.48550/ARXIV.2210.07316

[23] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. In *Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, December 9, 2016 (CEUR Workshop Proceedings, Vol. 1773)*, Tarek Richard Besold, Antoine Bordes, Artur S. d'Avila Garcez, and Greg Wayne (Eds.). CEUR-WS.org. https://ceur-ws.org/Vol-1773/CoCoNIPS_2016_paper9.pdf

[24] Joaquín Pérez-Iglesias, José R. Pérez-Agüera, Víctor Fresno, and Yuval Z. Feinstein. 2009. Integrating the Probabilistic Models BM25/BM25F into Lucene. *CoRR* abs/0911.5046 (2009). arXiv:0911.5046 http://arxiv.org/abs/0911.5046

[25] Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know What You Don't Know: Unanswerable Questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers*, Iryna Gurevych and Yusuke Miyao (Eds.). Association for Computational Linguistics, 784–789. doi:10.18653/V1/P18-2124

[26] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. 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 2019, Hong Kong, China, November 3-7, 2019*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, 3980–3990. doi:10.18653/V1/D19-1410

[27] Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. 2024. ARES: An Automated Evaluation Framework for Retrieval-Augmented Generation Systems. In *NAACL'24: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers).* Mexico City, Mexico, 338–354. https://arxiv.org/abs/2311.09476

[28] Elior Sulem, Jamaal Hay, and Dan Roth. 2022. Yes, No or IDK: The Challenge of Unanswerable Yes/No Questions. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, Marine Carpuat, Marie-Catherine de Marneffe, and Iván Vladimir Meza Ruíz (Eds.). Association for Computational Linguistics, 1075–1085. doi:10.18653/V1/2022.NAACL-MAIN.79

[29] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *ICLR'23: Proceedings of the 11th International Conference on Learning Representations.* Kigali, Rwanda. https://arxiv.org/pdf/2203.11171

[30] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-Instruct: Aligning Language Models with Self-Generated Instructions. In *ACL'23: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).* Toronto, Canada, 13484–13508. https://arxiv.org/abs/2212.10560

[31] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-Thought Prompting

Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (Eds.). http://papers.nips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html

[32] Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-Pack: Packaged Resources To Advance General Chinese Embedding. arXiv:2309.07597 [cs.CL]

[33] Tian Ye, Zicheng Xu, Yuanzhi Li, and Zeyuan Allen-Zhu. 2024. Physics of Language Models: Part 2.1, Grade-School Math and the Hidden Reasoning Process. *CoRR* abs/2407.20311 (2024). doi:10.48550/ARXIV.2407.20311 arXiv:2407.20311

[34] Asaf Yehudai, Boaz Carmeli, Yosi Mass, Ofir Arviv, Nathaniel Mills, Eyal Shnarch, and Leshem Choshen. 2024. Achieving Human Parity in Content-Grounded Datasets Generation. In *ICLR'24: Proceedings of the 12th International Conference on Learning Representations*. Vienna, Austria. https://arxiv.org/abs/2401.14367

[35] Xinyan Velocity Yu, Sewon Min, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. CREPE: Open-Domain Question Answering with False Presuppositions. In *ACL'23: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Toronto, Canada, 10457–10480. https://arxiv.org/abs/2211.17257

[36] Hongbang Yuan, Pengfei Cao, Zhuoran Jin, Yubo Chen, Daojian Zeng, Kang Liu, and Jun Zhao. 2024. Whispers that Shake Foundations: Analyzing and Mitigating False Premise Hallucinations in Large Language Models. In *ICLR'24: Proceedings of the 12th International Conference on Learning Representations*. Vienna, Austria. https://arxiv.org/abs/2402.19103

[37] Dun Zhang and FulongWang. 2024. Jasper and Stella: distillation of SOTA embedding models. *CoRR* abs/2412.19048 (2024). doi:10.48550/ARXIV.2412.19048 arXiv:2412.19048

[38] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *NeurIPS'23: Proceedings of the 37th International Conference on Neural Information Processing Systems*. New Orleans, LA, USA, 46595–46623. https://arxiv.org/abs/2306.05685

[39] Yanxu Zhu, Jinlin Xiao, Yuhang Wang, and Jitao Sang. 2024. KG-FPQ: Evaluating Factuality Hallucination in LLMs with Knowledge Graph-based False Premise Questions. arXiv:2407.05868 https://arxiv.org/abs/2407.05868

[40] Weronika Łajewska and Krisztian Balog. 2023. Towards Filling the Gap in Conversational Search: From Passage Retrieval to Conversational Response Generation. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23)*. ACM. doi:10.1145/3583780.3615132