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 six evaluation scenarios based on this framework. Experimental results show that human-written evidence and reference documents are crucial in most QA tasks, but in the news fact generation tasks, introducing human-written evidence leads to a decline in the discriminative power of evaluation. Compared to the LLM knowledge, search engine results are more important in most tasks, but they are less effective in the expert-validated QA task. Our dataset and code are available at https://anonymous.4open.science/r/UFO-813F.

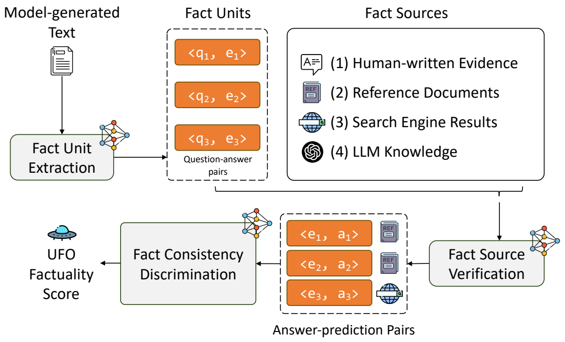


Figure 1: Our proposed factuality evaluation pipeline UFO. We integrate four fact sources within various evaluation scenarios to assess the factuality score.

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 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: (a) Integrates various fact sources flexibly.

[1] ChatGPT: https://chat.openai.com/chat, Bing Chat: https://copilot.microsoft.com/

(b) Uses a unified verification method for switching fact sources in specific tasks. (c) Combines different fact sources to enhance the factuality evaluation. In our framework, as shown in Figure 1, we first extract fact units from the model-generated text, including question-answer pairs. Then, we verify each fact 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.*, 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 six 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 (DP) (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 fact sources in different scenarios. Experimental results indicate that in most QA tasks, human-written evidence and reference documents enhance the DP of the evaluation pipeline. However, in news facts generation tasks, human-written evidence leads to a performance decline. Search engine results are generally more important than LLM knowledge but are less effective in expert-validated QA tasks. Although not the main focus of this paper, we evaluate nine existing LLMs: Bing Chat in "precise" generation mode, ChatGPT, LLaMA2-{7,13,70}B, LLaMA3-{8,70}B, and Qwen-{7,14}B. We discovered that the factuality score of ChatGPT is higher than Bing Chat in precise mode, yet comparable to LLaMA3-8B. The factuality score of LLaMA3 outperforms that of LLaMA2 and Qwen at a similar parameter scale. In open-source LLMs, increasing the scale of parameters can enhance factual accuracy.

Tasks	(1) Open-domain QA; (2) Web retrieval-based QA; (3) Expert-validated QA; (4) News fact generation; and (5) Retrieval-augmented QA.
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) $\langle S_{\text{lk}}, S_{\text{se}}\rangle$; (3) $\langle S_{\text{he}}, S_{\text{se}}, S_{\text{lk}}\rangle$; (4) $\langle S_{\text{rd}}, S_{\text{se}}, S_{\text{lk}}\rangle$; (5) $\langle S_{\text{he}}, S_{\text{rd}}, S_{\text{se}}, S_{\text{lk}}\rangle$; (6) $\langle S_{\text{rd}}, S_{\text{he}}, S_{\text{se}}, S_{\text{lk}}\rangle$.

Table 1: The tasks, fact sources, and evaluation scenarios 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 LLM factuality.
- We conduct a systematic analysis of the evaluation capabilities of four fact sources in six factuality evaluation scenarios and five tasks.
- We reveal that human-written evidence and reference documents are crucial for most QA tasks, while human-written evidence reduces the discriminative power of evaluation in news fact generation tasks. Search engine results are generally more effective than LLM knowledge, but LLM knowledge is more important in expert-validated QA 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 damage 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

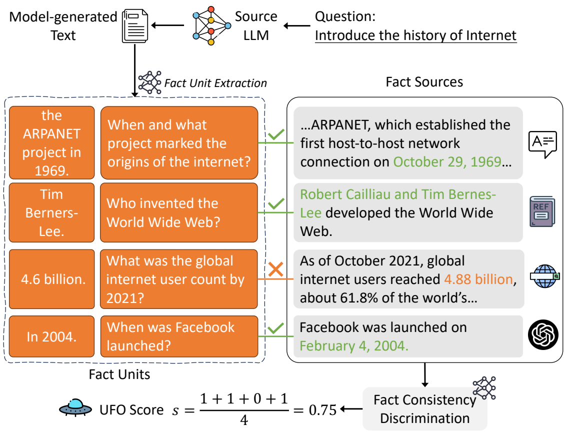
Factuality evaluation methods have evolved from traditional n-gram-based metrics to more sophisticated approaches leveraging PLMs and LLMs (Li et al., 2022). Initially, metrics such as BLEU (Papineni et al., 2002), ROUGE (Lin and Och, 2004), and METEOR (Banerjee and Lavie, 2005) assumed factual accuracy 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 question q_D sourced from a dataset D, a source LLM M generates a text passage $T_M(q_D)$. We define a list of fact sources, denoted as $S = \langle S^1, S^2, \cdots \rangle$. The objective is to assign a factuality score $s \in [0, 1]$ to the model-generated text $T_M(q_D)$. As we demonstrate in Figure 2, a higher score indicates greater consistency between the text $T_M(q_D)$ and the fact sources S, indicating higher factual accuracy of the source LLM M.



Model-generated Text
Source LLM
Question: Introduce the history of Internet
Fact Unit Extraction
Fact Sources
the ARPANET project in 1969.
When and what project marked the origins of the internet?
...ARPANET, which established the first host-to-host network connection on October 29, 1969...
Tim Berners-Lee.
Who invented the World Wide Web?
Robert Cailliau and Tim Bernes-Lee developed the World Wide Web.
REF
4.6 billion.
What was the global internet user count by 2021?
As of October 2021, global internet users reached 4.88 billion, about 61.8% of the world's...
In 2004.
When was Facebook launched?
Facebook was launched on February 4, 2004.
Fact Units
UFO Score $s = \frac{1+1+0+1}{4} = 0.75$
Fact Consistency Discrimination

Figure 2: A case of evaluation within the retrieval-augmented QA task where $S = \langle S_{\text{he}}, S_{\text{rd}}, S_{\text{se}}, S_{\text{lk}} \rangle$. Details of the generated text are omitted for clarity. The extracted answers are highlighted.

3.2 Fact Sources

Based on the origin of fact sources, we categorize them into four types: human-written evidence (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}) can be automatically collected or generated. These include text snippets retrieved from the web corpus and passages from the parameterized knowledge within LLMs. Specifically, for search engine results, we use the Google Search Engine API provided by Serper to retrieve 10 relevant document snippets based on the concatenation of the question q_i and the first 10 tokens of the model-generated text as keywords, forming the fact source S_{se}.[2] For LLM knowledge, we use one of the most advanced LLMs, gpt-4o-2024-05-13, to generate fact passages from its knowledge (Prompt A.1) to form the fact source S_{lk}.

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

[2] https://serper.dev/

3.3 UFO Evaluation Framework

Our evaluation pipeline includes three LLM-based modules: Fact Unit Extraction, Fact Source Verification, and Fact Consistency Discrimination. We apply LLaMA3-8B-Instruct in these three modules, and the prompts for these modules are provided in Appendix A.

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 $f_k = \langle q_k, e_k \rangle$. Benefiting from the potent language comprehension capabilities of LLMs, we introduce an LLM-based Fact Unit Extraction (FUE) method to extract the fact units. We follow the previous work (Min et al., 2023) and apply 4-shot demonstrations to the prompt in order to enhance the quality of extracted question-answer pairs, which is shown in Appendix A.2.

$$\{\langle q_1, e_1 \rangle, \cdots, \langle q_N, e_N \rangle\} = \text{FUE}(T_M(q_D)).$$

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 $\langle q_i, e_i \rangle$, 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 accurately extract answers from the fact source, we leverage the advanced context-understanding capabilities of LLMs. We instruct the LLM-based Fact Source Verification (FSV) module (Appendix A.3) to pinpoint the most relevant answers within the text, generating a "NOANS" text if no answer is found. This method directly prompts an LLM to retrieve answers from the passages from the fact sources, reducing inaccuracies during fact verification (Huang et al., 2023).

Answers are sequentially sought in each text passage of the fact source S^k until a suitable answer is found. If no text passage yields an answer, it indicates a mismatch with the fact source S^k, leading to a transition to the next fact source $S^{k+1} \in S$ for verification. Concretely, for a fact unit $\langle q_i, e_i \rangle$, we obtain the answer a_i using passage P_j^k from fact source S^k as follows:

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

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 (Appendix A.4), assigning a score of 0 or 1 to each fact unit $\langle q_i, e_i \rangle$. If no answer is extracted from all fact sources, the score for this fact unit is assigned a value of 0. Subsequently, we calculate the average score of all fact units as the factuality score of the model-generated text:

$$s_i = \text{FCD}(e_i, a_i) \in \{0, 1\}, \tag{2}$$

$$s = \frac{1}{N} \sum_{i=1}^{N} s_i. \tag{3}$$

3.4 Evaluation Criteria

Following existing studies (Sakai, 2006; Buckley and Voorhees, 2017), we measure the discriminative power (DP) of the evaluation metric.

Given the collection of source 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 source LLMs within the threshold. The PT indicates the percentage of cases where the pair of source LLMs cannot be distinguished within the given threshold. Thus, smaller values of MR and PT indicate a stronger discriminative power of the evaluation metric. To evaluate and compare the discriminative power with metrics clearly, we fix PT $= \alpha = 5\%$ and use a binary search method to find the threshold f. We then observe the value of DP $= 1 -$ MR as the success rate of distinguishing a pair of source LLMs, thereby assessing the discriminative power of the metric. The details of the pseudocode of DP measurement are provided in Appendix B.

3.5 Evaluation Scenarios

To assess the importance of all four fact sources across various tasks, we introduce six evaluation scenarios, each represented by an ordered list of fact sources S. (1) $S = \langle S_{\text{se}}, S_{\text{lk}} \rangle$. (2) $S = \langle S_{\text{lk}}, S_{\text{se}} \rangle$. (3) $S = \langle S_{\text{he}}, S_{\text{se}}, S_{\text{lk}} \rangle$. (4) $S = \langle S_{\text{rd}}, S_{\text{se}}, S_{\text{lk}} \rangle$. (5) $S = \langle S_{\text{he}}, S_{\text{rd}}, S_{\text{se}}, S_{\text{lk}} \rangle$. (6) $S = \langle S_{\text{rd}}, S_{\text{he}}, S_{\text{se}}, S_{\text{lk}} \rangle$. By comparing the discriminative power of a pair of evaluation scenarios, we can infer the importance of the fact sources:

(1) S_{se} and S_{lk}. In scenarios (1) and (2), we prioritize extracting passages from S_{se} or S_{lk}, respectively, to verify each fact unit. If the discriminative power of scenario (1) is higher, it indicates that S_{se} is more suitable for the factuality evaluation of this task, and vice versa.

(2) S_{he} and S_{rd}. In open-domain QA, web retrieval-based QA, and expert-validated QA tasks, human-written evidence S_{he} or reference documents S_{rd} might not always be provided. Consequently, the fact units in the model-generated text might not be fully verified by these fact sources. To determine the impact of these two fact sources, we introduce scenarios (3) to (6). Specifically, we fix the verification order of S_{se} and S_{lk} in order to leverage the external up-to-date facts and thoroughly verify facts. By comparing scenarios (1) and (3), we can infer the impact of S_{he} on the discriminative power of the evaluation pipeline. From the comparison of scenarios (1) and (4), we can infer the impact of S_{rd}. In the news fact generation and retrieval-augmented QA task, S_{he} and S_{rd} are both provided. To better explore the importance of S_{he} and S_{rd}, we compare the difference of discriminative power in scenarios (5) and (6) when all four fact sources are provided.

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 (tokenized by LLaMA2)						
Bing Chat	136.96	87.99	196.02	223.63	248.66	287.93
ChatGPT	118.03	106.75	127.53	384.94	369.65	173.16
llama2-7B	280.94	140.46	318.46	466.79	535.09	550.77
llama2-13B	325.79	184.31	351.66	509.80	572.72	525.34
llama2-70B	264.67	165.42	313.64	443.94	468.18	434.57
llama3-8B	236.89	97.09	315.77	489.17	568.98	511.35
llama3-70B	288.07	122.25	361.07	503.17	555.83	546.88
Qwen-7B	165.04	109.97	233.04	732.67	728.52	401.30
Qwen-14B	132.85	90.33	161.92	745.22	735.00	328.75
Avg. # of Extracted Facts Using LLaMA3-8B-Instruct						
Bing Chat	8.10	5.99	9.46	10.94	11.15	11.88
ChatGPT	7.37	7.38	7.39	13.62	12.36	8.78
llama2-7B	10.63	7.92	11.66	14.03	12.79	14.01
llama2-13B	11.97	9.66	12.53	13.23	13.02	15.66
llama2-70B	11.40	9.24	12.15	14.39	13.28	14.77
llama3-8B	10.30	6.59	11.47	14.24	13.80	14.03
llama3-70B	11.09	7.93	11.90	14.44	13.50	12.87
Qwen-7B	8.42	7.23	9.58	14.68	14.27	12.59
Qwen-14B	7.79	6.52	8.21	14.40	13.96	12.09
Avg. # of Extracted Facts Using gpt-3.5-turbo-0125						
Bing Chat	5.76	4.43	6.75	7.71	8.02	7.96
ChatGPT	5.91	5.96	5.90	8.61	8.33	6.71
llama2-7B	7.13	5.86	7.07	7.83	7.99	8.85
llama2-13B	8.04	7.07	7.70	8.08	8.31	8.87
llama2-70B	7.78	6.60	7.63	8.22	8.29	8.68
llama3-8B	6.91	4.96	7.63	8.21	8.59	8.23
llama3-70B	7.50	5.49	7.62	8.30	8.56	8.70
Qwen-7B	6.53	5.63	6.99	8.88	8.66	8.71
Qwen-14B	6.16	5.17	6.17	8.89	8.93	8.01

Table 2: Statistics of model-generated text from nine source LLMs 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 and Generation Tasks

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 the source LLMs to generate facts based on the question or write a news article with the first 30 tokens of the reference documents (Appendix A.5). Considering the available human-written evidence and reference documents in the datasets, we categorize the tasks presented in Table 1. We construct a golden answer G containing more facts for each task to compare with reference-based metrics. (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 $G = [a; S_{\mathrm{rd}}]$. **(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, the news summary is considered as the golden answer G and human-written evidence S_{he}, and the news stories are considered as reference documents S_{rd}. **(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 Source LLMs and Baselines

Source LLMs We evaluate nine 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.[3] 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-0125) for text generation.[4] (3) LLaMA (Touvron et al., 2023): We select three LLaMA2-series fine-tuned models (LLaMA2-{7,13,70}B-chat), and select two LLaMA3-series models (LLaMA3-{8,70}B-Instruct) for text generation. (4) To evaluate the difference in factuality between LLMs of similar parameter scales, we also evaluate Qwen1.5 (Bai et al., 2023) with two parameter scales ({7,14}B) and compare the performance with the other LLMs. The statistical data of the text generated by these source LLMs is demonstrated 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 human-written evidence and reference documents 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 (Buckley and Voorhees, 2017; Sakai, 2006) of the proposed evaluation pipeline UFO in each scenario. The experimental results are shown in the **first** and **second** part of Table 3. We have the following findings:

(1) Among all baselines, our proposed evaluation method achieves the best performance of discriminative power. For reference-based methods, the performance particularly relies on the quality of the golden answer G, especially the entities and fact-related keywords within the golden answer. The baselines using the question-generation and question-answering framework (QAGS and Q2) show relatively weaker discriminative power. This demonstrates that the proposed LLM evaluator outperforms PLM-based methods in extracting high-quality QA pairs and understanding the context. Reference-free baseline methods verify fact units with a fixed fact source, which means some fact units cannot be verified through the fact source. Our proposed method, utilizes a series of fact sources to thoroughly verify fact units, thereby enhancing the performance of discriminative power.

(2) In the open-domain QA and web retrieval-based QA task, we observe that the performance of scenario (1) outperforms (2), indicating that S_{se} is more effective on the discriminative power than S_{lk}. Meanwhile, in the expert-validated QA task, prioritizing S_{se} in the verification of fact sources degrades the discriminative power. In the TruthfulQA dataset, we notice that some facts are rather hard to verify through search engine results and

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

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

S_{he}	✗	✗	✓	✓	✓	✓
S_{rd}	✗	✓	✗	✓	✓	✓
Dataset	NQ	HotpotQA	TruthfulQA	CNN/DM	Multi-News	MS MARCO
1. Baseline methods						
BLEU-1	0.641	0.813	0.807	0.813	0.657	0.625
ROUGE-L	0.782	0.824	0.803	0.788	0.628	0.644
BERTScore-f1	0.769	0.852	0.744	0.759	0.623	0.689
BARTScore	0.837	0.866	0.721	0.766	0.616	0.677
QAGS	0.660	0.634	0.731	0.655	0.675	0.817
Q2	0.653	0.751	0.734	0.798	0.602	0.573
FacTool	0.910	0.917	0.817	0.904	0.816	0.832
FactScore	0.929	0.925	0.919	0.892	0.791	0.860
2. UFO (LLaMA3-8B-Instruct)						
① $\langle S_{\text{se}}, S_{\text{lk}} \rangle$	**0.945**	0.942	0.901	0.919	0.853	0.892
② $\langle S_{\text{lk}}, S_{\text{se}} \rangle$	0.932	0.933	0.924	0.899	0.844	0.885
③ $\langle S_{\text{he}}, S_{\text{se}}, S_{\text{lk}} \rangle$	-	-	**0.933**	0.907	0.839	0.909
④ $\langle S_{\text{rd}}, S_{\text{se}}, S_{\text{lk}} \rangle$	-	**0.952**	-	**0.930**	**0.864**	0.911
⑤ $\langle S_{\text{he}}, S_{\text{rd}}, S_{\text{se}}, S_{\text{lk}} \rangle$	-	-	-	0.921	0.854	0.917
⑥ $\langle S_{\text{rd}}, S_{\text{he}}, S_{\text{se}}, S_{\text{lk}} \rangle$	-	-	-	0.925	0.859	**0.920**
③ - ① ΔS_{he}	-	-	0.032	-0.012	-0.014	0.017
④ - ① ΔS_{rd}	-	0.010	-	0.011	0.011	0.019
① - ② ΔS_{se}	0.013	0.009	-0.023	0.020	0.009	0.007
② - ① ΔS_{lk}	-0.013	-0.009	0.023	-0.020	-0.009	-0.007
⑤ - ⑥ ΔS_{he}	-	-	-	-0.004	-0.005	-0.003
3. Incorporation of model-retrieved reference documents S_{rd} (LLaMA3-8B-Instruct)						
④ $\langle S_{\text{rd}}, S_{\text{se}}, S_{\text{lk}} \rangle$	**0.953**	**0.959**	0.928	**0.939**	**0.876**	0.904
⑤ $\langle S_{\text{he}}, S_{\text{rd}}, S_{\text{se}}, S_{\text{lk}} \rangle$	-	-	**0.941**	0.924	0.866	0.905
⑥ $\langle S_{\text{rd}}, S_{\text{he}}, S_{\text{se}}, S_{\text{lk}} \rangle$	-	-	0.936	0.935	0.872	**0.917**
4. UFO (gpt-3.5-turbo-0125)						
① $\langle S_{\text{se}}, S_{\text{lk}} \rangle$	**0.940**	0.941	0.889	0.912	0.846	0.880
② $\langle S_{\text{lk}}, S_{\text{se}} \rangle$	0.935	0.929	0.922	0.885	0.836	0.877
③ $\langle S_{\text{he}}, S_{\text{se}}, S_{\text{lk}} \rangle$	-	-	**0.931**	0.894	0.835	0.895
④ $\langle S_{\text{rd}}, S_{\text{se}}, S_{\text{lk}} \rangle$	-	**0.945**	-	**0.927**	**0.862**	0.899
⑤ $\langle S_{\text{he}}, S_{\text{rd}}, S_{\text{se}}, S_{\text{lk}} \rangle$	-	-	-	0.916	0.849	0.903
⑥ $\langle S_{\text{rd}}, S_{\text{he}}, S_{\text{se}}, S_{\text{lk}} \rangle$	-	-	-	0.920	0.859	**0.906**

Table 3: Discriminative power of evaluation metrics on six datasets under the condition $\alpha = 5\%$. The **first** part of the table shows the discriminative power of baseline methods. The **second** part of the table demonstrates all six studied scenarios in our proposed pipeline with the evaluator model LLaMA3-8B-Instruct. In the **third** part of the table, we incorporate reference documents retrieved by Bing Chat as part of the fact source S_{rd}. In the **fourth** part of the table, we test our pipeline with the evaluation model gpt-3.5-turbo-0125.

need detailed explanations from experts, such as "the signing date of the United States Declaration of Independence". In this case, the search results are possibly inaccurate or contradictory.

(3) For the news fact generation task, we find S_{rd} significantly improves discriminative power in scenario (4). Comparing between scenarios (1) and (2), we infer that S_{se} is more important than S_{lk}. Indeed, LLMs are prone to hallucinations regarding specific factual details when generating news-related facts based on their internal knowledge, which negatively impacts the verification of the given news text. In contrast, search engine results and reference documents provide many relevant and detailed news facts, thus improving the discriminative power. We also observe that human-written evidence often neglects the details of news facts and includes subjective comments, resulting in a negative impact on discriminative power in the verification of news fact generation task.

(4) We study the retrieval-augmented QA task in the MS MARCO dataset. The experimental results demonstrate that the scenario (6) outperforms the other scenarios. This indicates that both S_{he} and S_{rd} are crucial to the task, and the comparison with scenario (5) shows that S_{rd} presents more impact on the discriminative power of evaluation. From our observation, the reference documents clicked by users usually contain more comprehensive and accurate facts than human-written evidence.

Dataset		NQ		HotpotQA		TruthfulQA		CNN/DM		Multi-News		MS MARCO	
Models	Params	UFO	FT	UFO	FT	UFO	FT	UFO	FT	UFO	FT	UFO	FT
Bing Chat	N/A	0.738	0.634	0.615	0.707	0.640	0.699	0.627	0.796	0.725	0.800	0.784	0.795
ChatGPT	N/A	0.750	0.711	0.621	0.720	0.653	0.713	0.648	0.815	0.731	0.796	0.780	0.812
LLaMA2	7B	0.601	0.538	0.477	0.492	0.528	0.568	0.583	0.752	0.603	0.671	0.694	0.746
LLaMA2	13B	0.664	0.584	0.532	0.527	0.566	0.620	0.615	0.749	0.648	0.744	0.721	0.750
LLaMA2	70B	0.701	0.613	0.596	0.682	0.611	0.674	0.621	0.763	0.699	0.785	0.751	0.769
Qwen1.5	7B	0.682	0.613	0.555	0.531	0.549	0.590	0.617	0.779	0.654	0.749	0.730	0.766
Qwen1.5	14B	0.697	0.630	0.589	0.663	0.597	0.649	0.619	0.761	0.675	0.770	0.743	0.752
LLaMA3	8B	0.753	0.710	0.614	0.723	0.662	0.723	0.659	0.823	0.730	0.805	0.786	0.808
LLaMA3	70B	**0.808**	**0.742**	**0.652**	**0.760**	**0.680**	**0.755**	**0.691**	**0.846**	**0.769**	**0.818**	**0.814**	**0.837**

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. The highest factuality score is **bold**, and the second is underlined.

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 source LLM (*i.e.*, Bing Chat in our experiments). The discriminative power of the scenarios are shown in the **third** part of Table 3. We have the following findings:

(1) In NQ, HotpotQA, TruthfulQA, CNN/DM, and Multi-News datasets, the incorporation of model-retrieved documents raises the discriminative power performance. In open-domain QA, web retrieval-based QA, and expert-validated QA tasks, the retrieved documents contain entities and fact knowledge related to the question. In news fact generation tasks, the retriever accesses more comprehensive facts from reliable sources, thereby enhancing the discriminative power of the evaluation.

(2) Incorporation of retrieved reference documents slightly degrades the discriminative power in the retrieval-augmented QA task. Users click sufficient reference documents and provide answers, thus the model-retrieved documents may bring more noise, which contains irrelevant and redundant content to degrade the discriminative power.

5.3 Bias from LLM Evaluators

In Section 3.3, we propose three LLM-based modules and mitigate biases from LLMs. To assess the influence of selecting different LLM evaluators, we also test the proposed pipeline with gpt-3.5-turbo-0125 applied in the modules. The statistics of extracted facts are shown in Table 2, and the discriminative power performance is shown in the **fourth** part of Table 3. We observe that LLaMA3-8B extracts more facts, and applying LLaMA3-8B as the evaluator slightly enhances the discriminative power in most datasets. This indicates that LLaMA3-8B is more capable of capturing and extracting fine-grained facts, while the conclusion from the evaluation scenarios remains unchanged, indicating that the design of our proposed modules does not significantly introduce biases of LLMs.

5.4 Factuality Scores of LLMs

In addition to evaluating discriminative power, we also obtain the factuality scores of nine source LLMs on six 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 LLaMA3-70B achieves the best factuality score among all six datasets. Also, the factuality score of Bing Chat in "precise" mode is slightly lower than that of ChatGPT, and is close to the score of LLaMA3-8B. 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 Qwen) can enhance factual accuracy in all six datasets.

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 six evaluation scenarios show that for most QA tasks, human-written evidence and reference documents are crucial, but in the news fact generation tasks, introducing human-written evidence leads to a decline in performance. Compared to the LLM knowledge, search engine results are more important in most tasks, but they are less effective in the expert-validated QA task.

Limitations

In this work, we propose a unified and flexible factuality evaluation framework to analyze different fact sources. However, there are still several limitations:

(1) We prompt one of the most advanced LLMs (gpt-4o-2024-05-13) to generate passages as the fact source S_{lk}. However, over time, the LLM knowledge may become outdated. Meanwhile, the content of facts in the search engine results might be contradictory when the search query is unclear or ambiguous. In future work, we will explore the recognition and filtering of outdated content in LLM knowledge S_{lk}, and irrelevant or incorrect content in search engine results S_{se}.

(2) The discriminative power of the evaluation is obtained by constructions of source LLM pairs, thus it is influenced by the number of source LLMs. If the number of source LLMs is small, the calculation of discriminative power may be inaccurate. In our future work, we will evaluate more source LLMs to calculate the discriminative power of each scenario more precisely, thereby better discerning the importance of fact sources.

(3) We evaluate the text generated by Bing Chat, which we manually collected in December 2023 and released in our demonstrated anonymous link. However, due to the lack of a released checkpoint for the Bing Chat model, it may be difficult to reproduce the generated text that we collected at other times.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

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. *Preprint*, arXiv:2302.04023.

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. *Preprint*, arXiv:2005.14165.

Chris Buckley and Ellen M. Voorhees. 2017. Evaluating evaluation measure stability. *SIGIR Forum*, 51(2):235–242.

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*.

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. *Preprint*, arXiv:2301.00234.

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. *Preprint*, arXiv:2311.05232.

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. *Preprint*, arXiv:2203.05227.

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. *Preprint*, arXiv:2306.07906.

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. *Preprint*, arXiv:2203.11147.

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. *Preprint*, arXiv:2112.09332.

OpenAI. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

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.

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. *Preprint*, arXiv:2307.09288.

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. *Preprint*, arXiv:2303.18223.

A Prompt

A.1 LLM knowledge

We prompt gpt-4o-2024-05-13 to generate knowledge based on the model-generated text T. The details are shown in Table 5.

A.2 Fact Unit Extraction

Table 6 demonstrates the prompt we use in the fact unit extraction module. Following previous work (Min et al., 2023), we apply 4-shot demonstrations to enhance response quality.

A.3 Fact Source Verification

Given a passage from the fact source, we prompt the LLM to extract the answer from the passage. The prompt is shown in Table 7.

A.4 Fact Consistency Discrimination

We prompt the LLM to judge the consistency of two answers with a direct answer (yes or no). The details are demonstrated in Table 8.

A.5 Generation

For the open-domain QA, web retrieval-based QA, expert-validated QA, and retrieval-augmented QA tasks, we directly prompt the source LLMs to generate responses. The prompt is demonstrated in Table 9. For the news fact generation task, due to the lack of user query, we prompt the LLM to complete the article with the first 30 tokens in the first reference document. The prompt is demonstrated in Table 10.

B Pseudocode for DP Measurement

In Algorithm 1, we describe how we calculate the discriminative power of a given metric.

Given the following document: {model-generated text} The factuality of the document has not been evaluated. Your task is only to use your knowledge to serve as a fact source, and respond with a relevant, correct, and precise fact passage centered on the topic of the given document.

Table 5: Prompt for generating LLM knowledge based on the model-generated text.

Algorithm 1 Discriminative Power Measurement

```
$B \leftarrow 1000$
$\alpha \leftarrow 0.05$
$\epsilon \leftarrow 0.001$
$low \leftarrow 0$
$high \leftarrow 1$
$max\_iterations \leftarrow 20$
for $k = 1$ to $max\_iterations$ do
    $f \leftarrow (low + high)/2$
    for each $(M_i, M_j) \in M$ do
        $EQ(i, j) \leftarrow 0$
        $GT(i, j) \leftarrow 0$
        $GT(j, i) \leftarrow 0$
        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
    $MR_f \leftarrow \frac{\sum_{M_i,M_j} \min(GT(i,j),GT(j,i))}{B\sum_{M_i,M_j}}$
    $DP_f \leftarrow 1 - MR_f$
    $PT_f \leftarrow \frac{\sum_{M_i,M_j} EQ(i,j)}{B\sum_{M_i,M_j}}$
    if $PT_f < \alpha - \epsilon$ then
        $low \leftarrow f$
    else if $PT_f > \alpha + \epsilon$ then
        $high \leftarrow f$
    else
        break
    end if
end for
```

```
Please breakdown the following passage into independent atomic questions and answers.
<Passage>
He made his acting debut in the film The Moon is the Sun’s Dream (1992), and continued to appear in small and supporting roles throughout the 1990s.
<Atomic Q&A>
Question: What did he make his debut in?
Answer: He made his acting debut in the film.
Question: What is the name of the film in which he made his acting debut?
Answer: He made his acting debut in The Moon is the Sun’s Dream.
Question: When was The Moon is the Sun’s Dream released?
Answer: The Moon is the Sun’s Dream was released in 1992.
Question: What type of roles did he appear in after his acting debut?
Answer: After his acting debut, he appeared in small and supporting roles.
Question: When did he appear in small and supporting roles after his acting debut?
Answer: After his acting debut, he appeared in small and supporting roles throughout the 1990s.

Please breakdown the following passage into independent atomic questions and answers.
<Passage>
He is also a successful producer and engineer, having worked with a wide variety of artists, including Willie Nelson, Tim McGraw, and Taylor Swift.
<Atomic Q&A>
Question: What is his profession?
Answer: He is a producer and an engineer.
Question: Has he worked with a variety of artists?
Answer: Yes, he has worked with a wide variety of artists.
Question: Who is Willie Nelson?
Answer: Willie Nelson is an artist.
Question: Who is Tim McGraw?
Answer: Tim McGraw is an artist.
Question: Who is Taylor Swift?
Answer: Taylor Swift is an artist.

Please breakdown the following passage into independent atomic questions and answers.
<Passage>
In 1963, Collins became one of the third group of astronauts selected by NASA and he served as the back-up Command Module Pilot for the Gemini 7 mission.
<Atomic Q&A>
Question: What role did Collins become in 1963?
Answer: Collins became an astronaut.
Question: Which group of astronauts did Collins join?
Answer: Collins became one of the third group of astronauts.
Question: Who selected the third group of astronauts that Collins became a part of?
Answer: Collins became one of the third group of astronauts selected by NASA.
Question: When was Collins selected by NASA to be an astronaut?
Answer: Collins became one of the third group of astronauts selected by NASA in 1963.
Question: What was Collins’s role in the Gemini 7 mission?
Answer: He served as the Command Module Pilot for the Gemini 7 mission.
Question: What specific role did Collins serve in for the Gemini 7 mission?
Answer: He served as the back-up Command Module Pilot.

Please breakdown the following passage into independent atomic questions and answers.
<Passage>
In addition to his acting roles, Bateman has written and directed two short films and is currently in development on his feature debut.
<Atomic Q&A>
Question: Does Bateman have acting roles?
Answer: Yes, Bateman has acting roles.
Question: How many short films has Bateman written?
Answer: Bateman has written two short films.
Question: How many short films has Bateman directed?
Answer: Bateman has directed two short films.
Question: What has Bateman done in terms of writing and directing short films?
Answer: Bateman has written and directed two short films.
Question: What is Bateman currently working on?
Answer: Bateman is currently in development on his feature debut.

Please breakdown the following passage into independent atomic questions and answers.
<Passage>
{passage}
<Atomic Q&A>
```

Table 6: Prompt for fact unit extraction with 4-shot demonstrations.

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". evidence: {evidence} question: {question} your answer:

Table 7: Prompt for fact source verification.

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

Table 8: Prompt for fact consistency discrimination.

«System Prompt» Your task is to answer the question and introduce sufficient fact details based on the knowledge you possess. Your response must be in English. *«User»* {Question}

Table 9: Prompt for the open-domain QA, web retrieval-based QA, expert-validated QA, and retrieval-augmented QA tasks.

«System Prompt» You are an English news writer. *«User»* Please write an article starting exactly with: {Passage} Article:

Table 10: Prompt for the news fact generation task. We keep the first 30 tokens in the first reference document as the passage for generation.