

# EVER: Mitigating Hallucination in Large Language Models through Generation-Time Verification and Rectification

Anonymous ACL submission

## Abstract

Large Language Models (LLMs) have demonstrated remarkable proficiency in generating fluent text. However, they often encounter the challenge of generating inaccurate or hallucinated content. This issue is common in both non-retrieval-based generation and retrieval-augmented generation approaches, and existing post-hoc rectification methods may not address the accumulated hallucination errors that may be caused by the "snowballing" issue, especially in reasoning tasks. To tackle these challenges, we introduce a novel approach called Generation-Time Verification and Rectification (EVER). Instead of waiting until the end of the generation process to rectify hallucinations, EVER employs a generation-time, step-wise generation and hallucination rectification strategy. Apart from directly mitigating hallucination, we further demonstrate that both the EVER-rectified response and the original one can serve as preference data to enhance the factuality of the model through preference tuning. When compared to both retrieval-based and non-retrieval-based baselines, EVER demonstrates a significant improvement in generating trustworthy and factually accurate text across a diverse range of tasks, including biography generation and multi-hop reasoning.

## 1 Introduction

Recent years have witnessed remarkable progress in the field of Large Language Models (LLMs), which are increasingly adept at generating coherent, contextually fluent responses. Despite this, they are still prone to hallucination which is defined as the generated content is nonsensical or unfaithful to a reference content (Ji et al., 2023; Zhang et al., 2023b). Hallucination can be categorized into two types: intrinsic and extrinsic. Intrinsic hallucinations happen when the generated content is contradictory to the reference. Extrinsic hallucinations, meanwhile, are the content that, while seemingly plausible, cannot be verified by evidence, typically

appearing as imaginative concoctions or guesses made by the model (Min et al., 2023; Sun et al., 2023; Kandpal et al., 2023).

Due to the infrequent updates of an LLM’s parametric knowledge base, utilizing external knowledge has shown significant leap in enhancing factuality by providing up-to-date content (Lewis et al., 2020). Prior retrieval-based mitigation methods of LLM hallucination can be categorized into two categories: pre-generation, and post-generation methods. The pre-generation methods (Lewis et al., 2020; Vu et al., 2023; Asai et al., 2023) optimize the retrieved content to be more accurate, relevant and supportive. But these methods may still produce detailed factual errors, particularly in long-form generation if there is no mechanism for post-generation checks or revisions. Another line of work focuses on enhancing the attribution of text post-generation (Gao et al., 2022; Gou et al., 2023; Peng et al., 2023). However, these post-hoc editing methods do not account for the "snowballing" issue of hallucinations (Zhang et al., 2023a), where initial factual errors can lead to a series of accumulated errors, and they require increasingly complex revisions to mitigate its impact.

To address these challenges, we propose the REal-Time VERification and RECTification (EVER) framework. Instead of mitigating hallucination until the end of generation, EVER employs generation-time validation to identify both intrinsic and extrinsic hallucinations, mitigating these issues during the generation process to prevent error propagation. The process involves three stages: generation, validation, and rectification. First, a LLM generates an initial sentence based on a prompt, which may include externally retrieved knowledge, such as Retrieval-Augmented Generation (RAG) (Lewis et al., 2020). Then, it validates the correctness of each fact-related concept in the sentence by identifying intrinsic and extrinsic hallucinations. In the rectification stage, any detected errors are corrected

043  
044  
045  
046  
047  
048  
049  
050  
051  
052  
053  
054  
055  
056  
057  
058  
059  
060  
061  
062  
063  
064  
065  
066  
067  
068  
069  
070  
071  
072  
073  
074  
075  
076  
077  
078  
079  
080  
081  
082  
083

084 based on the type of hallucinations identified. The  
085 rectified sentence then undergoes another round of  
086 validation. If extrinsic hallucinations persist, de-  
087 pending on the task, we either flag the sentence  
088 with a warning to alert users to potential issues  
089 or abstain from answering the question, which en-  
090 hances the trustworthiness of the generated content.  
091 In addition to directly mitigating hallucination dur-  
092 ing the generation process, we further explore the  
093 utilization of the EVER-generated data to construct  
094 preference data pairs, essentially enhancing the fac-  
095 tuality of the model through preference tuning.

096 Our primary contribution of this paper is EVER,  
097 which introduces a novel approach to mitigate hal-  
098 lucinations in LLM. Compared to the state-of-the-  
099 art prior methods, our results demonstrate the ef-  
100 fectiveness of this approach in directly reducing  
101 hallucinations in two tasks: long-form biography  
102 generation and reasoning. Furthermore, we show  
103 the compatibility of EVER, which can serve as a  
104 complement to the traditional RAG method. Lastly,  
105 we demonstrate that EVER-rectified response can  
106 lead to better preference data and enhance the fac-  
107 tuality of LLM with preference tuning.

## 108 2 Generation-Time Verification and 109 Rectification

110 In this section, we firstly detail our method,  
111 GEneration-time VERification and Rectification  
112 (EVER), whose framework with one representa-  
113 tive example is shown in Figure 1. EVER aims to  
114 mitigate hallucinations in language model outputs  
115 by immediately validating each generated sentence  
116 during the generation period, which helps prevent  
117 error propagation. Secondly, as shown in Figure 2,  
118 we use EVER-rectified response to construct bet-  
119 ter preference data to align LLM to become more  
120 factual by using preference tuning.

### 121 2.1 Prompting-based Hallucination 122 Verification and Mitigation

123 We first present how to use EVER to directly mit-  
124 igate hallucination of LLM during the response  
125 generation period.

#### 126 2.1.1 Generation

127 The first stage is to generate the initial sentence  
128 given the prompt. Based on if an external knowl-  
129 edge is used in the prompt, we categorize the gen-  
130 eration method to two categories:

- 131 • **Non-retrieval Generation:** In non-retrieval gen-  
132 eration, the LLM is provided with a query and

133 is prompted to generate a response based solely  
134 on its internal knowledge without referring to  
135 external data sources.

- **Retrieval-Augmented Generation (RAG):** In  
136 RAG (Lewis et al., 2020), the LLM is presented  
137 with the context in the prompt. 138

139 After determining the generation category in  
140 EVER, we adopt a generation-time verification and  
141 rectification strategy to mitigate the "snowballing  
142 issue" in text generation (Zhang et al., 2023a;  
143 Varshney et al., 2023). This effect arises when early  
144 inaccuracies or hallucinations in the text result in  
145 compounded errors in subsequent sentences. By  
146 addressing hallucinations on a generation-time ba-  
147 sis, our strategy significantly reduces the likelihood  
148 of errors propagating throughout the entire text, en-  
149 suring that early hallucinations do not have a signif-  
150 icant impact on later generated content. Therefore,  
151 we transition to the validation and hallucination  
152 correction phases upon generating a new sentence.

#### 153 2.1.2 Concept-Level Validation

154 In the validation stage, we evaluate the generated  
155 sentence at a concept-level, with the goal of identi-  
156 fying the occurrence of hallucinations and classifi-  
157 cing them as either intrinsic or extrinsic halluci-  
158 nations. The entire validation phase includes three  
159 steps: key concepts identification, validation ques-  
160 tion generation, and support checking. We detail  
161 these steps as follows:

162 **Key Concepts Identification.** In key concepts  
163 identification step, we leverage the in-context learn-  
164 ing ability of the model to extract factual-related  
165 concepts from the generated sentence, such as  
166 dates, numbers, jobs, locations, etc. For exam-  
167 ple, as shown in Figure 1, in the sentence "Shin  
168 Jea-hwan is an artistic gymnast, born on Novem-  
169 ber 2, 1998, and has raised by a family of travel-  
170 ing circus performers.", we extract the concepts of  
171 "artistic gymnast", "November 2, 1998", and "trav-  
172 eling circus performers". Additionally, EVER is  
173 able to select customized check-worthy concepts,  
174 where check-worthiness is a classical challenge for  
175 fact-checking frameworks (Wright and Augenstein,  
176 2020; Abumansour and Zubiaga, 2023). This al-  
177 lows for a more targeted and efficient fact-checking  
178 process by focusing on the most critical and poten-  
179 tially erroneous information.

180 **Validation Question Generation.** Once the key  
181 concepts are identified, we will use the model to  
182 generate validation questions. These validation

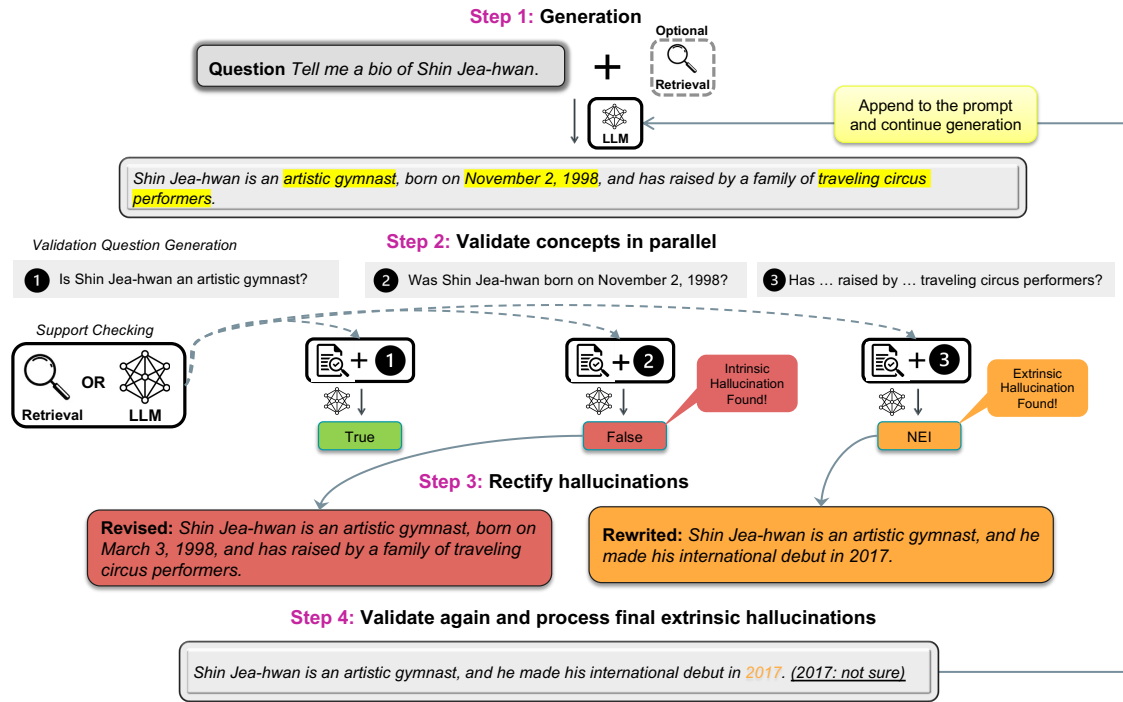


Figure 1: Overview of EVER pipeline in the biography generation task. EVER proactively identifies and rectifies concept-level hallucinations before each new sentence generation. Also, it flags any remaining extrinsic hallucinations after a single round of rectification, thereby enhancing the trustworthiness of the output.

183 questions are Yes/No questions constructed to verify the accuracy of the concepts in the initial sentence. For example, in Figure 1, for the extracted concept of "artistic gymnast", the corresponding validation question is "Is Shin Jea-hwan an artistic gymnast?"

184  
185  
186  
187  
188  
189 **Support Checking.** Then, in the last step, we use few-shot Chain of Thought (CoT) prompting (Wei et al., 2022) to guide the model to choose one of three flags for each validation question based on the evidence: True, False, or Not Enough Information (NEI). A True flag indicates that the evidence supports the generated concept, whereas a False flag signifies that the generated concept is in contradiction with the evidence, pointing towards an intrinsic hallucination. The NEI flag is assigned when no related evidence is found, suggesting the presence of an extrinsic hallucination. To compare the effect of retrieval on our method, we test on the following two strategies.

- 203 • **Self-query:** Based on the validation question, we prompt the LLM to directly answer the question by choosing from the three labels.
- 204
- 205
- 206 • **Evidence Retrieval:** This mode leverages external knowledge source to gather evidence that can help answer the validation question.
- 207
- 208

### 2.1.3 Rectifying Hallucination

209  
210  
211  
212  
213  
214  
215  
216  
217  
218  
219  
220  
221  
222  
223  
224  
225  
226  
227  
228

After the validation stage, if hallucination is detected, i.e., at least one validation question is assigned the flag False or NEI, EVER aims to rectify the corresponding sentence based on the evidence gathered, including two revision categories:

**Intrinsic Hallucination Revision.** Intrinsic Hallucinations refer to instances where the generated output contradicts the source content. These hallucination will be revised based on the evidence retrieved from last step. The primary objective is to align each entity or fact with verifiable truths.

**Extrinsic Hallucination Rewrite.** Extrinsic Hallucinations are defined as generated outputs that cannot be verified against the source content, meaning the output is neither supported nor refuted by the evidence. When confronted with such situations, the entire sentence undergoes a rewrite, taking into account feedback that pinpoints the issue and uses the retrieved evidence as a reference.

### 2.1.4 Processing the Remaining Extrinsic Hallucination

229  
230  
231  
232  
233  
234  
235

After completing the rectification phase, the refined sentence undergoes revalidation. If intrinsic hallucinations cannot be fully rectified with a single round of rectification, we conduct additional rounds. It's important to note that, in most scenar-

ios, one round of rectification is empirically sufficient to eliminate all intrinsic hallucinations (see detailed analysis in Appendix C). In such cases, if a sentence still exhibits extrinsic hallucinations, depending on the tasks, we will further refine it. For example, in short-form generation, if there is no other verified correct answers, we will abstain from answering the question to maintain honesty. In long-form generation, we will mark it with a final warning flag, "not sure," indicating the presence of extrinsic hallucination and enhancing the trustworthiness of the generated content. Acknowledging limitations and errors in generated content promotes transparency and a reliable user experience. Since completely rectifying all extrinsic hallucinations can be challenging, the warning signal effectively assists users in utilizing the generated content.

## 2.2 Enhancing Factuality of Model via Preference Tuning

In addition to directly rectify hallucination during the generation period, we extend the EVER framework to create better preference data to essentially enhance the factuality LLM by preference tuning. Here, as illustrated in Figure 2, the EVER-generated response  $y_{ever}$  can be naturally served as preferred response and the non-rectified response is used as dispreferred response. Formally, the preference data is defined as  $\mathcal{D} = \{x^{(i)}, y_w^{(i)}, y_l^{(i)}\}_{i=1}^N$ , where  $y_w^{(i)} = y_{ever}^{(i)}$  and  $y_l^{(i)}$  represent preferred and dispreferred responses given an input prompt  $x^{(i)}$ . Such preference data are then used to perform preference tuning by using direct preference optimization (DPO) (Rafailov et al., 2023), which are detailed as follows:

Specifically, in large language model, we first define a language model policy  $\pi_\theta$ , which can produce the response  $y$  with a conditional distribution  $\pi_\theta(y | x)$ . For each input prompt  $x$  and response  $y$ , we define a reward function  $r(x, y)$  to measure the generation quality of  $y$ . Our goal here is to maximize the average reward of outputs generated by the language model policy. Following a Bradley-Terry model (Bradley and Terry, 1952), DPO obtain each preference pair with the probability  $p(y_w \succ y_l)$ , which defined as:

$$p(y_w \succ y_l) = \sigma(r(x, y_w) - r(x, y_l)), \quad (1)$$

where  $\sigma(\cdot)$  is defined as a sigmoid function. Then, DPO achieves the maximum average reward by

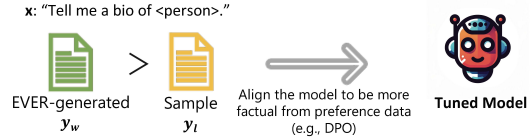


Figure 2: Illustration of using EVER-generated data to construct preference data pairs, which are then used to further finetune the model to enhance the factuality of model.

optimizing the following classification loss over the preference data as:

$$\mathcal{L}(\pi_\theta, \mathcal{D}) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[ \log \sigma \left( \alpha \log \frac{\pi_\theta(y_w | x)}{\pi_{\text{ref}}(y_w | x)} - \alpha \log \frac{\pi_\theta(y_l | x)}{\pi_{\text{ref}}(y_l | x)} \right) \right], \quad (2)$$

where  $\pi_{\text{ref}}(y | x)$  is defined as the reference policy, typically referring to the result after performing supervised fine-tuning.

## 3 Experiment of Prompting-based Hallucination Rectification

In this section, we evaluate the performance of EVER on three tasks, short-form QA (Appendix A), biography generation and reasoning, aiming to answer the following questions: (1) Can EVER effectively address the challenges we’ve identified for RAG and post-hoc edit methods? (2) Can EVER effectively reduce hallucination of LLMs compared to other baselines across different tasks? (3) Can EVER effectively increase the trustworthiness of generated texts? In practice, we apply one of the following variant of EVER based on different application scenarios:

- **EVER (NRG+SQ)**: The first variant is a non-retrieval method that involves non-retrieval sentence generation (NRG) combined with a self-query (SQ) approach during the support check in the validation phase.
- **EVER (NRG+ER)**: The second approach also employs a non-retrieval sentence generation approach, but it introduces evidence retrieval (ER) during the support check in the validation phase.
- **EVER (RAG+ER)**: The third variant enhances sentence generation with retrieval-augmented methods (RAG) and includes evidence retrieval during support checking.

### 3.1 Biography Generation Task

In this task, the LLM is prompted to generate factual long-form biographies (bio), where LLM

needs to ensure the accuracy of each atomic fact within the response.

### 3.1.1 Experimental Setup

**Dataset and Evaluation Metric.** We utilize the bio benchmark with 183 examples as proposed by (Min et al., 2023), our model is prompted with "Tell me a bio of <entity>." to generate a biography for a given entity. To evaluate the effectiveness of our method, we employ the FACTSCORE metric (Min et al., 2023). This metric leverages a retrieval-augmented language model ("ChatGPT + Retrieval"), for fact-checking the generated response, which has demonstrated that this metric aligns well with human evaluations. Furthermore, in line with other baseline settings, we retrieve evidence using Google Search.

**Evaluation Scenarios and Baselines.** We evaluate EVER in three scenarios: non-retrieval, retrieval-augmented rectification, and retrieval-augmented generation and rectification. Each scenario corresponds to a specific variant of EVER: EVER (NRG+SQ), EVER (NRG+ER), and EVER (RAG+ER), respectively.

In each scenario, we employ different baselines for evaluation. First, in the non-retrieval scenario, we compare EVER (NRG+SQ) with several models: 1) zero-shot generation models, including Llama 2 7B Chat, Llama 2 13B Chat (Touvron et al., 2023), InstructGPT (Ouyang et al., 2022), and GPT 3.5 Turbo; 2) a factuality-enhanced decoding method called Dola (Chuang et al., 2023); and 3) a chain of verification method called CoVE (Dhuliawala et al., 2023). Second, in the retrieval-based rectification scenario, we compare EVER (NRG+ER) with RRAR<sup>1</sup> (Gao et al., 2022). RRAR not only identifies attributions by using a search engine for outputs from various text generation models but also performs hallucination rectification. Third, for the RAG-like baselines, we compare EVER (RAG+ER) with vanilla RAG and Self-RAG (Asai et al., 2023). These models are trained to retrieve, generate, and critique to enhance the LLM’s output quality and factuality. Detailed descriptions of the baselines are discussed in Appendix E.

<sup>1</sup>While the original paper uses Bing Search and GPT-3, we adapted the code to match our experimental setup with Serper Google Search API and our chosen LLMs.

### 3.1.2 Results and Analysis

In Table 1, we report the performance on the biography generation task. Specifically, we have the following observations: first, compared with non-retrieval based scenario with retrieval based scenario, we observe that external knowledge retrieval significantly enhances the factuality of text generation. This trend indicates that retrieval mechanisms enrich the inherent knowledge of large language models with up-to-date and specific information, thereby improving the content’s accuracy.

Second, in comparison to other baselines of equivalent LLM scale, EVER exhibits superior performance in rectifying hallucinations across all scenarios, affirming the efficacy of its sentence-by-sentence generation, paired with generation-time verification and rectification. In particular, when retrieval is not utilized, EVER outperforms the post-hoc verification and revision method CoVe when applied to the same pretrained Llama 65B model. This effectiveness is further corroborated through a fine-grained comparison between EVER and RRAR. Here, we compare EVER and RRAR with respect to the rarity of the biography, as defined by the pageviews of their corresponding Wikipedia pages. The results in Figure 3 illustrate that, unlike RRAR, which cannot reduce hallucinations for more rare subjects, the sentence-by-sentence evidence retrieval validation in EVER maintains stable factual precision across varying rarities.

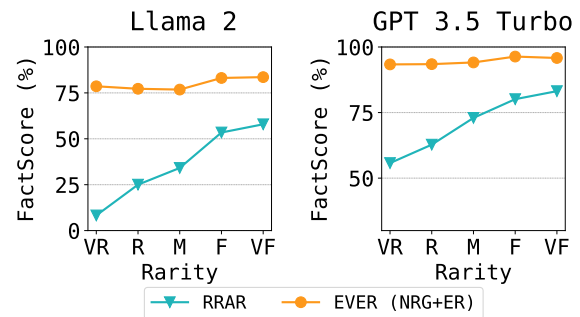


Figure 3: Comparison of our method and RRAR across examples with varying rarity distributions for the Llama 2 7B chat and GPT 3.5 Turbo models. "VR, R, M, F, VF" stands for "very rare", "rare", "medium", "frequent", and "very frequent", respectively.

Third, EVER could serve as a effective complementary method to the traditional retrieval-augmentation generation (RAG). Built upon traditional RAG, EVER (RAG+ER) demonstrates significant improvements over the conventional RAG approach. This demonstrates that EVER not only

effectively retrieves relevant information but also adeptly incorporates and refines this information within the generated content.

Table 1: Results on the biography generation task. "partial" indicates extracting and verifying the high-risk concepts. \*These numbers are from Asai et al. (2023). †We obtain the results from Dhuliawala et al. (2023). ‡The results are from Min et al. (2023)

LM Scale	Method	FACTSCORE (%)
<i>Non-Retrieval</i>		
InstructGPT	Zero-Shot <sup>‡</sup>	52.8
Llama 2 7B Chat	Zero-Shot	36.8
	Dola	36.8
	EVER (NRG+SQ)	46.7
Llama 2 13B Chat	Zero-Shot	40.3
	Dola	40.1
	EVER (NRG+SQ)	47.5
Llama 1 65B	Few-Shot <sup>†</sup>	55.9
	CoVe <sup>†</sup>	71.4
	EVER (NRG+SQ)	72.9
GPT 3.5 Turbo	Zero-Shot	71.8
	EVER (NRG+SQ) <sub>partial</sub>	73.9
	EVER (NRG+SQ)	<b>75.2</b>
<i>Retrieval-Augmented Rectification</i>		
Llama 2 7B Chat	RRAR	37.8
	EVER (NRG+ER)	76.9
Llama 2 13B Chat	RRAR	41.5
	EVER (NRG+ER)	79.5
GPT 3.5 Turbo	RRAR	74.3
	EVER (NRG+ER) <sub>partial</sub>	84.1
	EVER (NRG+ER)	<b>94.5</b>
<i>Retrieval-Augmented Generation and Rectification</i>		
PerplexityAI	RAG <sup>‡</sup>	71.2
Llama 2 7B Chat	RAG	79.4
	Self-RAG*	81.2
	EVER (RAG+ER)	86.4
Llama 2 13B Chat	RAG*	79.9
	Self-RAG*	80.2
	EVER (RAG+ER)	87.3
GPT 3.5 Turbo	RAG	92.7
	EVER (RAG+ER) <sub>partial</sub>	93.6
	EVER (RAG+ER)	<b>95.8</b>

### 3.2 Reasoning Task

The final task we evaluate is the reasoning task, where the phenomenon of "hallucination snowballing" frequently arises (Zhang et al., 2023a). By leveraging the Chain-of-Thought (CoT) prompting method (Wei et al., 2022), we present multi-hop questions that required LLMs to construct an accurate and factually correct reasoning chain to provide the correct answers.

### 3.2.1 Experimental Setup

**Datasets and Experiment Settings.** In this task, follow Trivedi et al. (2022), we use the test subset comprising 500 examples from the multi-hop question answering HotPotQA dataset (Yang et al., 2018). We calculate the exact match (EM) and F1 score, following (Yang et al., 2018; Gou et al., 2023). For other experiment settings, we use the same setting as in the biography generation task.

**Baselines.** Similar to the biography task, we evaluate all variants of EVER in the reasoning task. For each variant, we employ different baselines for evaluation. First, in the non-retrieval scenario, we compare EVER (NRG+SQ) with Few-shot CoT (Wei et al., 2022). Second, in the retrieval-based rectification scenario, we compare EVER (NRG+ER) with CRITIC (Gou et al., 2023). Third, we compare EVER (RAG+ER) with retrieval-based generation method IRCOT (Trivedi et al., 2022). See details about these baselines in Appendix F.

### 3.2.2 Results & Analysis

In Table 2, we report the results of EVER and other baselines on HotPotQA. According to the results, we demonstrate the superiority of EVER in improving the effectiveness of CoT prompting in reasoning tasks. Similar to the biography generation task, retrieval-based method significantly improves the performance compared with Few-Shot CoT. In addition, EVER (NRG+ER) outperforms CRITIC, likely because CRITIC, while capable of verifying the final answers to multi-hop reasoning questions, corrects the reasoning chain as a whole rather than step-by-step. This approach cannot mitigate the "snowballing" issue throughout the steps. Moreover, by integrating retrieved knowledge prior to generation and incorporating a validation phase after generation, EVER (RAG+ER) outperforms the IRCOT method. This indicates the importance of both pre-generation retrieval and post-generation validation in enhancing the accuracy and reliability of CoT-based reasoning.

### 3.3 Analysis

**Extrinsic Hallucination Analysis.** In the biography generation task, we conduct a human annotation analysis of the 300 instances that are classified as "Not Enough Info" (NEI). Here, we define three distinct categories of extrinsic hallucination, as showed in table 6. The most prevalent cases, found in 65% of cases, is that the evidence provided does not directly contain relevant informa-

Table 2: Results on the HotpotQA multi-hop reasoning dataset. "partial" indicates extracting and verifying the high-risk concepts. \*The result is from Gou et al. (2023).

Retrieval	Method	EM (%)	F1 (%)
N/A	Few-Shot CoT	32.6	46.8
	EVER (NRG+SQ) <sub>partial</sub>	32.9	47.9
	EVER (NRG+SQ)	<b>34.7</b>	<b>48.3</b>
Google	RRAR	34.5	46.7
	CRITIC*	40.3	52.9
Dataset	EVER (NRG+ER) <sub>partial</sub>	40.8	55.6
	EVER (NRG+ER)	<b>42.3</b>	<b>58.1</b>
Dataset	IRCoT	48.4	57.8
	EVER (RAG+ER) <sub>partial</sub>	48.2	58.1
	EVER (RAG+ER)	<b>51.4</b>	<b>61.2</b>

tion to support or contradict. The second most common error of the generated text, accounting for 15% of the instances, is that while the evidence is relevant, it requires additional inference. Also, 9% of cases involve subjective, opinion-based or interpretative content that is hard to classify objectively. Finally, our findings reveal that EVER incorrectly categorizes 11% of examples as "Not Enough Info" (NEI), despite these instances actually being supportive or contradictory. Nevertheless, the high accuracy of NEI-classified examples demonstrates both EVER's strong performance and the practicality of user warnings, cautioning against potential lack of factuality.

### 3.4 Efficiency Analysis.

Although the proposed active concept-level validation and rectification in EVER incurs time overheads, these overheads are typical in similar retrieval-based baselines. To further improve efficiency, we implemented two methods: 1) parallel validation of extracted concepts and 2) extracting and verifying only high-risk concepts. The parallel validation reduces the time required for verification by handling multiple concepts simultaneously. Next, by focusing on high-risk concepts of dates, jobs, numbers, and locations, we streamline the verification process, reducing the number of loops and runtime.

As Table 3 illustrates, all three EVER variants demonstrate lower runtime to those of other retrieval-based methods in biography generation by 76% and multi-hop reasoning by 40%. Our experiments demonstrate that the selective extraction reduces the average runtime by approximately 38%. This efficiency gain does not come at sacrificing

Table 3: Average runtime (s) comparison across different methods on the two datasets for the GPT 3.5 Turbo model. \*For CRITIC, involving up to three iterations, we calculate the average runtime.

Method	Biography	HotpotQA
RRAR	210.5	-
IRCoT	-	67.2
CRITIC*	-	83.8
EVER (NRG+SQ) <sub>partial</sub>	102.3	49.3
EVER (NRG+SQ)	195.7	73.6
EVER (NRG+ER) <sub>partial</sub>	85.2	53.2
EVER (NRG+ER)	141.8	86.9
EVER (RAG+ER) <sub>partial</sub>	<b>59.2</b>	<b>39.8</b>
EVER (RAG+ER)	95.4	62.8

performance, since EVER still outperforms other baselines (Table 1 and Table 2).

## 4 Experiment of Enhancing Factuality with Preference Tuning

Preference tuning on high-quality data has proven an effective method to improve factuality without introducing inference overhead (Tian et al., 2023a). In this section, we study the performance of finetuning language models by using the EVER-generated preference data pair to reduce hallucination.

### 4.1 Experimental Setup

**Datasets and Experiment Settings.** We adopt the aforementioned biography generation task and use the same 183 human entities as the test set to evaluate the fine-tuning result. Follow Tian et al. (2023a), we use the EVER-generated data as the preferred sample and other 20 randomly zero-shot generations for each human entity as the dispreferred sample. In total, we have 10,000 training preference pairs.

**Baselines.** We compare several methods, including the vanilla approach, which uses the SFT model output. Another baseline is FactTune-FS (Tian et al., 2023a), which samples 10 generations and runs DPO on  $\binom{10}{2}$  pairs with using FactScore to select the better one in each pair. Both of these two approaches compare with EVER-PREF (NRG+SQ). In addition, we use the vanilla RAG-generated data (RAG-PREF) as the chosen text as a baseline to create preference data pairs, which is then compared with the retrieval-based version of EVER-PREF.

#### 4.1.1 Results & Analysis.

As shown in Table 4, fine-tuning the Llama-2-7B-chat model on the biography generation task has

yielded insights into reducing hallucination in language models. Initially, finetuning using the generated data by retrieval-free and self-query versions of EVER demonstrates a reduction in hallucinations, as evidenced by the improvement in FactScore from the Vanilla baseline of 36.8% to 47.3% with EVER-PREF (NRG+SQ). This indicates that fine-tuning with retrieval-free methods enhances the factual accuracy of language models.

Further advancements are observed when finetuning incorporated text generated through retrieval mechanisms. Specifically, the utilization of more factual data by retrieval during fine-tuning, particularly with EVER-PREF (NRG+ER) and EVER-PREF (RAG+ER), increases the performance even further, achieving FactScores of 52.8% and 53.9% respectively. These results underscore the potential of fine-tuning language models with factually enriched datasets to mitigate hallucinations.

Table 4: Results of finetuning the Llama-2-7B-chat model on the biography generation task.

Method	FACTSCORE (%)
Vanilla	36.8
FactTune-FS	45.4
EVER-PREF (NRG+SQ)	<b>47.3</b>
RAG-PREF	50.2
EVER-PREF (NRG+ER)	52.8
EVER-PREF (RAG+ER)	<b>53.9</b>

## 5 Related Work

**Hallucination Detection.** Detecting hallucinations in LLMs is crucial for ensuring the reliability of generated content. To detect LLM hallucination, the first line of methods analyze the probability of tokens (Mielke et al., 2022; Kadavath et al., 2022; Varshney et al., 2023). Another line of methods leverage the inconsistency between multiple generated examples, including NLI-based approaches (Elaraby et al., 2023; Manakul et al., 2023) and QA-based methods (Manakul et al., 2023; Agrawal et al., 2023). In addition, Cohen et al. (2023) introduced a method in which one LM acts as an examiner, repeatedly cross-examining the outputs of the other LM to test their consistency. **Hallucination Mitigation.** A number of approaches have been developed to mitigate hallucination in LLMs. One line of work focuses on manipulating the model via decoding strategies (Chuang et al., 2023; Shi et al., 2023; Li et al., 2022, 2023) or preference fine-tuning (Tian et al., 2023b). Another line of work uses post-hoc edit methods, which can

be further divided into those involving retrieval (Peng et al., 2023; Menick et al., 2022; Gao et al., 2022; Chern et al., 2023; Yu et al., 2023; Varshney et al., 2023) and non-retrieval based strategies (Dhuliawala et al., 2023; Zhou et al., 2023). RAG is another approach to improve factuality by integrating external knowledge during the generation process (Lewis et al., 2020; Jiang et al., 2023; Asai et al., 2023). Yet, non-retrieval-based methods lack of updated information, RAG lacks of robustness to irrelevant and useless context, and post-hoc editing methods may not address the snowballing issue of hallucinations. Our proposed method, with step-by-step verification and rectification, effectively mitigates these challenges in prior work. In addition, we show that our proposed method EVER can be utilized to create better preference data to further finetune a LLM to be enhance its factuality.

**Reasoning Improvement.** Several studies aim to enhance LLMs’ performance in reasoning tasks. One line of works uses prompting strategies (Wei et al., 2022; Zhou et al., 2022; Kojima et al., 2022; Wang et al., 2022) to divide a difficult task into simpler ones and/or utilizes external tools to aid LLMs (Yao et al., 2022; Schick et al., 2023; Gao et al., 2023a; Yang et al., 2022), both of which are solving problems sequentially without checking the correctness of generation. Also, Gou et al. (2023); Zhao et al. (2023) involves post-generation verification. However, these works only focus on reasoning tasks, making it difficult to generalize to non-reasoning tasks. Additionally, they don’t improve the trustworthiness of generated texts. We take these challenges into consideration, and EVER utilizes general-purpose verification and rectification strategies that are suitable for various tasks. Furthermore, the user warning further enhances the trustworthiness of generated texts.

## 6 Conclusion

In this paper, we introduce the EVER, aiming to mitigate hallucination in LLMs. EVER effectively addresses both intrinsic and extrinsic hallucinations while also reducing the propagation of errors that may occur in sequential text generation. Our empirical results demonstrate that EVER significantly reduces hallucination in various tasks, including short-form QA, long-form biography generation, and reasoning. Moreover, EVER is able to generate better preference data pair to further finetune the model to reduce hallucination.



## 625 Limitation

626 This study acknowledges limitation in the EVER  
627 framework. Unlike conventional fact-checking pro-  
628 cess, which involves considering the information  
629 beyond the evidence (e.g., claimant, claim date,  
630 source, etc.) to check the factual accuracy, our  
631 focus is solely on enhancing text attribution to re-  
632 duce hallucinations. This only require an reference  
633 (which might be incorrect) that could support a  
634 fact.

## 635 Acknowledgement

636 We thank the Center for AI Safety and Google  
637 Cloud Research Credits program for supporting  
638 our computing needs.

## 639 References

- 640 Amani S Abumansour and Arkaitz Zubiaga. 2023.  
641 Check-worthy claim detection across topics for au-  
642 tomated fact-checking. *PeerJ Computer Science*,  
643 9:e1365.
- 644 Ayush Agrawal, Lester Mackey, and Adam Tauman  
645 Kalai. 2023. Do language models know when  
646 they’re hallucinating references? *arXiv preprint*  
647 *arXiv:2305.18248*.
- 648 Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and  
649 Hannaneh Hajishirzi. 2023. Self-rag: Learning to  
650 retrieve, generate, and critique through self-reflection.  
651 *arXiv preprint arXiv:2310.11511*.
- 652 Ralph Allan Bradley and Milton E Terry. 1952. Rank  
653 analysis of incomplete block designs: I. the method  
654 of paired comparisons. *Biometrika*, 39(3/4):324–  
655 345.
- 656 I Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua  
657 Feng, Chunting Zhou, Junxian He, Graham Neubig,  
658 Pengfei Liu, et al. 2023. Factool: Factuality detec-  
659 tion in generative ai—a tool augmented framework  
660 for multi-task and multi-domain scenarios. *arXiv*  
661 *preprint arXiv:2307.13528*.
- 662 Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon  
663 Kim, James Glass, and Pengcheng He. 2023. Dola:  
664 Decoding by contrasting layers improves factu-  
665 ality in large language models. *arXiv preprint*  
666 *arXiv:2309.03883*.
- 667 Roi Cohen, May Hamri, Mor Geva, and Amir Globerson.  
668 2023. Lm vs lm: Detecting factual errors via  
669 cross examination. *arXiv preprint arXiv:2305.13281*.
- 670 Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu,  
671 Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Ja-  
672 son Weston. 2023. Chain-of-verification reduces hal-  
673 lucination in large language models. *arXiv preprint*  
674 *arXiv:2309.11495*.

- Mohamed Elaraby, Mengyin Lu, Jacob Dunn, Xuey-  
ing Zhang, Yu Wang, and Shizhu Liu. 2023. Halo:  
Estimation and reduction of hallucinations in open-  
source weak large language models. *arXiv preprint*  
*arXiv:2308.11764*.
- Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony  
Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vin-  
cent Y Zhao, Ni Lao, Hongrae Lee, Da-Cheng  
Juan, et al. 2022. Attributed text generation via  
post-hoc research and revision. *arXiv preprint*  
*arXiv:2210.08726*.
- Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon,  
Pengfei Liu, Yiming Yang, Jamie Callan, and Gra-  
ham Neubig. 2023a. Pal: Program-aided language  
models. In *International Conference on Machine*  
*Learning*, pages 10764–10799. PMLR.
- Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen.  
2023b. Enabling large language models to generate  
text with citations. *arXiv preprint arXiv:2305.14627*.
- Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong  
Shen, Yujiu Yang, Nan Duan, and Weizhu Chen.  
2023. Critic: Large language models can self-correct  
with tool-interactive critiquing. *arXiv preprint*  
*arXiv:2305.11738*.
- 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 hallu-  
cination in natural language generation. *ACM Comput-*  
*ing Surveys*, 55(12):1–38.
- Zhengbao Jiang, Frank F Xu, Luyu Gao, Zhiqing  
Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang,  
Jamie Callan, and Graham Neubig. 2023. Ac-  
tive retrieval augmented generation. *arXiv preprint*  
*arXiv:2305.06983*.
- Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke  
Zettlemoyer. 2017. Triviaqa: A large scale distant-  
ly supervised challenge dataset for reading compre-  
hension. *arXiv preprint arXiv:1705.03551*.
- Saurav Kadavath, Tom Conerly, Amanda Askell, Tom  
Henighan, Dawn Drain, Ethan Perez, Nicholas  
Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli  
Tran-Johnson, et al. 2022. Language models  
(mostly) know what they know. *arXiv preprint*  
*arXiv:2207.05221*.
- Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric  
Wallace, and Colin Raffel. 2023. [Large language models struggle to learn long-tail knowledge](#). In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 15696–15707. PMLR.
- Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yu-  
taka Matsuo, and Yusuke Iwasawa. 2022. Large lan-  
guage models are zero-shot reasoners. *Advances in*  
*neural information processing systems*, 35:22199–  
22213.

675  
676  
677  
678  
679  
680  
681  
682  
683  
684  
685  
686  
687  
688  
689  
690  
691  
692  
693  
694  
695  
696  
697  
698  
699  
700  
701  
702  
703  
704  
705  
706  
707  
708  
709  
710  
711  
712  
713  
714  
715  
716  
717  
718  
719  
720  
721  
722  
723  
724  
725  
726  
727  
728  
729  
730

731	Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio	Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano	788
732	Petroni, Vladimir Karpukhin, Naman Goyal, Hein-	Ermon, Christopher D Manning, and Chelsea Finn.	789
733	rich Küttler, Mike Lewis, Wen-tau Yih, Tim Rock-	2023. Direct preference optimization: Your language	790
734	täschel, et al. 2020. Retrieval-augmented generation	model is secretly a reward model. <i>arXiv preprint</i>	791
735	for knowledge-intensive nlp tasks. <i>Advances in Neu-</i>	<i>arXiv:2305.18290</i> .	792
736	<i>ral Information Processing Systems</i> , 33:9459–9474.		
737	Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter	Timo Schick, Jane Dwivedi-Yu, Roberto Dessi, Roberta	793
738	Pfister, and Martin Wattenberg. 2023. Inference-time	Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola	794
739	intervention: Eliciting truthful answers from a lan-	Cancedda, and Thomas Scialom. 2023. Toolformer:	795
740	guage model. <i>arXiv preprint arXiv:2306.03341</i> .	Language models can teach themselves to use tools.	796
		<i>arXiv preprint arXiv:2302.04761</i> .	797
741	Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang,	Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia	798
742	Jason Eisner, Tatsunori Hashimoto, Luke Zettle-	Tsvetkov, Luke Zettlemoyer, and Scott Wen-tau	799
743	moyer, and Mike Lewis. 2022. Contrastive decoding:	Yih. 2023. Trusting your evidence: Hallucinate	800
744	Open-ended text generation as optimization. <i>arXiv</i>	less with context-aware decoding. <i>arXiv preprint</i>	801
745	<i>preprint arXiv:2210.15097</i> .	<i>arXiv:2305.14739</i> .	802
746	Potsawee Manakul, Adian Liusie, and Mark JF Gales.	Kai Sun, Yifan Ethan Xu, Hanwen Zha, Yue Liu, and	803
747	2023. Selfcheckgpt: Zero-resource black-box hal-	Xin Luna Dong. 2023. Head-to-tail: How knowl-	804
748	lucination detection for generative large language	edgeable are large language models (llm)? aka will	805
749	models. <i>arXiv preprint arXiv:2303.08896</i> .	llms replace knowledge graphs? <i>arXiv preprint</i>	806
		<i>arXiv:2308.10168</i> .	807
750	Jacob Menick, Maja Trebacz, Vladimir Mikulik,	Katherine Tian, Eric Mitchell, Huaxiu Yao, Christo-	808
751	John Aslanides, Francis Song, Martin Chadwick,	pher D Manning, and Chelsea Finn. 2023a. Fine-	809
752	Mia Glaese, Susannah Young, Lucy Campbell-	tuning language models for factuality. <i>arXiv preprint</i>	810
753	Gillingham, Geoffrey Irving, et al. 2022. Teaching	<i>arXiv:2311.08401</i> .	811
754	language models to support answers with verified		
755	quotes. <i>arXiv preprint arXiv:2203.11147</i> .		
756	Sabrina J Mielke, Arthur Szlam, Emily Dinan, and Y-	Katherine Tian, Eric Mitchell, Huaxiu Yao, Christo-	812
757	Lan Boureau. 2022. Reducing conversational agents’	pher D Manning, and Chelsea Finn. 2023b. Fine-	813
758	overconfidence through linguistic calibration. <i>Trans-</i>	tuning language models for factuality. <i>arXiv preprint</i>	814
759	<i>actions of the Association for Computational Linguis-</i>	<i>arXiv:2311.08401</i> .	815
760	<i>tics</i> , 10:857–872.		
761	Sewon Min, Danqi Chen, Hannaneh Hajishirzi, and	Hugo Touvron, Louis Martin, Kevin Stone, Peter Al-	816
762	Luke Zettlemoyer. 2019. <a href="#">A discrete hard EM ap-</a>	bert, Amjad Almahairi, Yasmine Babaei, Nikolay	817
763	<a href="#">proach for weakly supervised question answering.</a>	Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti	818
764	<i>Proceedings of the 2019 Conference on Empirical</i>	Bhosale, et al. 2023. Llama 2: Open founda-	819
765	<i>Methods in Natural Language Processing and the</i>	tion and fine-tuned chat models. <i>arXiv preprint</i>	820
766	<i>9th International Joint Conference on Natural Lan-</i>	<i>arXiv:2307.09288</i> .	821
767	<i>guage Processing (EMNLP-IJCNLP)</i> , pages 2851–		
768	2864, Hong Kong, China. Association for Computa-	Harsh Trivedi, Niranjan Balasubramanian, Tushar	822
769	tional Linguistics.	Khot, and Ashish Sabharwal. 2022. Interleav-	823
		ing retrieval with chain-of-thought reasoning for	824
770	Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike	knowledge-intensive multi-step questions. <i>arXiv</i>	825
771	Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer,	<i>preprint arXiv:2212.10509</i> .	826
772	Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023.		
773	Factscore: Fine-grained atomic evaluation of factual	Neeraj Varshney, Wenlin Yao, Hongming Zhang, Jian-	827
774	precision in long form text generation. <i>arXiv preprint</i>	shu Chen, and Dong Yu. 2023. A stitch in time saves	828
775	<i>arXiv:2305.14251</i> .	nine: Detecting and mitigating hallucinations of	829
		llms by validating low-confidence generation. <i>arXiv</i>	830
776	Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida,	<i>preprint arXiv:2307.03987</i> .	831
777	Carroll Wainwright, Pamela Mishkin, Chong Zhang,		
778	Sandhini Agarwal, Katarina Slama, Alex Ray, et al.	Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry	832
779	2022. Training language models to follow instruc-	Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny	833
780	tions with human feedback. <i>Advances in Neural</i>	Zhou, Quoc Le, et al. 2023. Freshllms: Refreshing	834
781	<i>Information Processing Systems</i> , 35:27730–27744.	large language models with search engine augmenta-	835
		tion. <i>arXiv preprint arXiv:2310.03214</i> .	836
782	Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng,	Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le,	837
783	Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou	Ed Chi, Sharan Narang, Aakanksha Chowdhery, and	838
784	Yu, Weizhu Chen, et al. 2023. Check your facts and	Denny Zhou. 2022. Self-consistency improves chain	839
785	try again: Improving large language models with	of thought reasoning in language models. <i>arXiv</i>	840
786	external knowledge and automated feedback. <i>arXiv</i>	<i>preprint arXiv:2203.11171</i> .	841
787	<i>preprint arXiv:2302.12813</i> .		

842	Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. <i>Advances in Neural Information Processing Systems</i> , 35:24824–24837.	<b>A Additional Experiment on Short-form QA Task</b>	897
843			898
844			
845		Honesty-tuned LLMs may exhibit over-conservatism due to an imbalanced trade-off between helpfulness and honesty (Ouyang et al., 2022). In this short-form QA task, we evaluate EVER’s ability to strike a better balance in this trade-off. Employing open-domain questions, EVER is designed to either abstain from answering or to modify answers depending on the context, aiming for generating more trustworthy text.	899
846			900
847	Orion Weller, Marc Marone, Nathaniel Weir, Dawn Lawrie, Daniel Khashabi, and Benjamin Van Durme. 2023. "according to..." prompting language models improves quoting from pre-training data. <i>arXiv preprint arXiv:2305.13252</i> .		901
848			902
849			903
850			904
851			905
852	Dustin Wright and Isabelle Augenstein. 2020. <b>Claim check-worthiness detection as positive unlabelled learning</b> . In <i>Findings of the Association for Computational Linguistics: EMNLP 2020</i> , pages 476–488, Online. Association for Computational Linguistics.	<b>A.1 Experimental Setup</b>	908
853			
854		<b>Dataset.</b> In this task, we use two short-form QA datasets, including TriviaQA-unfiltered (Joshi et al., 2017) and ALCE-Qampari QA (Gao et al., 2023b). For TriviaQA, we assume there is only one correct answer for each question. Since the test set of TriviaQA is not publicly available, we use the same test split from validation set as Min et al. (2019); Asai et al. (2023).	909
855			910
856			911
857	Kaiyu Yang, Jia Deng, and Danqi Chen. 2022. Generating natural language proofs with verifier-guided search. <i>arXiv preprint arXiv:2205.12443</i> .		912
858			913
859			914
860	Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. <i>arXiv preprint arXiv:1809.09600</i> .		915
861			916
862		<b>Evaluation Metric.</b> Following Schick et al. (2023), we evaluate performance based on whether gold answers are included in the model generations, rather than strictly requiring an exact string match. We report accuracy on the answered examples as $N_c / (N_{all} - N_{rej})$ , and the percentage of trustworthy examples as $(N_c + N_{rej}) / N_{all}$ , where $N_c$ , $N_{rej}$ , and $N_{all}$ represent the number of correct examples, abstention examples, and all examples, respectively. For Qampari QA, where the gold answer is a list of answers, we follow Gao et al. (2023b); Schick et al. (2023) in evaluating performance using the <i>recall@5</i> metric. Here, we consider recall to be 100% if the prediction includes at least 5 correct answers. Additionally, we assess the <i>precision</i> of the model’s prediction by checking for an exact string match with the gold answer list.	917
863			918
864			919
865	Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. <i>arXiv preprint arXiv:2210.03629</i> .		920
866			921
867			922
868			923
869	Wenhao Yu, Zhihan Zhang, Zhenwen Liang, Meng Jiang, and Ashish Sabharwal. 2023. Improving language models via plug-and-play retrieval feedback. <i>arXiv preprint arXiv:2305.14002</i> .		924
870			925
871			926
872			927
873	Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A Smith. 2023a. How language model hallucinations can snowball. <i>arXiv preprint arXiv:2305.13534</i> .		928
874			929
875			930
876			931
877	Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023b. Siren’s song in the ai ocean: A survey on hallucination in large language models. <i>arXiv preprint arXiv:2309.01219</i> .		932
878			933
879			934
880			935
881			936
882	Ruochen Zhao, Xingxuan Li, Shafiq Joty, Chengwei Qin, and Lidong Bing. 2023. Verify-and-edit: A knowledge-enhanced chain-of-thought framework. <i>arXiv preprint arXiv:2305.03268</i> .	<b>Baselines.</b> We evaluate EVER against two categories of baseline approaches: (1) zero-shot generation and vanilla retrieval-augmented generation, and (2) improvements to the baselines in category (1) by prompting LLMs to abstain from uncertain examples. In the zero-shot and RAG approaches with abstention prompting, LLMs respond with "Sorry, I don’t know" when unsure or when retrieved evidence is insufficient to answer, respectively. See detailed discussions in Appendix D.	937
883			938
884			939
885			940
886	Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, et al. 2022. Least-to-most prompting enables complex reasoning in large language models. <i>arXiv preprint arXiv:2205.10625</i> .	<b>Experiment Settings.</b> We employ two methods to retrieve relevant evidence: Google and the dataset. For each question, we retrieve the top 5 relevant	941
887			942
888			943
889			944
890			945
891			946
892	Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. 2023. Analyzing and mitigating object hallucination in large vision-language models. <i>arXiv preprint arXiv:2310.00754</i> .		
893			
894			
895			
896			

947	documents from the provided dataset. When using	995
948	Google, we retrieve a total of 10 relevant docu-	996
949	ments by querying both the question and the con-	997
950	catenation of the question and answer strings.	998
951	<b>A.2 Results and Analysis</b>	999
952	Table 5 reveals that traditional abstention	1000
953	prompting-based methods, as highlighted	1001
954	in Ouyang et al. (2022), tend to exhibit over-	1002
955	conservatism by refusing to answer a significant	1003
956	number of questions across datasets. In contrast,	
957	our EVER method stands out for its inclination	<b>E Biography Generation Baselines</b>
958	to provide correct answers rather than abstaining,	1004
959	significantly enhancing the helpfulness of the	
960	generated text. Additionally, EVER outperforms	• <b>Dola:</b> This decoding method leverages the ob-
961	other baselines in trustworthiness, as evidenced by	995
962	its higher trustful rate in Trivia QA. Furthermore,	1006
963	EVER demonstrates strong performance in	1007
964	producing higher correctness/factuality, showing	1008
965	higher accuracy, precision and recall compared	1009
966	to other baselines. Finally, EVER with evidence	1010
967	retrieval can also address the limitations of	1011
968	RAG. In the Trivia QA dataset, RAG performs	
969	even worse compared with zero-shot generation	• <b>CoVe:</b> In this non-retrieval-based pipeline,
970	when using the top-5 retrieved documents from	1012
971	the provided dataset as context, often due to	1013
972	the inclusion of irrelevant or misleading text.	1014
973	However, this issue can be effectively resolved by	1015
974	employing EVER. In summary, EVER effectively	1016
975	balances the trade-off between helpfulness and	
976	honesty, ensuring that the text it generates is both	• <b>RRAR:</b> This approach automatically at-
977	informative and reliable.	995
978	<b>B Extrinsic Hallucination</b>	1017
979	<b>C Multi-Round Rectification</b>	1018
980	We evaluate the effects of allowing multi-round rec-	1019
981	tification for GPT 3.5 Turbo model. The results in	1020
982	Table 7 shows that in general one round of rectifica-	1021
983	tion is sufficient for both tasks. Additional rounds	
984	of rectification yield negligible improvements in	• <b>Self-RAG:</b> This method improves an LM's
985	performance.	995
986	<b>D Short-form QA Baselines</b>	1022
987	Zero-shot involves generating texts solely based on	1023
988	the provided prompt without any additional contex-	1024
989	tual information. Retrieval Augmented Generation	1025
990	(RAG) incorporates an external knowledge in the	1026
991	prompt to enhance the generation process. RAG	1027
992	has two sources: relevant documents provided in	1028
993	the original datasets and relevant documents ob-	1029
994	tained through Google Search. For prompting,	
	we employ prompting engineering to increase the	<b>F Reasoning Baselines</b>
	trustworthiness of generated text by instructing the	1030
	model to respond with "I don't know" if there is no	
	answer within the context. The model's response	• <b>CRITIC:</b> This method enables LLMs to self-
	"I don't know" is considered an abstention. For the	995
	methods of zero-shot, zero-shot+prompting, RAG,	1032
	and RAG+prompting, as well as different datasets,	1033
	we use different prompts, which are listed in Ta-	1034
	ble 8 and Table 9.	1035
		1036
		• <b>IRCoT:</b> This work integrates retrieval into the
		995
		1037
		1038
		1039
		1040

Table 5: The results of GPT 3.5 turbo on the Trivia QA and Qampari QA datasets.

Method	Retrieval	Trivia QA			Qampari QA		
		Accuracy	% Trustful	% Abstention	Recall@5	Precision	% Abstention
Zero-shot	N/A	76.7	76.7	-	11.6	16.8	-
Zero-shot+prompting	N/A	80.4	79.0	11.7	11.4	33.5	46.0
<b>EVER (NRG+ER)</b>	Dataset	83.4	82.8	3.0	11.8	26.6	9.0
RAG	Dataset	71.3	71.3	-	22.8	35.2	-
RAG+prompting	Dataset	79.2	80.3	14.7	22.7	38.9	29.5
<b>EVER (RAG+ER)</b>	Dataset	82.3	86.8	5.3	<b>23.3</b>	<b>39.2</b>	1.0
RAG	Google	79.0	79.0	-	-	-	-
RAG+prompting	Google	81.3	82.0	10.0	-	-	-
<b>EVER (RAG+ER)</b>	Google	<b>84.9</b>	<b>87.7</b>	4.0	-	-	-

Table 6: The three categories of extrinsic hallucination identified by ChatGPT based on human annotations, along with their respective percentages. We also list one representative validation question and the corresponding evidence, where the extracted concept for each validation question are marked in yellow.

Category	Validation Question	Evidence
Not mention (65%)	Did notable achievements and impact in <b>Liga MX</b> earn Jorge Enríquez Garcías a debut for the Mexico national team?	... Jorge Enríquez first played for the Mexico national team at the 2011 CONCACAF U-20 Championship ...
Need further inference (15%)	Is Chris Johns one of the most dominant <b>featherweight champions</b> in boxing history?	... Chris John was The Ring’s #8-ranked featherweight in the world (and #10 pound-for-pound) ...
Subjective (9%)	Has Bobo Baldé left a <b>lasting impact</b> on the football world?	... Dianbobo "Bobo" Baldé (born 5 October 1975) is a former professional footballer who played as a defender ...
Misclassified examples (11%)		

Table 7: The results of multi-round rectification of EVER (NRG+ER) on the biography generation and reasoning tasks for GPT 3.5 Turbo.

# Rounds	FACTSCORE (%)	EM (%)	F1 (%)
1	94.5	42.3	58.1
2	94.7	43.5	57.8
3	95.2	43.1	59.4
4	93.8	42.6	58.3

## G Prompt Templates

Table 8: The prompts used to generate answers for the QampariQA dataset.

---

**Zero-shot**

Provide a list of accurate answers for the given question using only the provided context (some of which might be irrelevant). Separate answers by semicolons. For questions that have more than 5 answers, write at least 5 answers.

**Question:** ...

**Answer:**

**Zero-shot+prompting**

Provide a list of accurate answers for the given question using only the provided context (some of which might be irrelevant). Separate answers by semicolons. For questions that have more than 5 answers, write at least 5 answers. If there is no answer in the context, reply “sorry I don’t know”.

**Question:** ...

**Answer:**

**RAG**

**Context:** ...

Provide a list of accurate answers for the given question using only the provided context (some of which might be irrelevant). Separate answers by semicolons. For questions that have more than 5 answers, write at least 5 answers.

**Question:** ...

**Answer:**

**RAG+prompting**

**Context:** ...

Provide a list of accurate answers for the given question using only the provided context (some of which might be irrelevant). Separate answers by semicolons. For questions that have more than 5 answers, write at least 5 answers. If there is no answer in the context, reply “sorry I don’t know”.

**Question:** ...

**Answer:**

---

Table 9: The prompts used to generate answers for the TriviaQA dataset.

---

**Zero-shot**

Answer the following question.

**Question:** ...

**Answer:**

**Zero-shot+prompting**

Answer the following question based on the context. If there is no answer in the context, reply “sorry I don’t know”.

**Question:** ...

**Answer:**

**RAG**

**Context:** ...

Answer the following question based on the context.

**Question:** ...

**Answer:**

**RAG+prompting**

**Context:** ...

Answer the following question based on the context. If there is no answer in the context, reply “sorry I don’t know”.

**Question:** ...

**Answer:**

---

Table 10: The prompts used to extract concepts.

---

**Instruction:** Identify all objective factual concepts from the following sentence. Exclude the main subject and any subjective terms. Include all numerical details (such as times, quantities, etc.). Present your findings in a list separated by semicolons.

**Sentence:** Claude Monet (14 November 1840 – 26 December 1926) was a French painter born in Rue Laffitte, Paris, France, who along with his companions Auguste Renoir, Edgar Degas and Pierre-Auguste Renoir, is often referred to as the founder of Impressionism.

**Answer:** 14 November 1840; 26 December 1926; Rue Laffitte, Paris, France; French; painter; Auguste Renoir; Edgar Degas; Pierre-Auguste Renoir; founder of Impressionism

**Instruction:** Identify all objective factual concepts from the following sentence. Exclude the main subject and any subjective terms. Include all numerical details (such as times, quantities, etc.). Present your findings in a list separated by semicolons.

**Sentence:** Lee Min-ho has also won several awards for his outstanding performances in popular films like "Gangnam Blues" and "Bounty Hunters."

**Answer:** awards; popular films; Gangnam Blues; Bounty Hunters

**Instruction:** Identify all objective factual concepts from the following sentence. Exclude the main subject and any subjective terms. Include all numerical details (such as times, quantities, etc.). Present your findings in a list separated by semicolons.

**Sentence:** Pablo Escobar, often referred to as "El Patrón," was a Colombian drug lord and the leader of the Medellín Cartel, dominating the cocaine trade during the 1970s and 1980s.

**Answer:** El Patrón; Colombian; drug lord; Medellín Cartel; cocaine trade; 1970s; 1980s

**Instruction:** Identify all objective factual concepts from the following sentence. Exclude the main subject and any subjective terms. Include all numerical details (such as times, quantities, etc.). Present your findings in a list separated by semicolons.

**Sentence:** Meryl Streep earned widespread acclaim for her performances in films like "The Iron Lady," "Doubt," and "Julie & Julia."

**Answer:** The Iron Lady; Doubt; Julie & Julia

**Instruction:** Identify all objective factual concepts from the following sentence. Exclude the main subject and any subjective terms. Include all numerical details (such as times, quantities, etc.). Present your findings in a list separated by semicolons.

**Sentence:** {sentence}

**Answer:**

---

Table 11: The prompts used to generate validation questions for smaller models, such as Llama 2 7B/13B Chat. For GPT-3.5, we use zero-shot with the same instruction.

---

**Sentence:** Leonardo da Vincian, an Italian polymath of the High Renaissance who was active as a painter, draughtsman, engineer, scientist, theorist, sculptor, and architect, was born in Vinci, Italy, on 15 April 1452.  
For the above sentence about "Leonardo da Vinci", generate a yes/no question WITHOUT any pronouns about the entity of "15 April 1452". The question MUST contain the entity.

**Question:** Was Leonardo da Vinci born on 15 April 1452?

**Sentence:** Wolfgang Amadeus Mozart, during his brief lifetime, composed more than 600 works, many of which are acknowledged as the pinnacles of symphonic, concertante, chamber, operatic, and choral music.  
For the above sentence about "Wolfgang Amadeus Mozart", generate a yes/no question WITHOUT any pronouns about the entity of "more than 600 works". The question MUST contain the entity.

**Question:** Did Wolfgang Amadeus Mozart compose more than 600 works during his lifetime?

**Sentence:** Frida Kahlo, a renowned Mexican artist, is best known for her self-portraits and works like "The wounded deer" and "The Two Fridas".  
For the above sentence about "Frida Kahlo", generate a yes/no question WITHOUT any pronouns about the entity of "The Two Fridas". The question MUST contain the entity.

**Question:** Did Frida Kahlo create "The Two Fridas"?

**Sentence:** {sentence}  
For the above sentence about "{topic}", generate a yes/no question WITHOUT any pronouns about the entity of "{topic}". The question MUST contain the entity.

**Question:**

---

Table 12: The prompts used to do support checking with evidence retrieval.

---

Based on the evidence, answer the following question by selecting one of these options: True, False, or Not Enough Information. YOU MUST PROVIDE THE REASONING FIRST BEFORE MAKING A DECISION.

**Evidence:** Jane Austen - BritishLiteratureArchive.org: Jane Austen (16 December 1775 – 18 July 1817) was an English novelist known for her novels that critique the British landed gentry of the 18th century.

**Question:** Was Jane Austen an English novelist?

**Answer:** The evidence presents Austen as an English novelist. The claim is consistent with this information. Therefore, the decision is True.

Based on the evidence, answer the following question by selecting one of these options: True, False, or Not Enough Information. YOU MUST PROVIDE THE REASONING FIRST BEFORE MAKING A DECISION.

**Evidence:** Ada Lovelace - WomenInTechHistory.com: Ada Lovelace (10 December 1815 – 27 November 1852) was an English mathematician and writer, chiefly known for her work on Charles Babbage’s proposed mechanical general-purpose computer, the Analytical Engine.

**Question:** Is Ada Lovelace regarded as the first computer programmer?

**Answer:** The evidence describes Ada’s significant work on the Analytical Engine, a proposed mechanical computer by Charles Babbage. However, it doesn’t explicitly state that she is considered the first computer programmer. Therefore, the decision is Not Enough Information.

Based on the evidence, answer the following question by selecting one of these options: True, False, or Not Enough Information. YOU MUST PROVIDE THE REASONING FIRST BEFORE MAKING A DECISION.

**Evidence:** Leonardo da Vinci - RenaissanceMasters.org: Leonardo da Vinci (15 April 1452 – 2 May 1519) was an Italian polymath of the Renaissance era, known for his works in painting, science, mathematics, and various other fields.

**Question:** Was Leonardo da Vinci a 17th-century composer known for his operas?

**Answer:** The evidence introduces da Vinci as an Italian polymath from the Renaissance era, acclaimed for his contributions in painting, science, and other areas. The claim erroneously describes him as a 17th-century composer, which doesn’t align with the known facts. Therefore, the decision is False.

Based on the evidence, answer the following question by selecting one of these options: True, False, or Not Enough Information. Multiple sources of evidence are presented, each separated by a semicolon. YOU MUST PROVIDE THE REASONING FIRST BEFORE MAKING A DECISION.

**Evidence:** {evidence}

**Question:** {validation question}

**Answer:**

---



Table 13: The prompts used to do support checking with self-query. We use an "According to" prompting strategy to better recall memorized facts in LMs (Weller et al., 2023).

---

Answer the following question by selecting one of these options: True, False, or Not Enough Information. YOU MUST PROVIDE THE REASONING FIRST BEFORE MAKING A DECISION.

**Question:** Was Jane Austen an English novelist?

**Answer:** According to Wikipedia, Jane Austen (1775-1817) was an English novelist who is best known for her six major novels, including "Pride and Prejudice," "Sense and Sensibility," and "Emma." Therefore, the decision is True.

**Question:** Was Jane Austen an English novelist?

**Answer:** According to Wikipedia, Jane Austen (1775-1817) was an English novelist who is best known for her six major novels, including "Pride and Prejudice," "Sense and Sensibility," and "Emma." Therefore, the decision is True.

Answer the following question by selecting one of these options: True, False, or Not Enough Information. YOU MUST PROVIDE THE REASONING FIRST BEFORE MAKING A DECISION.

**Question:** Was Jane Austen an English novelist?

**Answer:** According to Wikipedia, Jane Austen (1775-1817) was an English novelist who is best known for her six major novels, including "Pride and Prejudice," "Sense and Sensibility," and "Emma." Therefore, the decision is True.

Answer the following question by selecting one of these options: True, False, or Not Enough Information. YOU MUST PROVIDE THE REASONING FIRST BEFORE MAKING A DECISION.

**Question:** Is Ada Lovelace regarded as the first computer programmer?

**Answer:** According to Wikipedia, Ada Lovelace (1815-1852) was an English mathematician and writer, known for her work on Charles Babbage's early mechanical general-purpose computer, the Analytical Engine. No further information about her high school love is mentioned on Wikipedia. Therefore, the decision is Not Enough Information.

Answer the following question by selecting one of these options: True, False, or Not Enough Information. YOU MUST PROVIDE THE REASONING FIRST BEFORE MAKING A DECISION.

**Question:** Was Leonardo da Vinci a 17th-century composer known for his operas?

**Answer:** According to Wikipedia, Leonardo da Vinci as an Italian polymath from the Renaissance era, acclaimed for his contributions in painting, science, and other areas. The claim erroneously describes him as a 17th-century composer, which doesn't align with the known facts. Therefore, the decision is False.

Answer the following question by selecting one of these options: True, False, or Not Enough Information. YOU MUST PROVIDE THE REASONING FIRST BEFORE MAKING A DECISION.

**Question:** {validation question}

**Answer:** According to Wikipedia,

---