Tug-of-War Between Knowledge: Exploring and Resolving Knowledge Conflicts in Retrieval-Augmented Language Models

Zhuoran Jin[1,2], **Pengfei Cao**[1,2], **Yubo Chen**[1,2,*], **Kang Liu**[1,2,3],
Xiaojian Jiang[4], **Jiexin Xu**[4], **Qiuxia Li**[4], **Jun Zhao**[1,2]
[1] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[2] The Laboratory of Cognition and Decision Intelligence for Complex Systems,
Institute of Automation, Chinese Academy of Sciences, Beijing, China
[3] Shanghai Artificial Intelligence Laboratory [4] China Merchants Bank
{zhuoran.jin, pengfei.cao, yubo.chen, kliu, jzhao}@nlpr.ia.ac.cn

Abstract

Retrieval-augmented language models (RALMs) have demonstrated significant potential in refining and expanding their **internal memory** by retrieving evidence from **external sources**. However, RALMs will inevitably encounter **knowledge conflicts** when integrating their internal memory with external sources. Knowledge conflicts can ensnare RALMs in a tug-of-war between knowledge, limiting their practical applicability. In this paper, we focus on exploring and resolving knowledge conflicts in RALMs. First, we present an evaluation framework for assessing knowledge conflicts across various dimensions. Then, we investigate the behavior and preference of RALMs from the following two perspectives: (1) Conflicts between internal memory and external sources: We find that stronger RALMs emerge with the **Dunning-Kruger effect**, persistently favoring their faulty internal memory even when correct evidence is provided. Besides, RALMs exhibit an **availability bias** towards common knowledge; (2) Conflicts between truthful, irrelevant and misleading evidence: We reveal that RALMs follow the principle of **majority rule**, leaning towards placing trust in evidence that appears more frequently. Moreover, we find that RALMs exhibit **confirmation bias**, and are more willing to choose evidence that is consistent with their internal memory. To solve the challenge of knowledge conflicts, we propose a method called **C**onflict-**D**isentangle **C**ontrastive **D**ecoding (**CD^2**) to better calibrate the model's confidence. Experimental results demonstrate that our CD^2 can effectively resolve knowledge conflicts in RALMs.

Keywords: Knowledge Conflicts, Retrieval-Augmented Language Models, Internal Memory, External Sources

1. Introduction

Large language models (LLMs) (Brown et al., 2020; OpenAI, 2023) have memorized a substantial amount of factual knowledge during pretraining, and encapsulated this knowledge within their parameters as **internal memory** or **parametric knowledge** (Zhu and Li, 2023; Sun et al., 2023). Nevertheless, the internal memory may sometimes be inaccurate or outdated, making LLMs prone to hallucination (Ji et al., 2023), where generated responses may seem plausible but are fictional. As shown in Figure 1, for the question "*Who won the latest Nobel Prize in Physics?*", its correct answer is "*Pierre Agostini*". However, relying solely on unfactual internal memory to answer this question may lead to incorrect answers, such as "*Benjamin List*".

To address the issue of hallucination, one promising solution is to employ retrieval-augmented language models (RALMs) (Lewis et al., 2020; Izacard and Grave, 2021; Ram et al., 2023; Shi et al., 2023b; Jin et al., 2023). RALMs first retrieve a handful of reference evidence from **external sources** or **non-parametric knowledge**, and subsequently integrate these external sources with their internal memory for generation. However, owing to challenges like misinformation, divergent perspectives and the evolving nature of knowledge (Jang et al., 2022), **knowledge conflicts** (Longpre et al., 2021; Chen et al., 2022; Xie et al., 2023) widely exist in RALMs. As illustrated in Figure 1, the model's in-

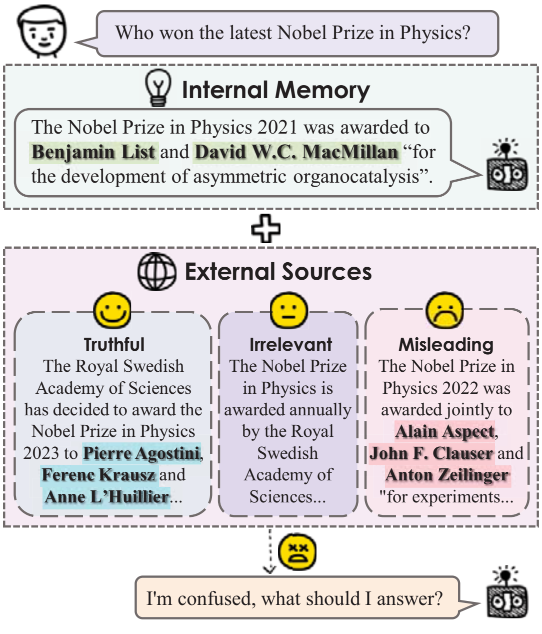
Who won the latest Nobel Prize in Physics?
Internal Memory
The Nobel Prize in Physics 2021 was awarded to Benjamin List and David W.C. MacMillan "for the development of asymmetric organocatalysis".
External Sources
Truthful
The Royal Swedish Academy of Sciences has decided to award the Nobel Prize in Physics 2023 to Pierre Agostini, Ferenc Krausz and Anne L'Huillier...
Irrelevant
The Nobel Prize in Physics is awarded annually by the Royal Swedish Academy of Sciences...
Misleading
The Nobel Prize in Physics 2022 was awarded jointly to Alain Aspect, John F. Clauser and Anton Zeilinger "for experiments...
I'm confused, what should I answer?

Figure 1: An example of knowledge conflicts.

[*] Corresponding author.

LREC-COLING 2024, pages 16867–16878
20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

ternal memory ("*Benjamin List*") may conflict with external sources ("*Pierre Agostini*"), causing interference. Moreover, knowledge conflicts also arise among external sources due to the presence of **truthful** ("*Pierre Agostini*"), **irrelevant** and **misleading** ("*Alain Aspect*") evidence. *Knowledge conflicts can ensnare RALMs in a tug-of-war between knowledge*, making them confusing and potentially limiting their practical applicability. Therefore, it is necessary to **explore** and **resolve** knowledge conflicts in retrieval-augmented language models.

In this paper, we introduce an evaluation framework that conducts a thorough investigation into knowledge conflicts: (1) We reconstruct four question answering datasets to assess knowledge conflicts from open-domain (*i.e.*, single evidence), entity-centric (*i.e.*, question with entity popularity), and multi-hop (*i.e.*, multiple pieces of evidence) perspectives; (2) We employ seven open-source LLMs (*e.g.*, FLAN-T5, LLaMA2) and one API-access LLM (*e.g.*, ChatGPT) to analyze the influence of model sizes and capabilities under conflicts; (3) We evaluate RALMs from three key dimensions: **correctness** (*i.e.*, whether RALMs can answer questions correctly based on internal memory or external sources?), **faithfulness** (*i.e.*, whether RALMs refer to external sources in answering, and which type of evidence is preferred?) and **memorization** (*i.e.*, how well RALMs stick to their internal memory?).

Based on this, we systematically investigate the behavior and preference of RALMs in the conflicting situation from the following two perspectives:

Knowledge Conflicts Between Internal Memory and External Sources. (1) When we provide knowledge that conflicts with the internal memory of RALMs as external sources, we observe that RALMs **easily change** their internal beliefs to believe the external knowledge. *As model sizes and capabilities expand, the model gains greater confidence in its internal memory and has a certain ability to perceive conflicts*; (2) Then, we come to wonder *whether the model has different preferences for various types of knowledge?* We show that RALMs exhibit an **availability bias** towards common knowledge, *preferring knowledge that is easily accessible in memory*. For those long-tail knowledge, the model depends on external sources in most cases. As the knowledge becomes more common, the model gradually shifts towards placing more trust in its internal memory; (3) Besides, we further conduct an in-depth analysis from the viewpoint of correct and incorrect internal memory. We find that *capable LLMs may lack confidence in their correct internal memory, but will stubbornly persist in their incorrect internal memory*. For instance, even when provided with correct external evidence, ChatGPT often persists in trusting its incorrect internal memory for more than half the time. This phenomenon aligns with the **Dunning-Kruger effect** in human psychology (Kruger and Dunning, 1999; Dunning, 2011; Singh et al., 2023), wherein individuals tend to overestimate their abilities when they have limited competence in specific domains.

Knowledge Conflicts Between Truthful, Irrelevant and Misleading Evidence. (1) RALMs often struggle to discern truthful evidence from misleading evidence, and at times, they may be distracted by irrelevant data. RALMs follow the principle of **majority rule**, *leaning towards placing trust in evidence that appears more frequently*; (2) Furthermore, we consider a more interesting scenario where external sources contain evidence supporting the model's internal memory. To this end, we induce the model's internal memory and integrate it into external sources. We find RALMs exhibit **confirmation bias** (Nickerson, 1998; Xie et al., 2023), *being more inclined to choose evidence that is consistent with their own internal memory, regardless of whether it is correct or incorrect*; (3) We also examine knowledge conflicts in multi-hop reasoning scenarios, where RALMs must integrate multiple pieces of evidence to answer the question. *As the count of conflicting hops increases, the model faces greater difficulty in reasoning the correct answer.*

To solve the challenge of knowledge conflicts, we propose a novel method called **C**onflict-**D**isentangle **C**ontrastive **D**ecoding (**CD^2**) to better calibrate the model's confidence. For knowledge conflicts between internal memory and external sources, to mitigate the impact of the RALM's incorrect internal memory, we leverage contrastive decoding to amplify the difference between output logits with and without external sources. To address knowledge conflicts between truthful, irrelevant and misleading evidence, we adopt **fact-aware** instruction tuning to enable the RALM to be aware of truthful and misleading evidence. In detail, we train an **expert** LM to generate truthful answers and an **amateur** LM to generate misleading answers, and then utilize contrastive decoding to maximize the difference between expert logits and amateur logits.

Our contributions are summarized as follows:

- We introduce an evaluation framework that conducts a thorough investigation into knowledge conflicts in RALMs. We conduct experiments with eight models on four QA datasets, and then evaluate them from three dimensions: correctness, faithfulness and memorization.
- We investigate the behavior and preference of RALMs in the tug-of-war between knowledge. We find that powerful RALMs emerge with the Dunning-Kruger effect when internal memory and external sources conflict. Moreover, RALMs follow the principle of majority rule and exhibit confirmation bias when facing truthful, irrelevant and misleading evidence.

- We propose a novel method called Conflict-Disentangle Contrastive Decoding (CD^2) to better calibrate the model's confidence. Experimental results demonstrate that our CD^2 can effectively resolve knowledge conflicts in RALMs, with an average Recall improvement of 2.35% and 2.41% on two datasets. Our code will be publicly available at https://github.com/jinzhuoran/KConflict/.

2. Related Work

2.1. Retrieval-Augmented Language Models

Retrieval-augmented language models (RALMs) have shown remarkable potential in mitigating hallucinations (Chen et al., 2023a) and expanding knowledge boundaries (Ren et al., 2023). Previous methods (Guu et al., 2020; Lewis et al., 2020; Izacard and Grave, 2021; Borgeaud et al., 2022b) focus on pre-training a smaller LM to better utilize retrieved evidence. Recently, some work (Yu et al., 2023; Ram et al., 2023; Shi et al., 2023b) has unleashed the in-context learning capabilities of LLMs, simply prepending retrieved evidence to the input without updating any parameters. In this paper, we keep in line with the in-context learning paradigm. Thus, we adopt frozen LLMs to extract answers from a handful of retrieved passages. We do not consider those RALMs like RAG (Lewis et al., 2020) and FiD (Izacard and Grave, 2021), because their limited size makes it difficult to answer questions correctly relying solely on their internal memory.

2.2. Knowledge Conflicts

Knowledge conflicts (Longpre et al., 2021; Chen et al., 2022; Shi et al., 2023a; Xie et al., 2023; Wang et al., 2023; Neeman et al., 2023) have recently garnered the attention of researchers. Longpre et al. (2021) propose an entity substitution framework to create entity-based conflicts by replacing a gold entity mention with an alternate entity. They show that models frequently rely on their parametric knowledge, generating answers not present in the given evidence. Chen et al. (2022) consider a more realistic scenario, where models consider multiple evidence passages. Different from Longpre et al. (2021), they find that when provided with a high recall retriever, models rely almost exclusively on the evidence passages rather than internal memory. Xie et al. (2023) argue that LLMs do not trust entity substitution-based conflicting evidence, possibly because of the incoherence of the evidence. To this end, they instruct LLMs to generate coherent conflicting evidence, and reveal that when both supportive and contradictory evidence to their parametric memory are present, LLMs tend to cling to their parametric memory. However, most of the existing studies only consider the situation where the model encounters conflicts when its internal memory is correct. In this paper, we separately investigate the behavior of RALMs in cases of both correct and incorrect internal memory and uncover some unique phenomena. Besides, we also propose an effective method to solve knowledge conflicts in RALMs.

3. Evaluation Framework

3.1. Datasets

Following previous work (Chen et al., 2022; Xie et al., 2023), we adopt the question answering (QA) task to explore knowledge conflicts in the open-book setting. We choose four QA datasets: NQ (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017), PopQA (Mallen et al., 2023) and MuSiQue (Trivedi et al., 2022). NQ and TriviaQA are both commonly used open-domain QA datasets. PopQA is an entity-centric QA dataset including truly long-tail distribution. PopQA is constructed from triples in Wikidata where the subject entity in each question has popularity. MuSiQue is a multi-hop QA dataset with 2-4 hop questions. For each dataset, we randomly sample 500 or 1000 questions from the original testing set as the evaluation data. We only select questions for which supporting evidence can be retrieved. All the datasets use Wikipedia (Petroni et al., 2021) as their external source.

3.2. Analyzed Models

To analyze the impact of model sizes and capabilities, we adopt seven open-source LLMs, including FLAN-T5-XL 3B (Chung et al., 2022), FLAN-T5-XXL 11B, FLAN-UL2 20B (Tay et al., 2023), Baichuan2 7B (Yang et al., 2023), Baichuan2 13B, LLaMA2 7B (Touvron et al., 2023) and LLaMA2 13B. Besides, we use the `gpt-3.5-turbo` version of ChatGPT as the API-access LLM. We use in-context learning to let RALMs answer the questions conditioned on $M \in \{4, 8, 16\}$ demonstrations and $K \in \{3, 5, 10, 20\}$ retrieved evidence.

3.3. Evaluation Metrics

Correctness According to the previous study, we use EM and F1 to evaluate whether the model can provide correct answers. Since LLMs tend to produce verbose yet accurate responses, which may lead to lower EM and F1 scores. Therefore, following Adlakha et al. (2023), we adopt Recall (R) to compute the proportion of tokens in the gold reference that are present in the prediction.

Faithfulness To assess the extent to which the model's predictions depend on retrieved evidence,

Model	Internal Memory			Correct Memory w/ C			Incorrect Memory w/ C			IMR - CMR
	EM	F1	R	Mem R	Con R	MR	Mem R	Con R	MR	
					NQ					
FLAN-T5-XL	10.34	16.68	19.02	14.42	82.87	7.20	6.41	74.75	1.32	-5.88
FLAN-T5-XXL	14.43	21.29	23.44	14.36	82.53	9.83	7.86	77.63	2.83	-7.00
FLAN-UL2	22.92	29.69	30.11	11.33	**86.13**	7.11	10.22	**76.40**	4.75	-2.36
Baichuan2 7B	21.88	30.18	32.47	16.51	83.40	10.31	16.82	65.71	10.20	-0.11
Baichuan2 13B	26.02	34.40	37.33	22.34	77.31	15.91	21.19	58.68	12.76	-3.15
LLaMA2 7B	29.04	39.24	41.72	19.04	82.03	13.10	22.95	65.77	12.92	-0.18
LLaMA2 13B	**34.86**	**46.27**	50.26	15.30	84.90	8.47	26.02	60.27	17.53	9.06
ChatGPT	25.77	38.56	**59.84**	**27.68**	80.40	**18.14**	**39.23**	62.58	**50.69**	**32.55**
					TriviaQA					
FLAN-T5-XL	27.43	32.76	33.83	16.74	76.94	16.10	8.97	84.37	8.84	-7.26
FLAN-T5-XXL	35.67	41.14	41.89	17.89	80.14	17.85	8.88	83.96	8.54	-9.31
FLAN-UL2	47.47	53.42	54.82	16.10	82.86	15.92	19.50	**84.60**	19.70	3.78
Baichuan2 7B	56.58	60.71	62.32	16.76	**82.30**	15.96	18.55	76.03	15.90	-0.06
Baichuan2 13B	60.10	64.49	66.63	22.46	78.89	**21.27**	19.54	73.71	18.94	-2.33
LLaMA2 7B	61.58	65.46	66.60	19.16	79.14	19.10	20.33	76.67	17.39	-1.71
LLaMA2 13B	**67.20**	**72.01**	74.04	16.76	81.90	16.16	18.87	75.14	18.39	2.23
ChatGPT	56.60	66.30	**82.88**	**26.84**	76.17	19.57	**39.72**	69.87	**39.41**	**19.84**

Table 1: Experimental results under knowledge conflicts between internal memory and external sources on NQ and TriviaQA. Mem R denotes the Recall between predictions and internal memory predictions. Con R denotes the Recall between predictions and conflicting references. IMR - CMR denotes the difference in memorization ratio between incorrect memory and correct memory. Bold denotes the highest results.

we adopt K-Precision (KP) (Adlakha et al., 2023) to calculate the proportion of tokens in the prediction that are present in external evidence. K-Precision can also be used to quantify the preference for truthful, irrelevant and misleading evidence.

Memorization We leverage the memorization ratio (Longpre et al., 2021; Chen et al., 2022; Xie et al., 2023) $MR = \frac{f_m}{f_m+f_s}$ to measure how often the model sticks to its internal memory, where f_m is the frequency of relying on internal memory, f_s is the frequency of relying on external sources.

3.4. Memory Induction and Conflict Generation

We adopt closed-book QA to induce the model's internal memory and perform greedy decoding of the answers. Then we prompt the model to further generate evidence supporting its internal memory based on the elicited answers. To construct more confusing and coherent conflicting knowledge, we distill counterfactuals (Xie et al., 2023; Chen et al., 2023b) with LLMs. We unleash the creative power of ChatGPT with a temperature $\tau = 1.0$, enabling it to create counterfactual answers and conflicting evidence based on the given questions, reference answers, and supporting evidence.

4. Conflicts Between Internal Memory and External Sources

In this section, we investigate knowledge conflicts between internal memory and external sources. Specifically, we first elicit the model's internal memory through closed-book QA, then classify the internal memory according to whether it is correct or not. For questions that the model's internal memory can answer correctly/incorrectly, we provide $K = 3$ conflicting incorrect/correct external evidence.

RALMs easily change their internal memory to believe the conflicting external knowledge. As shown in Table 1, we conduct experiments on NQ and TriviaQA. We can find that after providing conflicting evidence, Con R (Recall of conflicting references) is much higher than Mem R (Recall of internal memory). This demonstrates that *RALMs tend to trust external knowledge, even when it conflicts with their initial internal memory.*

The greater the model's capability, the more confident it becomes. Intuitively, we can observe that as the model's capacity increases, the correctness (*e.g.*, EM, F1 and R) of its internal memory also improves. Meanwhile, we find that along with the increase in correctness, there is also an improvement in the MR. This indicates that *as model sizes and capabilities expand, the model gains greater confidence in its internal memory*. To conduct a comprehensive analysis of the changes in the model's confidence score, we categorize the model's behavior under conflict into four distinct groups: (1) Change Inco: Modifying incorrect internal memory based on conflicting evidence; (2) Sustain Inco: Preserving incorrect internal memory despite conflicting evidence; (3) Change Corr: Modifying correct internal memory based on conflicting evidence; (4) Sustain Corr: Preserving correct internal memory despite conflicting evidence. As shown in Figure 2, FLAN-T5 consistently increases

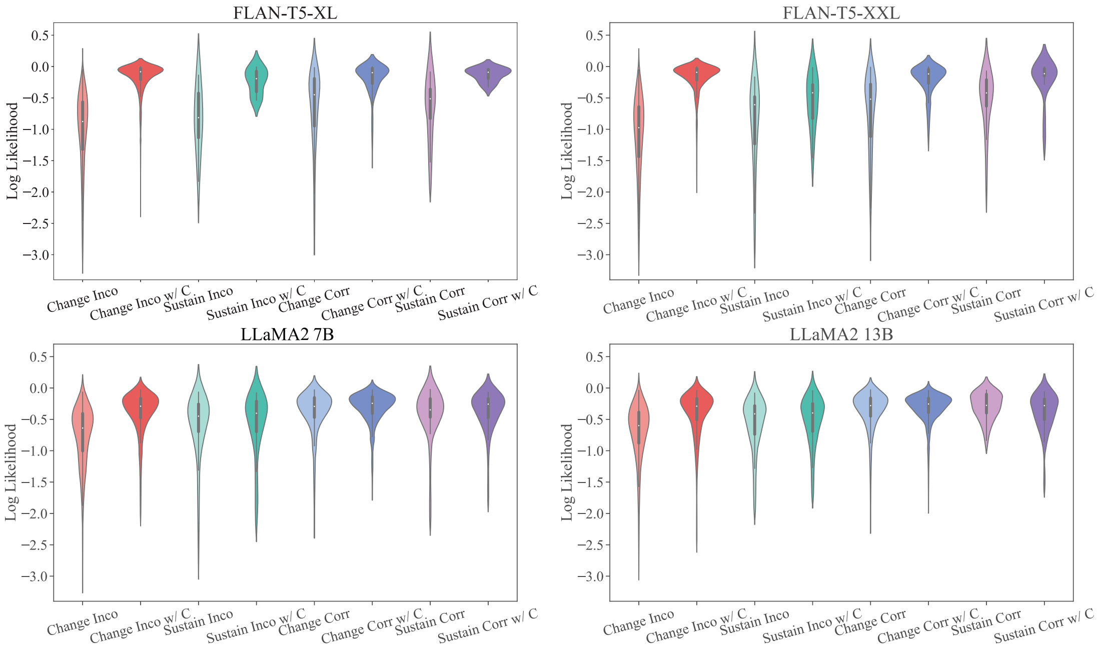

Figure 2: Confidence score distributions of different models under knowledge conflicts between internal memory and external sources on NQ. w/ C denotes providing conflicting evidence to the model. The wider the violin plot, the denser the data points. The larger the log likelihood, the higher the confidence score.

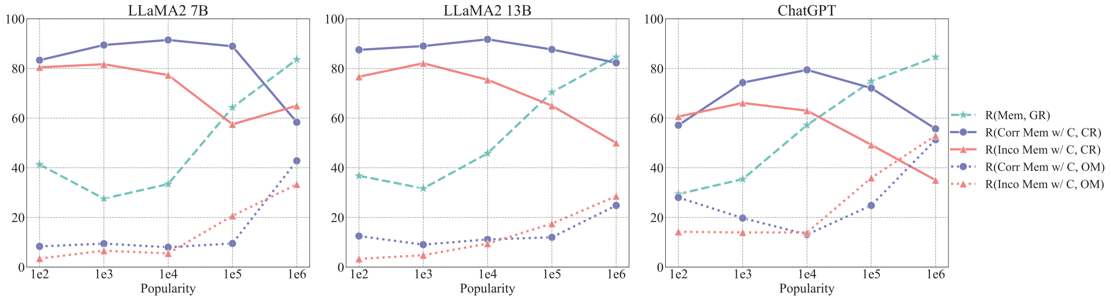

Figure 3: Recall under knowledge conflicts with various entity popularities on PopQA. GR denotes the gold references, CR denotes the conflicting references, and OM denotes the internal memory predictions.

its confidence score when encountering conflicts, indicating that FLAN-T5 lacks confidence in internal beliefs and prefers to trust external evidence. On the contrary, we observe that LLaMA2 significantly improves confidence scores only when modifying its incorrect internal memory based on truthful evidence. This proves that as the model's capabilities improve, *the model has a certain ability to perceive conflicts and will not blindly trust conflicting evidence*. However, if the model is not calibrated well to resolve conflicts, it will still answer incorrectly.

The Dunning-Kruger effect in human psychology merges in capable LLMs. Based on the above findings, a natural question arises: *is the model more confident on its correct internal memory or its incorrect internal memory?* As shown in Table 1, we are surprised to discover that with the increasing capabilities of the model, the memorization ratio for incorrect beliefs gradually surpasses that of correct beliefs. For instance, less capable LLMs like FLAN-T5-XL and FLAN-T5-XXL seldom adhere to their incorrect memory, whereas more capable LLMs like ChatGPT frequently rely on their incorrect internal memory for over half of the time.

We argue that *there is a phenomenon emerging among those **capable** models: they may lack confidence in their correct internal memory, but will stubbornly persist in their incorrect internal memory.* This phenomenon aligns with the **Dunning-Kruger effect** in human psychology (Kruger and Dunning, 1999; Dunning, 2011; Singh et al., 2023), wherein individuals with limited competence in a specific domain tend to overestimate their abilities. As depicted in Figure 2, when the model preserves incorrect memory despite conflicting evidence, there are notably high confidence scores present. This also proves that the model is overconfident in some

Model	EM	F1	R	Con R ↓	Tru KP	Mis KP ↓	Irr KP ↓	Corr MR	Inco MR
FLAN-T5-XL	52.55	62.81	66.95	-	84.00	-	62.70	62.81	3.45
w/ Conflict	30.23	39.84	43.98	49.70	62.00	68.50	33.01	45.45	3.32
FLAN-T5-XXL	52.30	63.83	70.00	-	85.86	-	61.34	65.24	4.19
w/ Conflict	29.34	39.07	43.40	52.54	61.22	70.72	31.83	47.56	4.84
FLAN-UL2	58.55	68.42	71.04	-	85.66	-	60.23	75.11	3.81
w/ Conflict	**31.51**	39.35	41.76	53.12	58.34	71.40	**30.71**	48.93	5.99
Baichuan2 7B	43.75	55.49	60.26	-	79.70	-	62.57	67.14	14.88
w/ Conflict	30.87	**41.25**	45.71	**44.32**	64.61	**62.47**	35.88	51.07	12.50
Baichuan2 13B	48.09	59.81	64.96	-	81.73	-	60.30	72.32	15.18
w/ Conflict	27.68	37.28	42.60	47.75	61.33	65.91	33.19	46.73	14.06
LLaMA2 7B	48.21	59.00	61.94	-	80.50	-	62.29	69.78	16.20
w/ Conflict	30.36	39.60	42.59	49.60	61.40	66.43	34.25	49.84	14.90
LLaMA2 13B	45.28	56.79	62.99	-	82.42	-	62.78	71.10	19.59
w/ Conflict	29.85	39.93	46.14	45.73	65.85	65.41	35.73	50.13	16.28
ChatGPT	20.66	36.93	72.00	-	78.98	-	69.94	84.96	60.00
w/ Conflict	16.58	31.07	**61.05**	46.65	**70.82**	67.75	43.41	**69.21**	**51.78**

Table 2: Experimental results under knowledge conflicts between truthful, irrelevant and misleading evidence on NQ. Con R is the Recall between the predictions and conflicting references. Tru KP, Mis KP and Irr KP represent the K-Precision of truthful, misleading and irrelevant evidence. Underline means the highest or lowest (↓) results without conflicts. Bold means the highest or lowest (↓) results under conflicts.

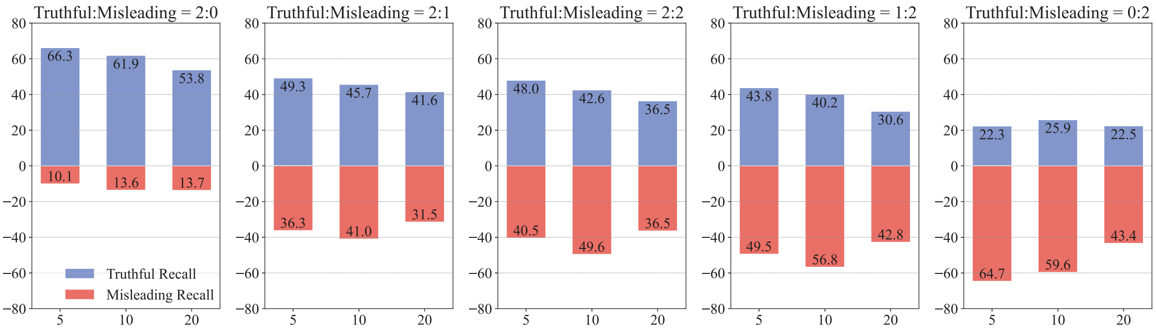

Figure 4: Recall of LLaMA2 7B with various amounts of evidence and different conflicting ratios on NQ.

incorrect internal memory (Slobodkin et al., 2023).

RALMs exhibit an availability bias towards common knowledge. We aim to explore whether the model's preference varies when faced with knowledge conflicts of different popularities. We conduct experiments on entity-centric PopQA with a knowledge popularity distribution ranging from 1e2 to 1e6. As depicted in Figure 3, for those long-tail knowledge, the model depends on external sources in most cases. However, as the knowledge popularity increases, the model begins to lean towards trusting its internal memory. This indicates when facing conflicts, the model exhibits an availability bias towards commonly known knowledge, more willing to believe in knowledge that they can easily recall.

5. Conflicts Between Truthful, Irrelevant and Misleading Evidence

In this section, we investigate knowledge conflicts between truthful, irrelevant and misleading evidence. First, we provide the model with $K = 10$ evidence (including truthful and irrelevant evidence) and then prompt it to perform open-book QA. Then, we construct conflicts by introducing misleading evidence, ensuring that the number of misleading evidence equals that of truthful evidence.

RALMs often struggle to discern truthful evidence from misleading evidence. As shown in Table 2, after encountering conflicts, the correctness of RALMs will decline to a certain extent. Among them, the Recall of FLAN-UL2 drops the most (↓ 29.28%), and the Recall of ChatGPT drops the least (↓ 10.95%). This also confirms our previous argument that the more capable model possesses a certain ability to perceive conflicts. However, for most RALMs, their K-Precision for misleading evidence is higher than that for truthful evidence, indicating that *they still have difficulty distinguishing truthful evidence from misleading evidence.* Furthermore, we also observe that they can be easily distracted by irrelevant evidence.

RALMs follow the principle of majority rule. We aim to delve deeper into the criteria the model employs when selecting evidence as the reference for generating answers. We provide LLaMA2 7B with different amounts of evidence ($K \in \{5, 10, 20\}$) and set different conflicting ratios (truthful evidence: misleading evidence $\in \{2:0, 2:1, 2:2, 1:2, 0:2\}$). As depicted in Figure 4, the model leans towards placing trust in evidence that appears more frequently. Besides, we find that *more external evidence is not necessarily better*. Excessively long context may cause the model to lose focus, making it challenging to provide accurate answers.

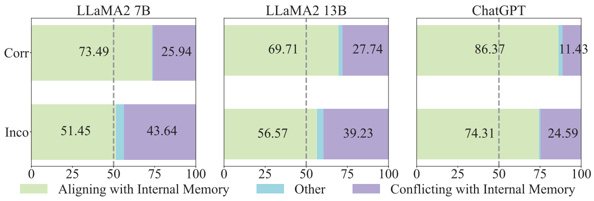

Figure 5: Frequency of choosing evidence aligning with or conflicting with internal memory on NQ.

RALMs exhibit confirmation bias when partial external evidence aligns with their internal memory. Furthermore, we consider a more interesting scenario if external sources contain evidence supporting the model's internal memory. For example, if the model has incorrect internal memory, then we induce the model's internal memory and add it to external sources as misleading evidence. Instead, we induce the model's correct internal memory as truthful evidence. As depicted in Figure 5, RALMs exhibit **confirmation bias** (Nickerson, 1998), *displaying a stronger inclination to select evidence that aligns with their own internal memory.* This is consistent with the observations of Xie et al. (2023). Besides, we also find that *as model sizes and capabilities expand, the model gains greater confidence in its internal memory, especially when the internal memory is provided as external evidence*, which further supports our findings in Section 4.

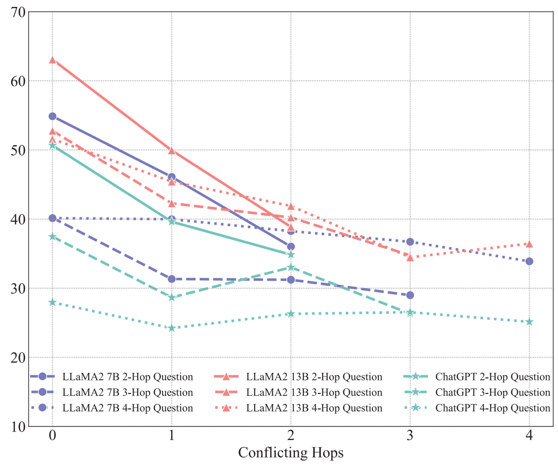

Figure 6: Recall of different models with various conflicting hops on MuSiQue.

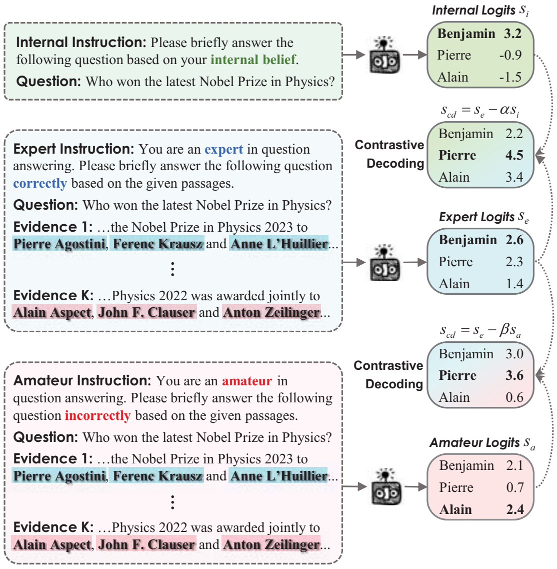
Internal Logits s_i
Internal Instruction: Please briefly answer the following question based on your internal belief.
Question: Who won the latest Nobel Prize in Physics?
Benjamin 3.2
Pierre -0.9
Alain -1.5
$s_{cd} = s_e - \alpha s_i$
Contrastive Decoding
Benjamin 2.2
Pierre 4.5
Alain 3.4
Expert Instruction: You are an expert in question answering. Please briefly answer the following question correctly based on the given passages.
Question: Who won the latest Nobel Prize in Physics?
Evidence 1: …the Nobel Prize in Physics 2023 to Pierre Agostini, Ferenc Krausz and Anne L'Huillier…
⋮
Evidence K: …Physics 2022 was awarded jointly to Alain Aspect, John F. Clauser and Anton Zeilinger…
Expert Logits s_e
Benjamin 2.6
Pierre 2.3
Alain 1.4
$s_{cd} = s_e - \beta s_a$
Contrastive Decoding
Benjamin 3.0
Pierre 3.6
Alain 0.6
Amateur Instruction: You are an amateur in question answering. Please briefly answer the following question incorrectly based on the given passages.
Question: Who won the latest Nobel Prize in Physics?
Evidence 1: …the Nobel Prize in Physics 2023 to Pierre Agostini, Ferenc Krausz and Anne L'Huillier…
⋮
Evidence K: …Physics 2022 was awarded jointly to Alain Aspect, John F. Clauser and Anton Zeilinger…
Amateur Logits s_a
Benjamin 2.1
Pierre 0.7
Alain 2.4

Figure 7: An illustration of Conflict-Disentangle Contrastive Decoding method.

As the number of conflicting hops increases, the model faces greater difficulty in reasoning. To investigate knowledge conflicts in multi-hop reasoning scenarios, we conduct experiments on MuSiQue. For multi-hop questions, the model needs to reason based on multiple pieces of evidence. Therefore, we incrementally raise the count of misleading evidence until it matches the count of truthful evidence. As illustrated in Figure 6, it is evident that *an increase in the number of conflicting hops hinders the model's ability to reason accurately.* Moreover, we also find that ChatGPT does not perform well in scenarios involving multi-hop reasoning with multiple pieces of evidence.

6. Conflict-Disentangle Contrastive Decoding

Based on our investigations above, we believe that a capable RALM has a certain ability to **perceive** conflicts. However, the model still struggles to answer correctly as it lacks proper calibration to **resolve** conflicts. As shown in Figure 7, after providing the external sources, the logit of token "*Pierre*" increases significantly, while the logit of token "*Benjamin*" decreases. Although the knowledge conflict does lead to a shift in the model's confidence, it still tends to produce an inaccurate answer because of the strong influence of its incorrect internal memory. Therefore, we focus on better calibrating the RALM's confidence to resolve knowledge conflicts.

6.1. Method

Contrastive decoding (Li et al., 2023; Shi et al., 2023a), a search objective for generating fluent and

Method	NQ-Conf						TriviaQA-Conf					
	EM	F1	R	Tru KP	Mis KP ↓	Irr KP ↓	EM	F1	R	Tru KP	Mis KP ↓	Irr KP ↓
Closed-book	31.38	41.86	44.88	-	-	-	65.42	68.89	70.14	-	-	-
In-context	30.36	39.60	42.59	61.40	66.43	34.25	55.38	59.57	60.87	73.36	54.14	19.65
Finetune$_{\text{Orig}}$	46.05	54.34	55.50	72.68	55.36	33.29	64.56	69.96	71.54	82.87	45.81	**19.61**
Finetune$_{\text{Conf}}$	58.67	68.17	69.77	83.06	40.04	33.51	74.18	80.19	81.81	91.67	32.74	23.29
CD^2 $\beta = 0.5$	**61.35**	**70.83**	**72.42**	**85.15**	**39.62**	**32.55**	**77.76**	**82.47**	**83.93**	**93.03**	**31.60**	23.51

Table 3: Performance of LLaMA2 7B on NQ-Conf and TriviaQA-Conf.

Dataset	Method	EM	F1	R
NQ-Inco	In-context	31.55	41.87	43.63
	+ CD^2 $\alpha = 0.3$	**32.29**	**43.07**	44.98
	+ CD^2 $\alpha = 0.5$	32.10	43.04	**45.68**
	+ CD^2 $\alpha = 0.7$	31.73	42.19	44.36
TriviaQA-Inco	In-context	51.38	55.57	56.61
	+ CD^2 $\alpha = 0.3$	**52.41**	57.46	58.97
	+ CD^2 $\alpha = 0.5$	**52.41**	**57.72**	**59.31**
	+ CD^2 $\alpha = 0.7$	50.34	56.13	58.07

Table 4: Performance of LLaMA2 7B on NQ-Inco and TriviaQA-Inco.

diverse text, aims to search for the higher-quality output token that maximizes the difference between expert (positive) logits and amateur (negative) logits. Inspired by this, we propose a novel method called **C**onflict-**D**isentangle **C**ontrastive **D**ecoding (**CD^2**) to maximize the difference between various logits under knowledge conflicts and calibrate the model's confidence shown in Figure 7.

For knowledge conflicts between internal memory and external sources, to mitigate the impact of the RALM's incorrect internal memory, we predict the output token y_t by amplifying the difference between expert logits with external sources s_e and internal logits without external sources s_i:

$$y_t = \operatorname{argmax}\left(s_e\left(y_t \mid d, x, y_{<t}\right) - \alpha s_i\left(y_t \mid x, y_{<t}\right)\right),$$

where d denotes the external sources, x denotes the question, y denotes the answer. Our method can be applied to LLMs like LLaMA2 during the inference stage without additional training.

For knowledge conflicts between truthful, irrelevant and misleading evidence, we first adopt **fact-aware** instruction tuning to enable the RALM to be aware of truthful and misleading evidence in retrieved external sources. As shown in Figure 7, we train an **expert** LM to generate truthful answers normally, and an **amateur** LM to generate misleading answers deliberately. Then, we predict the output token y_t by maximizing the difference between expert logits s_e and amateur logits s_a:

$$y_t = \operatorname{argmax}\left(s_e\left(y_t \mid d, x, y_{<t}\right) - \beta s_a\left(y_t \mid d, x, y_{<t}\right)\right).$$

6.2. Experiment Setups

We employ LLaMA2 7B as the backbone model. We conduct experiments on two widely used QA datasets: NQ and TriviaQA. For knowledge conflicts between internal memory and external sources, we create NQ-Inco and TriviaQA-Inco datasets, including questions that the model's internal memory answers incorrectly. For knowledge conflicts between truthful, irrelevant and misleading evidence, we reconstruct NQ-Conf and TriviaQA-Conf datasets following the setting in Section 5. We provide the model with $K = 10$ evidence. Then we compare our CD^2 with the following baselines: (1) Closed book: We do not provide external sources and directly let the model answer the question; (2) In-context: We let the model answer the question based on external sources; (3) Finetune$_{\text{Orig}}$: We finetune the model on the original training set; (4) Finetune$_{\text{Conf}}$: We finetune the model on the conflict-enhanced training set (*i.e.*, knowledge conflicts exist in the external sources). Our CD^2 also adopts fact-aware instruction tuning on the conflict-enhanced training set. For all methods that require training, we apply LoRA (Hu et al., 2022) to finetune the model on NQ with 3,000 training examples. We set learning rate=5e-5, epoch=5, batch size=4, lora rank=8 and lora alpha=16. All experiments are conducted with NVIDIA RTX A6000 GPUs.

6.3. Results

On the one hand, as shown in Table 4, CD^2 can be directly applied to LLaMA2 7B without updating model parameters. We can observe that CD^2 leads to a 2.05% and 2.70% Recall improvement on NQ-Inco and TriviaQA-Inco respectively. This highlights the efficacy of our method in mitigating the model's incorrect internal memory and enhancing its attention towards external sources. On the other hand, as shown in Table 3, CD^2 enables the model to effectively distinguish truthful evidence from misleading evidence. Compared with In-context, CD^2 achieves a notable 29.83% and 23.06% Recall improvement on NQ-Conf and TriviaQA-Conf, while Mis KP decreases significantly.

7. Conclusion

In this paper, we focus on exploring and resolving knowledge conflicts in RALMs. First, we present an evaluation framework for assessing knowledge conflicts across various dimensions. Then, we investigate the behavior and preference of RALMs from the following two perspectives: (1) Conflicts between internal memory and external sources: We find that capable RALMs emerge with the **Dunning-Kruger effect**, persistently favoring their incorrect internal memory even when correct evidence is provided. Moreover, RALMs exhibit an **availability bias** towards commonly known knowledge; (2) Conflicts between truthful, irrelevant and misleading evidence: We reveal that RALMs follow the principle of **majority rule**, preferring evidence that appears more frequently. Furthermore, we find that RALMs exhibit **confirmation bias**, and are more willing to choose evidence that is consistent with their internal memory. To address the issue of knowledge conflicts, we propose a method called **C**onflict-**D**isentangle **C**ontrastive **D**ecoding (**CD^2**) to better calibrate the model's confidence. Experimental results demonstrate that CD^2 can effectively resolve knowledge conflicts. We hope our work can provide useful insights for further research.

Limitations

For further study, we conclude some limitations of our work as follows:

- We mainly investigate knowledge conflicts in RALMs from two perspectives: model performance and confidence score. In the future, we will delve into exploring the model's mechanisms to understand the issue of knowledge conflicts. For example, we can analyze whether knowledge conflicts will inhibit or stimulate the activation values of certain neurons.
- In this paper, we resolve the knowledge conflicts in RALMs by calibrating the open-source model's logits. How to mitigate the knowledge conflicts in black-box LLMs which are unable to access output logits remains to be studied.

We hope that our findings will better promote the practical application of RALMs.

Acknowledgements

This work is supported by the National Key Research and Development Program of China (No.2020AAA0106400), the National Natural Science Foundation of China (No.62176257). This work is also supported by the Strategic Priority Research Program of Chinese Academy of Sciences (Grant No.XDA27020100), the Youth Innovation Promotion Association CAS, and Yunnan Provincial Major Science and Technology Special Plan Projects (No.202202AD080004).

Bibliographical References

Vaibhav Adlakha, Parishad BehnamGhader, Xing Han Lu, Nicholas Meade, and Siva Reddy. 2023. Evaluating correctness and faithfulness of instruction-following models for question answering. *arXiv preprint arXiv:2307.16877*.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego De Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack Rae, Erich Elsen, and Laurent Sifre. 2022a. Improving language models by retrieving from trillions of tokens. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 2206–2240. PMLR.

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. 2022b. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pages 2206–2240. PMLR.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Hung-Ting Chen, Michael Zhang, and Eunsol Choi. 2022. Rich knowledge sources bring complex knowledge conflicts: Recalibrating models to reflect conflicting evidence. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2292–2307, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2023a. Benchmarking large language models in retrieval-augmented generation. *arXiv preprint arXiv:2309.01431*.

Zeming Chen, Qiyue Gao, Antoine Bosselut, Ashish Sabharwal, and Kyle Richardson. 2023b. DISCO: Distilling counterfactuals with large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5514–5528, Toronto, Canada. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

David Dunning. 2011. The dunning–kruger effect: On being ignorant of one's own ignorance. In *Advances in experimental social psychology*, volume 44, pages 247–296. Elsevier.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880, Online. Association for Computational Linguistics.

Joel Jang, Seonghyeon Ye, Changho Lee, Sohee Yang, Joongbo Shin, Janghoon Han, Gyeonghun Kim, and Minjoon Seo. 2022. TemporalWiki: A lifelong benchmark for training and evaluating ever-evolving language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6237–6250, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Zhuoran Jin, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. 2023. Instructor: Instructing unsupervised conversational dense retrieval with large language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

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

Justin Kruger and David Dunning. 1999. Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of personality and social psychology*, 77(6):1121.

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

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2023. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12286–12312, Toronto, Canada. Association for Computational Linguistics.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7052–7063, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ella Neeman, Roee Aharoni, Or Honovich, Leshem Choshen, Idan Szpektor, and Omri Abend. 2023. DisentQA: Disentangling parametric and contextual knowledge with counterfactual question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10056–10070, Toronto, Canada. Association for Computational Linguistics.

Raymond S Nickerson. 1998. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, 2(2):175–220.

OpenAI. 2023. Gpt-4 technical report.

Cheng Qian, Xinran Zhao, and Sherry Tongshuang Wu. 2023. " merge conflicts!" exploring the impacts of external distractors to parametric knowledge graphs. *arXiv preprint arXiv:2309.08594*.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *arXiv preprint arXiv:2302.00083*.

Ruiyang Ren, Yuhao Wang, Yingqi Qu, Wayne Xin Zhao, Jing Liu, Hao Tian, Hua Wu, Ji-Rong Wen, and Haifeng Wang. 2023. Investigating the factual knowledge boundary of large language models with retrieval augmentation. *arXiv preprint arXiv:2307.11019*.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Scott Wen-tau Yih. 2023a. Trusting your evidence: Hallucinate less with context-aware decoding. *arXiv preprint arXiv:2305.14739*.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2023b. Replug: Retrieval-augmented black-box language models. *arXiv preprint arXiv:2301.12652*.

Aniket Kumar Singh, Suman Devkota, Bishal Lamichhane, Uttam Dhakal, and Chandra Dhakal. 2023. The confidence-competence gap in large language models: A cognitive study. *arXiv preprint arXiv:2309.16145*.

Aviv Slobodkin, Omer Goldman, Avi Caciularu, Ido Dagan, and Shauli Ravfogel. 2023. The curious case of hallucinatory unanswerablity: Finding truths in the hidden states of over-confident large language models. *arXiv preprint arXiv:2310.11877*.

Kai Sun, Yifan Ethan Xu, Hanwen Zha, Yue Liu, and Xin Luna Dong. 2023. Head-to-tail: How knowledgeable are large language models (llm)? aka will llms replace knowledge graphs? *arXiv preprint arXiv:2308.10168*.

Yi Tay, Mostafa Dehghani, Vinh Q. Tran, Xavier Garcia, Jason Wei, Xuezhi Wang, Hyung Won Chung, Dara Bahri, Tal Schuster, Steven Zheng, Denny Zhou, Neil Houlsby, and Donald Metzler. 2023. UL2: Unifying language learning paradigms. In *The Eleventh International Conference on Learning Representations*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Yike Wang, Shangbin Feng, Heng Wang, Weijia Shi, Vidhisha Balachandran, Tianxing He, and Yulia Tsvetkov. 2023. Resolving knowledge conflicts in large language models. *arXiv preprint arXiv:2310.00935*.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2023. Adaptive chameleon or stubborn sloth: Unraveling the behavior of large language models in knowledge conflicts. *arXiv preprint arXiv:2305.13300*.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, Fan Yang, et al. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Wenhao Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, Soumya Sanyal, Chenguang Zhu, Michael Zeng, and Meng Jiang. 2023. Generate rather than retrieve: Large language models are strong context generators. In *The Eleventh International Conference on Learning Representations*.

Zeyuan Allen Zhu and Yuanzhi Li. 2023. Physics of language models: Part 3.1, knowledge storage and extraction.

Language Resource References

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

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*, 7:453–466.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822, Toronto, Canada. Association for Computational Linguistics.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. 2021. KILT: a benchmark for knowledge intensive language tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2523–2544, Online. Association for Computational Linguistics.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. Musique: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554.