UFO: a Unified and Flexible Framework for Evaluating Factuality of Large Language Models

Anonymous ACL submission

Abstract

Large language models (LLMs) may generate text that lacks consistency with human knowledge, leading to factual inaccuracies or *hallucination*. Existing research for evaluating the factuality of LLMs involves extracting fact claims using an LLM and verifying them against a predefined fact source. However, these evaluation metrics are task-specific, and not scalable, and the substitutability of fact sources in different tasks is under-explored. To address these challenges, we categorize four available fact sources: human-written evidence, reference documents, search engine results, and LLM knowledge, along with five text generation tasks containing six representative datasets. Then, we propose UFO, an LLM-based unified and flexible evaluation framework to verify facts against plug-and-play fact sources. We implement five evaluation scenarios based on this framework. Experimental results show that for most QA tasks, human-written evidence and reference documents are crucial, and they can substitute for each other in retrieval-augmented QA tasks. In news fact generation tasks, search engine results and LLM knowledge are essential. Our dataset and code are available at https://anonymous.4open.science/r/UFO-813F.

1 Introduction

The advancement of large language models (LLMs) has facilitated the development of generative artificial intelligence (Zhao et al., 2023). Many LLM-based applications have been released, such as ChatGPT and Bing Chat (also known as Bing Copilot), which gradually change people's working habits.[1] However, LLMs tend to generate factually inaccurate texts, which lack consistency with human knowledge, and degrade the usability of the model-generated text. Such a shortcoming of

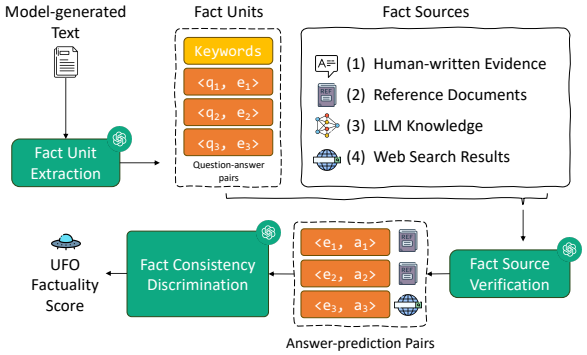

Figure 1: Our proposed factuality evaluation pipeline UFO . We integrate four fact sources within various evaluation scenarios to assess the factuality score.

LLMs is well-known as *hallucination* (Bang et al., 2023; Ji et al., 2023). The quality of datasets and training paradigms are concerned as the potential factors causing hallucinations in LLMs (Li et al., 2022). How to detect and measure the hallucinations in model-generated texts has received increasing attention.

Current automatic evaluation metrics employ a specific fact source to evaluate the factuality of LLMs for certain tasks. However, there is still a lack of analysis on the applicability of different fact sources in various tasks. Considering the establishment of a new task, the fact sources relied upon by previous evaluation methods may not be applicable. It's important to consider whether alternative fact sources can be utilized. For example, when a new QA task arises, collecting human-written evidence can be extremely costly. In such cases, whether search results from a search engine can be used as a substitute for human-written evidence as a fact source remains unexplored.

To address the issue, we propose UFO , a **U**nified and **F**lexible framework for factuality evaluati**O**n, which allows for: (a) Flexibly integration of various fact sources. (b) A unified verification method that enables switching fact sources in specific tasks.

[1] ChatGPT: https://chat.openai.com/chat, Bing Chat: https://www.bing.com/new

(c) The combination of different fact sources to enhance the factuality evaluation. In our framework shown in Figure 1, we first extract fact units from the text, including verifiable question-answer pairs and keywords of model-generated text. Then, for each fact, we verify it against the set of fact sources until a matching answer is found. Finally, we assign a binary matching score to each fact.

With the support of this evaluation framework, we can systematically analyze the evaluation capabilities of different fact sources across various scenarios in existing evaluation tasks. Specifically, we consider four different fact sources: **(1) Human-written evidence**. This corresponds to some text generation tasks with labeled data. For example, expert-validated QA tasks often provide human-written answers for evaluation. **(2) Reference documents**. Many recent studies, *e.g.*, WebBrain (Qian et al., 2023), WebGPT (Nakano et al., 2022), GopherCite (Menick et al., 2022), WebCPM (Qin et al., 2023), WebGLM (Liu et al., 2023), ALCE (Gao et al., 2023) and Bing Chat, have reported that leveraging reference documents can facilitate LLMs generation of more factual text. Therefore, such reference documents can also be a fact source for factuality evaluation. **(3) Search engine results**. When humans are asked to check the factuality of a text, they usually make judgments by turning to search engines. **(4) LLM knowledge**. Existing studies (Fu et al., 2023) suggest that advanced LLMs (such as GPT-4) can serve as a fact source for verification.

We design five evaluation scenarios where different fact sources and their combinations are used, summarized in Table 1, to demonstrate the flexibility of UFO . In each evaluation scenario, we compute the discriminative power (Sakai, 2006) of our proposed framework and compare it with eight baseline metrics. We experiment with these evaluation scenarios over five text-generation tasks, including Open-domain QA, Web Retrieval-based QA, Expert-Validated QA, News Fact Generation, and Retrieval-Augmented QA, to investigate the importance of data sources in different task scenarios. The experimental results demonstrated that in most QA tasks, obtaining human-written evidence and reference documents enhances the discriminative power of the evaluation pipeline. In the news fact generation task, we only require the search engine results and LLM knowledge to verify facts. In the retrieval-augmented QA task, the positive effects derived from two fact sources are comparable, thus allowing them to be substituted for each other. Although not the main focus of this paper, we evaluate six existing LLMs: Bing Chat in "precise" generation mode, ChatGPT, LLaMA-7b, LLaMA-13b, Vicuna-7b, and Vicuna-13b. We discovered that the factuality score of Bing Chat in precise mode is lower than that of ChatGPT, yet comparable to Vicuna-13b. In open-source LLMs, increasing the scale of model parameters can enhance factual accuracy.

Fact Sources	(1) Human-written evidence (S_{he}); (2) Reference documents (S_{rd}); (3) Search engine results (S_{se}); (4) LLM knowledge (S_{lk}).
Evaluation Scenarios	(1) $\langle S_{\text{se}}, S_{\text{lk}} \rangle$; (2) $\langle S_{\text{he}}, S_{\text{se}}, S_{\text{lk}} \rangle$; (3) $\langle S_{\text{rd}}, S_{\text{se}}, S_{\text{lk}} \rangle$; (4) $\langle S_{\text{he}}, S_{\text{rd}}, S_{\text{se}}, S_{\text{lk}} \rangle$; (5) $\langle S_{\text{rd}}, S_{\text{he}}, S_{\text{se}}, S_{\text{lk}} \rangle$.
Tasks	(1) Open-domain QA; (2) Web Retrieval-based QA; (3) Expert-Validated QA; (4) News Fact Generation; and (5) Retrieval-Augmented QA.

Table 1: The fact sources, evaluation scenarios, and tasks we study in the paper.

Our contributions can be summarized as follows:

• We propose UFO , a pipeline integrating flexible plug-and-play fact sources with unified verification methods for evaluating the factuality of LLMs.

• We conduct a systematic analysis of the evaluation capabilities of four fact sources across five factuality evaluation scenarios and five tasks.

• We reveal that human-written evidence and reference documents are essential in QA tasks, while search engine results and LLM knowledge are crucial in news fact generation tasks.

2 Related Work

2.1 Text Generation and Hallucination

The advancement of text generation has been propelled by pre-trained language models (PLMs) like BART (Lewis et al., 2020), T5 (Raffel et al., 2020), and GPT-2 (Radford et al., 2019), utilizing structures that range from encoder-decoder to decoder-only configurations. The emergence of LLMs such as GPT-3 (Brown et al., 2020), characterized by their vast parameter counts and extensive training data, marked a significant evolution. These LLMs exhibit "Emergent Abilities" (Wei et al., 2022a) like In-Context Learning (Dong et al., 2023) and Chain-of-Thought Reasoning (Wei et al., 2022b). Despite these advancements, a challenge is the generation of text that deviates from human knowledge,

known as *hallucination* (Bang et al., 2023; Li et al., 2022). Even the latest LLMs, such as GPT-4 (OpenAI, 2023), still suffer from hallucinations, which greatly damages the factuality of the generated text.

In this paper, we propose a unified and flexible pipeline UFO to evaluate the factuality of the generated texts, which can detect hallucinations in various text generation tasks.

2.2 Factuality Evaluation

Factuality evaluation methods have evolved from traditional n-gram-based metrics to more sophisticated approaches leveraging PLMs and LLMs (Li et al., 2022). Initially, metrics such as BLEU (Papineni et al., 2002), ROUGE (Lin and Och, 2004), and METEOR (Banerjee and Lavie, 2005) assumed factual accuracy correlated with n-gram overlaps. Later, metrics like BERTScore (Zhang et al., 2019) utilizing contextual embeddings, and BARTScore (Yuan et al., 2021) employing generative scoring, captured deep semantic information between texts for evaluating factuality consistency. QAGS (Wang et al., 2020) further innovates by combining entity extraction with PLM-based question generation and answering, while Q^2 (Honovich et al., 2021) leverages natural language inference (NLI) for entailment analysis. More recently, LLM-based metrics such as FactScore (Min et al., 2023) and FacTool (Chern et al., 2023) utilize LLM's reasoning ability, extracting and verifying facts against sources like Wikipedia dumps.

Different from previous studies, our proposed pipeline UFO integrates human-written evidence, reference documents, search engine results, and LLM knowledge for factuality evaluation.

3 Methodology

3.1 Problem Statement

Given a query q_D sourced from a dataset D, an evaluated LLM M generates a text passage $T_M(q_D)$. We define a collection of fact sources, denoted as S. The objective is to assign a factuality score $s \in [0, 1]$ to the model-generated text $T_M(q_D)$. A higher score denotes a greater consistency between the text $T_M(q_D)$ and the fact sources S, indicating higher factual accuracy of the LLM M.

3.2 Fact Sources

Based on the origin of fact sources, we categorize them into four types: human-written evidence

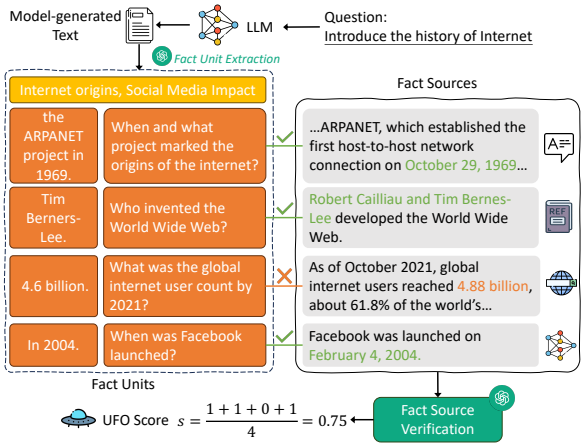

Figure 2: A case of evaluating Vicuna-generated text within the retrieval-augmented QA task where $S_D = \{S_{\text{he}}, S_{\text{rd}}\}$. Details of the generated text are omitted for clarity. The extracted answers are underlined.

(S_{he}), reference documents (S_{rd}), search engine results (S_{se}), and LLM knowledge (S_{lk}). Each type of fact source contains a series of text passages $\{P^1, P^2, \cdots\}$. The first two types of fact sources (S_{he} and S_{rd}) are provided by established datasets and require some cost to collect, such as responses and evidence written by users, and selected reference documents while they browse web pages. The latter two (S_{se} and S_{lk}) are fact sources relevant to specifically generated questions, independent of any particular dataset. These include text snippets retrieved from the web corpus and answers from the parameterized knowledge within LLMs.

For a given question, it might not be possible to obtain an answer from a certain fact source. Therefore, in an evaluation scenario, we predefine a sequence of fact sources $S = \langle S^1, S^2, \cdots \rangle$, and systematically verify each until a matched answer is extracted.

3.3 UFO Evaluation Framework

Our evaluation pipeline includes three LLM-based modules: Fact Unit Extraction, Fact Source Verification, and Fact Consistency Discrimination. We employ OpenAI's ChatGPT API (`gpt-3.5-turbo-1106`) for these modules.

3.3.1 Fact Unit Extraction

LLMs can generate a text with several sentences for a given input, but not all the generated sentences are fact-related. Therefore, our first problem is to determine the smallest unit for factuality evaluation. We start by analyzing the process of factuality evaluation performed by humans. When faced with

a text, humans will first focus on entities and their relevant descriptions that may cause factual errors. Then, they will ask a series of questions about the factuality of these descriptions. For example, when a text describes the date of birth D of a famous person X, a common question is *"when was X born?"*. Finally, by comparing the golden answer D' (from their knowledge or Internet) with D, the factuality of the description can be evaluated.

Based on these analyses, we consider an entity-centric question q_k and its corresponding answer e_k can be used as a basic fact unit $f_k = \langle q_k, e_k \rangle$. However, a generated question may have a weak relationship with its context. Therefore, we generate concise keywords of model-generated text. For instance, a model-generated text is related to the demand for vinyl records in the Oxfam Charity Shop. the generated question, *"What has led to a rise in demand for vinyl records?"* yields search engine results that discuss the recent global music trends. Thus, it is necessary to generate concise keywords t for the model-generated text to refine the search engine's capability in retrieving content specifically relevant to the Oxfam Charity Shop.

Benefiting from the potent language comprehension capabilities of LLMs, we introduce an LLM-based Fact Unit Extraction (FUE) method to extract the fact units (Prompt 1) and generate the keywords (Prompt 2) from the model-generated text. These prompts are provided in Appendix A.

$$\{t, \langle q_1, e_1 \rangle, \cdots, \langle q_p, e_p \rangle\} = \text{FUE}(T_{M_i}(q_D)).$$

Next, we will utilize the fact source sequence S in different scenarios to evaluate the factual accuracy of these fact units.

3.3.2 Fact Source Verification

To verify the accuracy of a given fact unit $\langle q_i, e_i \rangle$ against the keywords t, our target is to identify the correct answer a_i to the question q_i using a specific text passage P_j^k from a fact source S^k. However, not all text passages in the fact source are relevant to the question. To enhance the relevance of extracted answers, we employ the advanced language comprehension abilities of LLMs. We instruct the LLM-based Answer Extraction (AE) module to pinpoint the most relevant answers within the text, generating a "[NOANS]" text if no answer is found. This method involves directly prompting an LLM to retrieve answers from human-written evidence S_{he} and reference documents S_{rd} (Prompt 3), reducing inaccuracies during fact verification (Huang et al., 2023). When dealing with search engine results S_{se} and LLM knowledge S_{lk}, we prompt the model to first check if an answer is available in the search results before resorting to its internal knowledge (Prompt 4). Answers are sequentially sought in each text passage of the fact source until a suitable answer is found. If no text passage yields an answer, it indicates a mismatch with the fact source, leading to a transition to the next fact source $S^{k+1} \in S$ for verification.

Concretely, for a fact unit $\langle q_i, e_i \rangle$ and keywords t, we obtain the answer a_i using passage P_j^k from fact source S^k as follows:

$$a_i = \text{AE}(t, P_j^k, q_i), \tag{1}$$

$$\text{AE}(t, P_m^k, q_i) = \text{[NOANS]}, \tag{2}$$

$$m \in \{1, \cdots, j-1\}. \tag{3}$$

3.3.3 Fact Consistency Discrimination

Given the answer e_i extracted from the model-generated text and the answer a_i extracted from fact sources, our objective is to determine whether the two answers are factually consistent. To achieve this, we employ an LLM-based fact consistency discrimination (FCD) module (Prompt 5), assigning a score of 0 or 1 to each fact unit $\langle q_i, e_i \rangle$. Subsequently, we calculate the average score of all fact units as the factuality score of the model-generated text:

$$s_i = \text{FCD}(e_i, a_i) \in \{0, 1\}, \tag{4}$$

$$s = \frac{1}{N} \sum_{i=1}^{N} s_i. \tag{5}$$

3.4 Evaluation Criteria

We measure the discriminative power (DP) of the evaluation metric, as described by Sakai (2006). Given the collection of evaluated LLMs M and all pairs $(M_i, M_j) \subset M$, we bootstrap sample the evaluation score on M_i and M_j. Then, given a threshold value f, we obtain minority rate (MR) and proportion of ties (PT) values. The MR represents the failure rate of distinguishing the evaluation score differences between a pair of LLMs within the threshold. The PT indicates the percentage of cases where the pair of LLMs cannot be distinguished within the threshold. The smaller the values of MR and PT, the stronger the discriminative power of the evaluation metric. Finally, we have the MR-PT curve as the discriminative power of the evaluation metric. The details of the

pseudocode of DP measurement are given in Appendix B.

3.5 Evaluation Scenarios

To assess the importance of each fact source across various tasks, we introduce five evaluation scenarios, each represented by an ordered list of fact sources S. (1) $S = \langle S_{\text{se}}, S_{\text{lk}} \rangle$. (2) $S = \langle S_{\text{he}}, S_{\text{se}}, S_{\text{lk}} \rangle$. (3) $S = \langle S_{\text{rd}}, S_{\text{se}}, S_{\text{lk}} \rangle$. (4) $S = \langle S_{\text{he}}, S_{\text{rd}}, S_{\text{se}}, S_{\text{lk}} \rangle$. (5) $S = \langle S_{\text{rd}}, S_{\text{he}}, S_{\text{se}}, S_{\text{lk}} \rangle$. To thoroughly verify facts in the text and mitigate the hallucination of LLM knowledge, we retain and fix the verification order of S_{se} and S_{lk}. By comparing the DP in scenarios (1), (2), and (3), we can infer the impact of fact sources. Comparing the DP in scenarios (4) and (5) reveals the effects of changing the verification order of fact sources.

Moreover, LLMs incorporating web search modules, such as Bing Chat, have been able to generate text while providing retrieved reference documents. In Section 5.2, we will discuss the impact of using these referenced documents as the supplementary fact source S_{rd} in evaluation scenarios.

S_{he}	✗	✗	✓	✓	✓	✓
S_{rd}	✗	✓	✗	✓	✓	✓
Dataset	NQ	HQA	TQA	C/D	M-N	MS
Avg. # of Tokens						
Bing Chat	136.96	87.99	196.02	223.64	248.66	287.93
ChatGPT	398.36	336.26	393.41	534.74	531.17	561.91
llama-7b	455.14	427.46	453.58	433.66	436.47	459.84
llama-13b	431.99	406.42	432.24	422.61	424.83	455.64
vicuna-7b	353.72	332.16	387.88	398.84	401.92	413.28
vicuna-13b	341.00	327.38	346.32	366.60	369.31	398.76
Avg. # of Sentences						
Bing Chat	5.15	3.42	7.50	8.00	8.20	10.54
ChatGPT	12.46	10.05	12.62	15.47	15.57	18.66
llama-7b	17.34	15.80	18.41	15.73	16.21	18.85
llama-13b	15.88	14.54	16.70	14.64	14.86	17.19
vicuna-7b	12.25	11.39	14.50	13.03	13.17	15.98
vicuna-13b	11.72	10.85	12.31	12.16	12.10	14.77
Avg. # of Facts Extracted Using *ChatGPT*						
Bing Chat	4.12	3.08	4.78	5.32	5.64	5.85
ChatGPT	5.60	5.17	5.41	5.94	5.92	6.04
llama-7b	4.86	4.97	5.06	5.15	5.18	5.30
llama-13b	5.14	5.11	5.00	4.96	5.28	5.11
vicuna-7b	5.76	5.66	5.34	6.04	6.18	5.46
vicuna-13b	5.51	5.56	5.26	5.90	5.66	5.62

Table 2: Statistics of model-generated text on six datasets. "HQA", "TQA", "C/D", "M-N", and "MS" are abbreviations of "HotpotQA", "TruthfulQA", "CNN/DM", "Multi-News" and "MS MARCO".

4 Experiments

4.1 Datasets

Considering the available human-written evidence and reference documents, we categorize tasks presented in Table 1. We carry out our evaluation pipeline on six datasets: NQ (Lee et al., 2019), HotpotQA (Yang et al., 2018), TruthfulQA (Lin et al., 2022), CNN/DM (Hermann et al., 2015), Multi-News (Fabbri et al., 2019), and MS MARCO (Bajaj et al., 2016). We collect 200 samples from each dataset and prompt evaluated LLMs to generate verifiable facts in sufficient detail (Prompt 6).

To compare with reference-based metrics, we construct a golden answer G containing more facts for each task. **(1) Open-domain QA**: In the NQ dataset, we concatenated the provided short answers to form G. **(2) Web Retrieval-based QA**: In the HotpotQA dataset, we combined the short answer and the reference documents as the golden answer $G = [a; S_{\text{rd}}]$. **(3) Expert-validated QA**: In the TruthfulQA dataset, all provided human-written correct answers and best answers were considered as the fact source S_{he}, forming the golden answer G. **(4) News Fact Generation**: For the CNN/DM and Multi-News datasets, we first prompted ChatGPT (Prompt 7) to generate a title of the given summary (considered as S_{he}). Then, we used the generated title to prompt the evaluated LLMs (Prompt 6) to generate an introduction centered around the facts. **(5) Retrieval-Augmented QA**: In the MS MARCO dataset, the answer a was regarded as S_{he}, and all user-clicked documents were considered as S_{rd}. The answer and the selected documents were concatenated to form G.

4.2 Baselines

Evaluated Models We evaluate six existing LLMs with varying parameter scales in our experiments: (1) Bing Chat is a GPT-4-based model specifically tailored for web searches. For this model, we choose the "Precise" generation mode to test the factuality when the model is expected to generate the most accurate and detailed fact units.[2] In each provided URL, we extract all the <p> tags of the corresponding web page. Subsequently, we divide the text into multiple passages, each containing no more than 1024 tokens. (2) ChatGPT: we utilized OpenAI's ChatGPT API (`gpt-3.5-turbo-1106`) for text generation.[3] (3) LLaMA (Touvron et al., 2023): We select two

[2] https://blogs.bing.com/search/march_2023/Confirmed-the-new-Bing-runs-on-OpenAI%E2%80%99s-GPT-4

[3] https://platform.openai.com/docs/api-reference/chat

LLaMA-series fine-tuned models for text generation (LLaMA-2-7b-chat and LLaMA-2-13b-chat). (4) Vicuna (Chiang et al., 2023): Vicuna is a chat assistant developed by fine-tuning LLaMA-2 foundation model with user-shared conversations collected from ShareGPT.[4] We select Vicuna-7b-v1.5 and Vicuna-13b-v1.5 to generate text. The statistical data of the text generated by these LLMs is given in Table 2.

Baseline Evaluation Metrics We compare our proposed pipeline with both reference-based and reference-free metrics.

(1) Reference-based metrics. Such metrics require a golden answer G and calculate the consistency with the model-generated text. BLEU (Papineni et al., 2002) and ROUGE (Lin and Och, 2004) are used to measure the token-level term overlap. BERTScore (Zhang et al., 2019) and BARTScore (Yuan et al., 2021) are model-based metrics to evaluate passage-level similarity. QAGS (Wang et al., 2020) and Q^2 (Honovich et al., 2021) are the most relevant PLM-based and NLI-based metrics to evaluate factuality.

(2) Reference-free metrics. FactScore (Min et al., 2023) first breaks down the model-generated text into several claims. Subsequently, these claims are verified through Wikipedia dumps. In this study, we form all golden answers as the corpus for FactScore verification. FacTool (Chern et al., 2023) performs the verification of each claim by employing a search engine and derives factuality scores at the claim level.

5 Results and Analysis

5.1 Discriminative Power Results

Our goal is to evaluate the discriminative power (Sakai, 2006) of the proposed evaluation pipeline UFO in each scenario. For simplicity, in the scenario such as $S = \langle S_{\text{se}}, S_{\text{lk}} \rangle$, we name our framework "ufo(se+lk)". The experimental results are shown in Figure 3. Each point on the curve represents the values of MR and PT calculated at a given threshold f. We have the following findings:

(1) Among baseline metrics, BARTScore demonstrates minimal variance with a notably low MR value, while QA-based metrics like QAGS and Q^2 show suboptimal discriminative power across all tasks. It indicates that PLM-based methods are particularly reliant on the quality of the golden answer, especially its entities and relationships. FactScore (Min et al., 2023) verifies each extracted claim against a predefined fact source and enhances the LLM-based method's capability through In-Context Learning with demonstrations. Thus it shows relatively high discriminative power with additional time and token usage. We will discuss the API and time usage in Section 5.4.

(2) HotpotQA and TruthfulQA dataset provide S_{rd} and S_{he} respectively. However, the proposed pipeline UFO in the scenario $\langle S_{\text{rd}}, S_{\text{se}}, S_{\text{lk}} \rangle$ in HotpotQA resulted in weaker DP. It suggests that the quality of S_{rd} is inferior to search engine results S_{se} and LLM knowledge S_{lk}. Search engines can retrieve more relevant details for entities involved in multi-hop reasoning. In TruthfulQA, there is a significant presence of questions with confusion, hence the fact source S_{he} has higher quality. UFO in the scenario $\langle S_{\text{he}}, S_{\text{se}}, S_{\text{lk}} \rangle$ show a substantial increase in DP compared to the scenario without S_{he}. This also indicates the necessity of incorporating human-written evidence in expert-validated QA tasks.

(3) CNN/DM, Multi-News, and MS MARCO provide both S_{he} and S_{rd}. However, in the task of news fact generation, UFO in the scenario $S = \langle S_{\text{se}}, S_{\text{lk}} \rangle$ achieves the highest DP, indicating that both S_{he} and S_{rd} exhibit a negative impact on DP. In fact, within provided documents and summaries, the factual details are considerably limited and not specific enough. Thus, employing search engines enables precise retrieval of factual details based on specific extracted questions. In the retrieval-augmented QA task, S_{he} and S_{rd} significantly enhance DP. Moreover, changing the order of verification between S_{he} and S_{rd} does not significantly affect the DP value. This reflects a high degree of consistency between user-clicked reference documents and human-written evidence. It suggests that for this task, reference documents and human-written evidence can be substituted for each other, and we can rely solely on user-clicked reference documents without the need for collecting human-written evidence.

5.2 Effect of Model-Retrieved Documents

Some existing LLMs provide retrieved reference documents during text generation. We incorporate these as part of S_{rd} to evaluate the certain LLM (*i.e.*, Bing Chat in our experiments). The discriminative power of LLM-based evaluation metrics is shown in Figure 4. We have the following findings:

(1) In the NQ, HotpotQA, and TruthfulQA

[4] https://sharegpt.com/

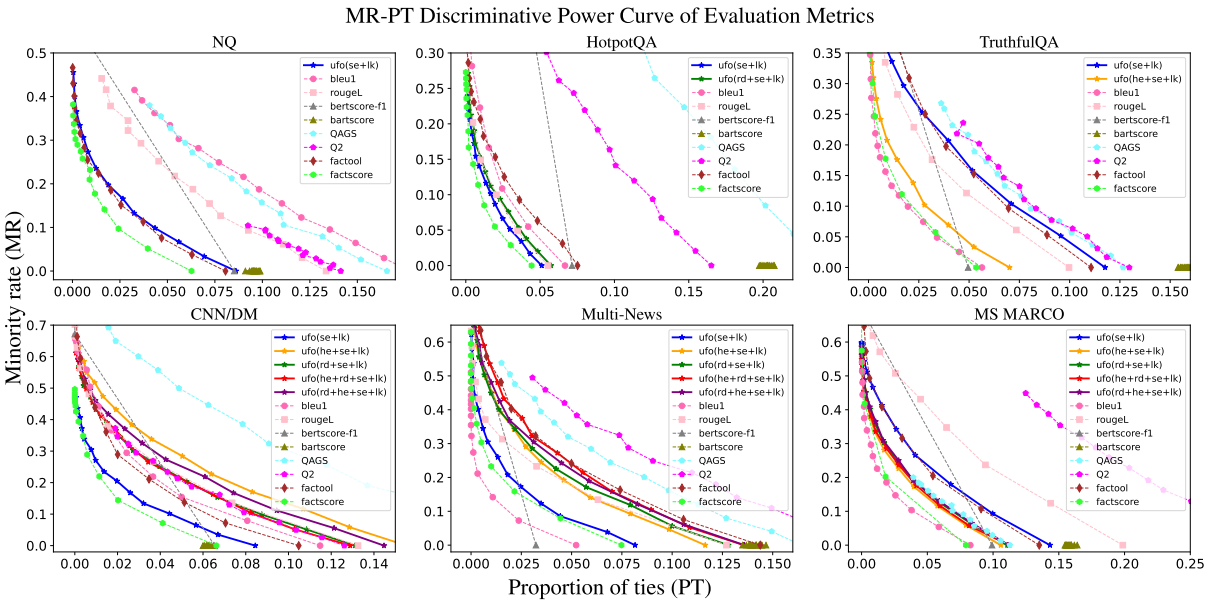

Figure 3: MR-PT discriminative power curve of evaluation metrics on datasets. The closer the curve is to the bottom-left corner, the better the evaluation metric is. Our proposed model curve is represented by a **solid line**, while the baseline model curve is depicted using a **dashed line**.

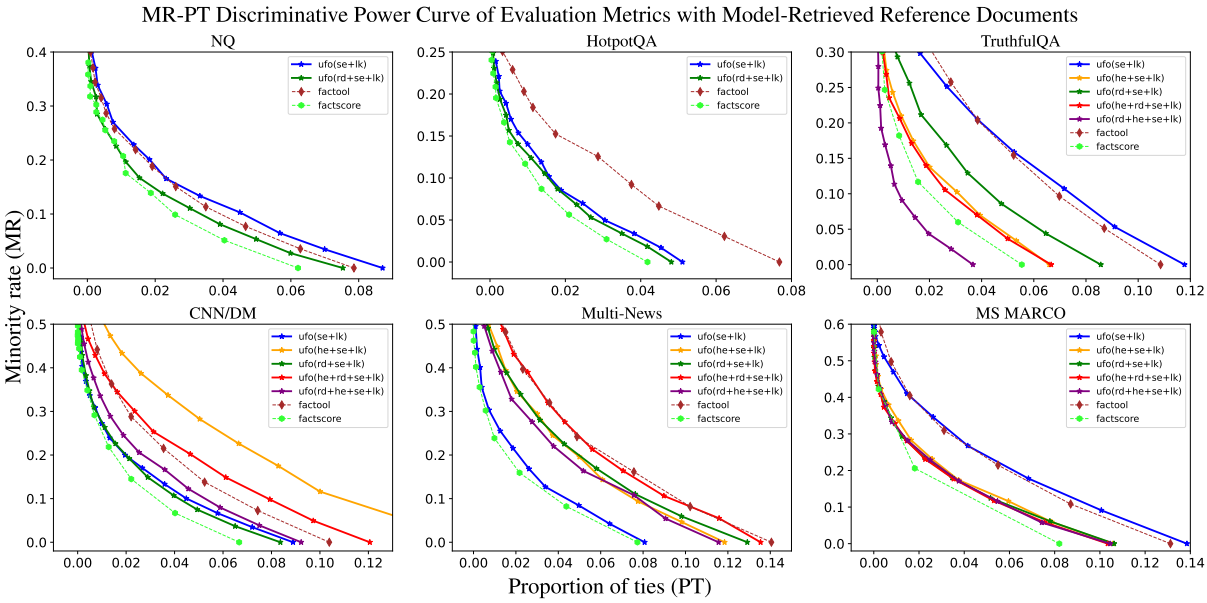

Figure 4: MR-PT discriminative power curve of evaluation metrics on datasets. The closer the curve is to the bottom-left corner, the better the evaluation metric is. We incorporate reference documents retrieved by Bing Chat as part of the fact source S_{rd}. Non-LLM-based methods are omitted for clarity.

Scenarios	Pearson ↑		Spearman ↑		#Avg. Tokens ↓		#Avg. Time ↓	
	FacTool	FactScore	FacTool	FactScore	FacTool	FactScore	FacTool	FactScore
$\langle S_{\text{se}}, S_{\text{lk}} \rangle$	**0.296**	0.272	**0.308**	0.283	+1.20%	-6.71%	**-3.44%**	**-22.31%**
$\langle S_{\text{he}}, S_{\text{se}}, S_{\text{lk}} \rangle$	0.275	0.269	0.276	0.270	**-3.79%**	**-12.21%**	-2.03%	-19.26%
$\langle S_{\text{rd}}, S_{\text{se}}, S_{\text{lk}} \rangle$	0.279	0.278	0.283	0.290	+12.33%	-9.97%	-2.31%	-16.02%
$\langle S_{\text{he}}, S_{\text{rd}}, S_{\text{se}}, S_{\text{lk}} \rangle$	0.263	0.269	0.271	0.281	+6.25%	-11.03%	+2.43%	-18.75%
$\langle S_{\text{rd}}, S_{\text{he}}, S_{\text{se}}, S_{\text{lk}} \rangle$	0.267	**0.283**	0.271	**0.295**	+11.20%	-10.35%	+7.13%	-19.63%

Table 3: Pearson's and Spearman's correlation coefficients for UFO and the LLM-based baseline evaluation metrics under five evaluation scenarios. All p-values are less than 0.01. Additionally, we compared the usage of ChatGPT API tokens and the average time required for evaluating samples. The best results are marked **bold**.

Dataset	NQ		HotpotQA		TruthfulQA		CNN/DM		Multi-News		MS MARCO	
Metrics	UFO	FT	UFO	FT	UFO	FT	UFO	FT	UFO	FT	UFO	FT
Bing Chat	0.752	0.615	0.630	0.709	0.649	0.594	0.628	0.742	0.685	0.745	0.725	0.787
ChatGPT	**0.762**	**0.776**	**0.635**	**0.725**	0.662	**0.700**	**0.669**	**0.806**	**0.708**	**0.806**	**0.765**	**0.845**
llama-7b	0.610	0.480	0.465	0.453	0.630	0.560	0.607	0.673	0.589	0.689	0.711	0.757
llama-13b	0.674	0.596	0.537	0.537	0.597	0.564	0.573	0.688	0.661	0.731	0.735	0.731
vicuna-7b	0.670	0.631	0.515	0.562	0.692	0.610	0.603	0.714	0.593	0.688	0.662	0.755
vicuna-13b	0.676	0.658	0.514	0.570	**0.717**	0.664	0.652	0.740	0.646	0.731	0.739	0.778

Table 4: Factuality scores of our proposed evaluation framework UFO in the scenario of $S = \langle S_{\text{he}}, S_{\text{rd}}, S_{\text{se}}, S_{\text{lk}} \rangle$ and FacTool (abbreviated to "FT") on six datasets. In the evaluation of the group of evaluated LLMs, the highest factuality score is **bold**, and the second highest score is underlined.

datasets, incorporating retrieved reference documents in the evaluation scenarios enhances the DP. Specifically on the TruthfulQA dataset, we further observe that UFO in the scenario $\langle S_{\text{rd}}, S_{\text{he}}, S_{\text{se}}, S_{\text{lk}} \rangle$ significantly boosts discriminative power and surpasses FactScore. This implies that the model accurately retrieves relevant reference documents based on easily confused facts during text generation.

(2) In the CNN/DM dataset, UFO in the scenario $\langle S_{\text{rd}}, S_{\text{se}}, S_{\text{lk}} \rangle$ shows a slight improvement when incorporating model-retrieved documents compared to the use of search engines and LLM knowledge $\langle S_{\text{se}}, S_{\text{lk}} \rangle$. It suggests that retrieved reference documents serve as a beneficial complement to search engine results in this task.

(3) In the MS MARCO dataset, the enhancement of DP brought by incorporating retrieved reference documents is minimal, indicating a factual consistency between human-written evidence, clicked reference documents, and retrieved reference documents.

5.3 Factuality Scores of LLMs

In addition to evaluating discriminative power, we also calculated the factuality scores of evaluated LLMs on various datasets. Under the evaluation scenario $S = \langle S_{\text{he}}, S_{\text{rd}}, S_{\text{se}}, S_{\text{lk}} \rangle$, the comparative experimental results between our proposed framework UFO and FacTool are presented in Table 4. Both evaluation methods show that in most datasets, the factuality score of Bing Chat in "precise" mode is slightly lower than that of ChatGPT, but close to the score of Vicuna-13b. This implies that hallucinations occur during the retrieval-augmented generation process, thereby reducing the factual accuracy of the generated text. We also observe that increasing the parameter scale of open-source LLMs (LLaMA and Vicuna) can enhance factual accuracy in most datasets.

5.4 Cost of LLM-based Metrics

Our proposed pipeline sequentially extracts answers from listed text passages in the fact source. To assess the efficiency of our proposed evaluation pipeline, we compared the average evaluation time, API token costs, and the correlation coefficient with existing LLM-based evaluation metrics. Experimental results (shown in Table 3) demonstrate that our proposed pipeline achieves the highest correlation coefficient with FacTool and FactScore in the scenarios where the fact sources are set to $\langle S_{\text{se}}, S_{\text{lk}} \rangle$ and $\langle S_{\text{rd}}, S_{\text{he}}, S_{\text{se}}, S_{\text{lk}} \rangle$ respectively.

Besides, in comparison to FactScore, our proposed evaluation pipeline reduces token consumption by about 10% and time cost by about 20%, while our proposed pipeline maintains a relatively comparable discriminative power. Compared to FacTool, in most five evaluation scenarios, we achieved greater discriminative power with an affordable additional cost of about 10% in all tasks. It implies that the incorporation of fact sources can enhance the discriminative power of evaluation metrics.

6 Conclusion

In this paper, we propose UFO, a factuality evaluation pipeline incorporating flexible plug-and-play fact sources: human-written evidence, reference documents, search engine results, and LLM knowledge with unified verification methods. Experimental results on five evaluation scenarios show that open-domain QA, web retrieval-based QA, and expert-validated QA tasks require high-quality human-written evidence and model-retrieved reference documents, and retrieval-augmented QA needs either human-written evidence or user-clicked reference documents, while news fact generation tasks rely on search engine results and LLM knowledge.

Limitations

Though our proposed pipeline analyzes different fact sources, there are still several limitations: (1) We utilize ChatGPT (`gpt-3.5-turbo-1106`) in the modules of our proposed evaluation framework, therefore the updating of the LLM will affect both the cost and effectiveness of our evaluation. (2) Currently, our approach has not yet been integrated with the training process of LLMs. In future work, we will consider incorporating the factuality score evaluated by our framework into the training process of the LLMs through reinforcement learning methods.

Ethics Statement

The datasets used in this paper are available publicly online. Particularly, any data involving sensitive information has been anonymized, ensuring that it cannot be traced back to individuals.

References

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, et al. 2016. Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization@ACL 2005, Ann Arbor, Michigan, USA, June 29, 2005*, pages 65–72. Association for Computational Linguistics.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher 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.

I-Chun Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, Pengfei Liu, et al. 2023. Factool: Factuality detection in generative ai–a tool augmented framework for multi-task and multi-domain scenarios. *arXiv preprint arXiv:2307.13528*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. 2023. A survey on in-context learning.

Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1074–1084, Florence, Italy. Association for Computational Linguistics.

Xue-Yong Fu, Md. Tahmid Rahman Laskar, Cheng Chen, and Shashi Bhushan TN. 2023. Are large language models reliable judges? A study on the factuality evaluation capabilities of llms. *CoRR*, abs/2311.00681.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6465–6488, Singapore. Association for Computational Linguistics.

Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 1693–1701, Cambridge, MA, USA. MIT Press.

Or Honovich, Leshem Choshen, Roee Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. 2021. q^2: Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7856–7870, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12):248:1–248:38.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6086–6096, Florence, Italy. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Wei Li, Wenhao Wu, Moye Chen, Jiachen Liu, Xinyan Xiao, and Hua Wu. 2022. Faithfulness in natural language generation: A systematic survey of analysis, evaluation and optimization methods.

Chin-Yew Lin and Franz Josef Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, 21-26 July, 2004, Barcelona, Spain*, pages 605–612. ACL.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Xiao Liu, Hanyu Lai, Hao Yu, Yifan Xu, Aohan Zeng, Zhengxiao Du, Peng Zhang, Yuxiao Dong, and Jie Tang. 2023. Webglm: Towards an efficient web-enhanced question answering system with human preferences.

Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, Francis Song, Martin Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, and Nat McAleese. 2022. Teaching language models to support answers with verified quotes.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 12076–12100. Association for Computational Linguistics.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2022. Webgpt: Browser-assisted question-answering with human feedback.

OpenAI. 2023. Gpt-4 technical report.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

Hongjing Qian, Yutao Zhu, Zhicheng Dou, Haoqi Gu, Xinyu Zhang, Zheng Liu, Ruofei Lai, Zhao Cao, Jian-Yun Nie, and Ji-Rong Wen. 2023. Webbrain: Learning to generate factually correct articles for queries by grounding on large web corpus.

Yujia Qin, Zihan Cai, Dian Jin, Lan Yan, Shihao Liang, Kunlun Zhu, Yankai Lin, Xu Han, Ning Ding, Huadong Wang, Ruobing Xie, Fanchao Qi, Zhiyuan Liu, Maosong Sun, and Jie Zhou. 2023. WebCPM: Interactive web search for Chinese long-form question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8968–8988, Toronto, Canada. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Tetsuya Sakai. 2006. Evaluating evaluation metrics based on the bootstrap. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '06, page 525–532, New York, NY, USA. Association for Computing Machinery.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022a. Emergent abilities of large language models. *Trans. Mach. Learn. Res.*, 2022.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022b. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 27263–27277.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with BERT. *CoRR*, abs/1904.09675.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models.

A Prompt

Prompt 1: Fact unit extraction using ChatGPT

Your task is to segment a given document into several atomic claims. For each claim, you need to generate several questions related to it and extract an answer for each question from that claim. Your output is a JSON list. Each element includes the question, the answer, and the sentence from the document containing the atomic claims.
You MUST only respond in the JSON List format as described below. DO NOT RESPOND WITH ANYTHING ELSE. ADDING ANY OTHER EXTRA NOTES THAT VIOLATE THE RESPONSE FORMAT IS BANNED. START YOUR RESPONSE WITH '['. [Response Format] ["question": "Informative question", "answer": "A concise phrase under 10 words", "sentence": "Sentence containing the answer.", ...]
document: {document}

Prompt 2: Generate keywords for model-generated text

Generate keywords for the following document. Do not provide any explanations.
document: {document}
keywords:

Prompt 3: LLM-based answer extraction with S_{he} and S_{rd}

You are an answer-extraction expert. Your task is to extract a short answer from the evidence to the question. Directly answer without any explanations. If the evidence is irrelevant to the question, respond ONLY with "NOANS".
keywords: {keywords}
evidence: {evidence}
question: {question}
your answer:

Prompt 4: LLM-based answer extraction with S_{se} and S_{lk}

You are a question-answering expert. You are given a question, keywords, and some snippets. Your task is to output a short answer to the question based on the snippets or the knowledge you possess, while your answer is factually consistent with the given keywords. If your answer is based on the snippets, you should provide the indices of the snippets. If there is no relevant snippet, you should answer with the knowledge you possess, and the output index is [-1]. If you are uncertain about the correctness and timeliness of your answer, your answer should be formed as [NOANS] instead. An example output format: [<your answer>]; [<index1>, <index2>, ...]. Your output MUST begin with '['. DO NOT GIVE ANY EXPLANATIONS.
question: {question}
keywords: {keywords}
snippets: {snippets}

Prompt 5: LLM-based fact consistency discrimination

Your task is to judge whether the following two answers are factually consistent. Directly answer yes or no.
Answer 1: $\{e_i\}$
Answer 2: $\{a_i\}$

Prompt 6: Generate long-form text

You have been presented with the following title. Your task is to provide a comprehensive introduction to the query topic with sufficient verifiable facts based on the knowledge you possess. Your output must be in English.
Title: {title}
Introduction:

Prompt 7: Generate title for model-generated text

Generate a summarized title for the following document. Do not provide any explanations.
document: {document}
title:

B Pseudocode for DP Measurement

Algorithm 1 Discriminative Power Measurement with MR-PT Curve

for each $threshold \in \{0, 0.01, \cdots, 0.20\}$ **do**
  $f \leftarrow threshold$
  $B \leftarrow 1000$
  for each $(M_i, M_j) \in M$ **do**
    for $b = 1$ to B **do**
      $Q_i = mean(Bootstrap(M_i))$
      $Q_j = mean(Bootstrap(M_j))$
      $m = f * \max(Q_i, Q_j)$
      if $|Q_i - Q_j| < m$ **then**
        $EQ(i,j) \leftarrow EQ(i,j) + 1$
      else if $Q_i > Q_j$ **then**
        $GT(i,j) \leftarrow GT(i,j) + 1$
      else
        $GT(j,i) \leftarrow GT(j,i) + 1$
      end if
    end for
  end for
  $MR_f \leftarrow \frac{\sum_{M_i,M_j} \min(GT(i,j),GT(j,i))}{B \sum_{M_i,M_j}}$
  $PT_f \leftarrow \frac{\sum_{M_i,M_j} EQ(i,j)}{B \sum_{M_i,M_j}}$
end for