A Survey of Large Language Models Attribution

Anonymous ACL submission

Abstract

Open-domain generative systems have gained significant attention in the field of conversational AI (e.g., generative search engines). In this paper, we present a comprehensive review of the attribution mechanisms employed by these systems, particularly with large language models. While attribution or citation improves factuality and verifiability, issues like ambiguous knowledge reservoirs, inherent biases, and the drawbacks of excessive attribution can hinder the effectiveness of these systems. The purpose of this survey is to provide valuable implications for researchers, helping in the refinement of attribution methodologies to improve the reliability and veracity of responses generated by open-domain generative systems. We believe that this field is still in its early stages; therefore, we maintain a repository to keep track of ongoing studies at AnonymousURL.

1 Introduction

Since the emergence of open-domain generative systems driven by Large Language Models (LLMs) (Anil et al., 2023; OpenAI, 2023), addressing the coherent generation of potentially inaccurate or fabricated content has been a persistent challenge in Natural Language Processing (NLP) (Rawte et al., 2023; Ye et al., 2023a; Zhang et al., 2023b). These problems are commonly referred to within the community as hallucination problems in which generated content presents distorted or invented facts that lack credible sources (Peskoff and Stewart, 2023). This becomes particularly obvious in scenarios involving information-seeking and knowledge-based question-answering, where users rely on these systems for expert knowledge (Malaviya et al., 2023).

The essence of the hallucination problem may stem from the fact that pre-trained models are sourced from vast, unfiltered real-world texts (Penedo et al., 2023). These human-generated texts inherently contain inconsistencies and falsehoods. The objective of pre-training is merely to predict the next word, without explicitly modeling the veracity of the generated content. Even after utilizing reinforcement learning from human feedback (Ouyang et al., 2022), models can still exhibit external hallucinations (Bai et al., 2022). To address the issue of external hallucinations, researchers have begun to employ measures like external references to enhance the authenticity and reliability of chatbots (Thoppilan et al., 2022; Menick et al., 2022; Nakano et al., 2021). The distinction between explicit attribution and learning from human feedback lies not only in the need for human verification and compliance but also in recognizing that generated content might become outdated or invalid over time. As shown in Figure 1, attribution can leverage real-time information to ensure relevance and accuracy. However, the fundamental challenge of attribution revolves around two essential requirements (Liu et al., 2023):

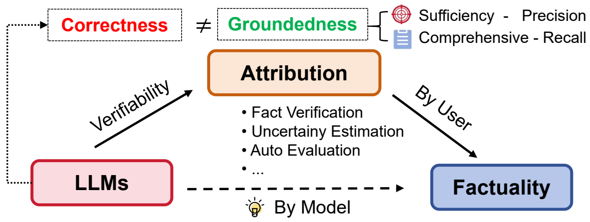

Figure 1: By providing attribution, both developers and users can view the possible source of an answer and evaluate factuality and reliability to form their own assessment. Attribution as a more realistic way to reduce hallucinations bypasses the task of directly determining the "truthfulness" of statements, a feat difficult to achieve except for the most basic queries.

1. **Comprehensive Attribution or Citation (High Recall).** All claims and statements (except debatable or subjective text, e.g., abstained text) made by the model-generated

content should be fully supported by appropriate references.

2. **Sufficiency Attribution or Citation (High Precision).** Every reference should directly support its associated claim or statement.

With these requirements in mind, we can break down the main ways models handle attribution into three types (see examples in Figure 2):

1. **Direct Model-driven Attribution.** The LLM itself provides the attribution for its answer. However, this type often poses a challenge as not only might the answers be hallucinated, but the attributions themselves can also be (Agrawal et al., 2023). Although ChatGPT provides correct or partially correct answers about 50.6% of the time, the suggested references were only present 14% of the time (Zuccon et al., 2023).

2. **Post-retrieval Answering.** This approach is rooted in the idea of explicitly retrieving information and then letting the model answer based on these retrieved data. But retrieval does not inherently equate to attribution (Gao et al., 2023b). Issues arise when the boundaries between internal knowledge of the model and externally retrieved information become blurred, leading to potential knowledge conflicts (Xie et al., 2023). Retrieval can also be used as a specialized tool allowing the model to trigger it independently, similar to the `Browse with Bing` in ChatGPT. [1]

3. **Post-generation Attribution.** The system first provides an answer and then conducts a search using both the question and the answer for attribution. The answer is then modified if necessary and appropriately attributed. Modern search engines like Bing Chat [2] have already incorporated such attribution. However, studies have shown that only 51.5% of the content generated from four generative search engines was entirely supported by their cited references (Liu et al., 2023). This form of attribution is particularly lacking in high-risk professional fields such as medicine and law, with research revealing a significant number of incomplete attributions (35% and 31%, respectively); furthermore, many attributions were derived from unreliable sources and 51% of them were evaluated as unreliable by experts (Malaviya et al., 2023).

[1]https://openai.com/blog/chatgpt-plugins
[2]https://www.bing.com/new

Moving beyond general discussions on text hallucinations (Zhang et al., 2023b; Ye et al., 2023a; Rawte et al., 2023), our study delves deeper into the attribution of LLMs. As shown in Figure 3, we explore its origins, the technology underpinning it, and the criteria for its assessment. Additionally, we touch upon challenges such as biases and the potential for excessive citations. We believe that by focusing on these attribution issues, we can make the models more trustworthy and easier to understand. Our goal with this study is to shed light on attribution in a way that is clearer and encourages deeper thought on the topic.

2 Task Definition

Attribution refers to the capacity of an entity, such as a language model, to generate and provide evidence, often in the form of references or citations, that substantiates the claims or statements it produces. This evidence is derived from identifiable sources, ensuring that the claims can be logically inferred from a foundational corpus, making them comprehensible and verifiable by a general audience. Attribution itself is related to search tasks (Page et al., 1999; Tay et al., 2022) where only several web pages are returned. However, the primary purposes of attribution include enabling users to validate the claims made by the model, promoting the generation of text that closely aligns with the cited sources to enhance accuracy and reduce misinformation or hallucination, and establishing a structured framework for evaluating the completeness and relevance of the supporting evidence in relation to the presented claims.

The accuracy of attribution centers on *whether the produced statement is entirely backed by the referenced source*. For example, Rashkin et al. (2021) propose the *Attributed to Identified Sources* (AIS) evaluation framework to assess whether a particular statement is supported by provided evidence. Bohnet et al. (2022) further propose attributed question answering, where the model takes a question and produces a paired response of an answer string and its supporting evidence from a specific corpus, such as paragraphs.

Formally, consider a query q (or an instruction, a prompt) and a corpus of text passages $\mathcal{D}$.

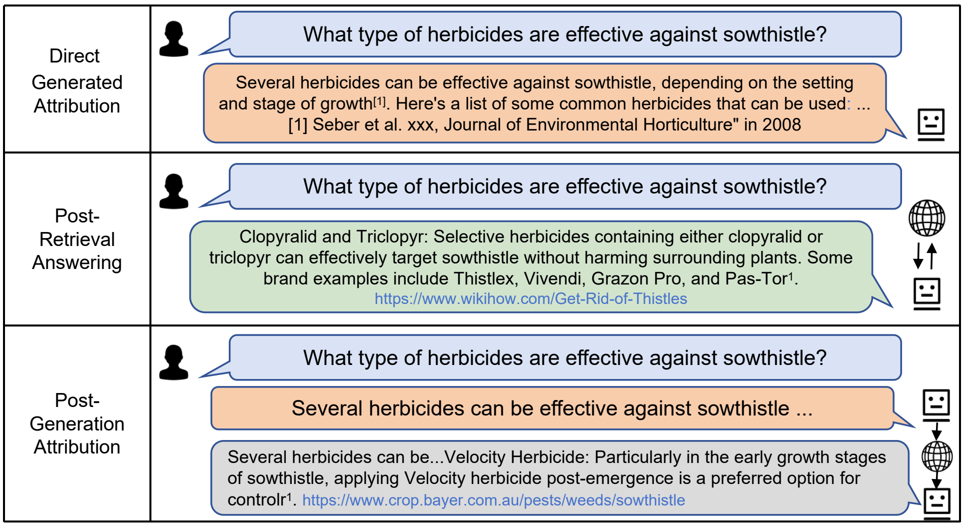

Figure 2: Three ways to attribute model-generated content. In direct model-driven attribution, the reference document is derived from model itself and is used to cite generated answer. In post-retrieval answering, model generates answer with citations based on retrieved documents. In post-generation attribution, an answer is first generated then then the answer is modified again to add references for attribution.

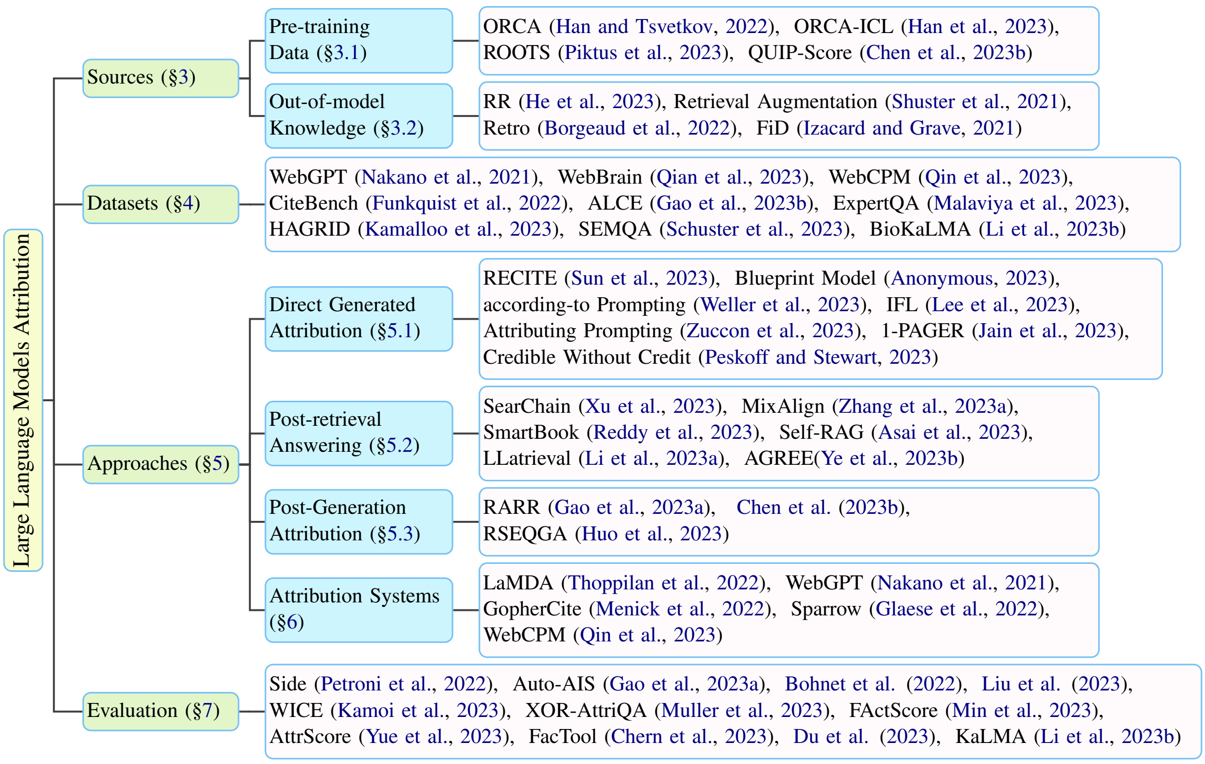

Figure 3: Taxonomy of large language models attribution.

The objective of the system is to produce an output $\mathcal{S}$, where $\mathcal{S}$ is a set of n distinct statements: $s_1, s_2, \ldots, s_n$. Each statement s_i is associated with a set of citations $\mathcal{C}_i$. This set $\mathcal{C}_i$ is defined as $\mathcal{C}_i = \{c_{i,1}, c_{i,2}, \ldots\}$, where each $c_{i,j}$ is a passage from the corpus $\mathcal{D}$. For practical applications, the output from LLMs can be segmented into individual statements using sentence boundaries. This approach is utilized because a single sentence typically encapsulates a coherent statement while maintaining brevity, facilitating easy verification. In terms of representation, citations may be enclosed within square brackets, for instance, `[1][2]`. It should be noted, however, that these citations can also be applied at the phrase level, rather than exclusively at the sentence level. It is important to highlight that the task configurations discussed in this paper are distinct from the generation of citation texts found in scholarly articles or wikipedia, where the citing and cited documents are usually used as inputs (Fetahu et al., 2016; Xing et al., 2020; Wu et al., 2021; Gu and Hahnloser, 2022).

3 Sources of Attribution

3.1 Pre-training Data

LLMs are typically trained on extensive corpora collected from various sources, predominantly the web. This vast amount of pre-training data forms the bedrock on which these models develop their understanding and capabilities. However, due to the scale of the data involved, manual inspection is often unfeasible, leading to potential inaccuracies, biases, and other undesirable artifacts in the data (Piktus et al., 2023). Despite these challenges, LLMs tend to perform well on a wide array of downstream tasks, even with little to no task-specific tuning. This performance hints at the ability of models to either memorize or reason through patterns present in the data. However, the specific patterns or the extent to which they are memorized or reasoned through, especially in different downstream tasks, remain somewhat elusive.

The concept of attribution in this context refers to tracing back the behavior of the model on a particular task to specific portions of the pre-training data (Han and Tsvetkov, 2022; Weller et al., 2023). By identifying a subset of pre-training data that significantly influences the model behavior on a downstream task, researchers aim to provide a clearer understanding of how the pre-training data impacts the model's performance (Han et al., 2023). This kind of attribution is essential for interpreting the model, providing insights into whether the model is capturing task-relevant patterns or merely memorizing data. Furthermore, it aids in enhancing the trustworthiness of the model by offering a clearer picture of how the model operates and what sources of data significantly contribute to its performance. Through such attribution methodologies, researchers aim to bridge the understanding gap, offering a pathway towards better interpretability, trustworthiness, and eventually, the improvement of LLMs in handling various NLP tasks.

3.2 Out-of-model Knowledge

This source reveals methods to leverage out-of-model knowledge (e.g., web, knowledge graph) for attribution to enhance the capabilities of models (Shuster et al., 2021; Li et al., 2023b). Primary among these methods is the retrieval-augmented generation technique (Lewis et al., 2020) which uses an encoder-decoder mechanism to encode questions and decode answers, augmented with documents or passages from extensive unstructured datasets. Furthermore, retrieval-enhanced language models are highlighted, which improve performance by fetching k-most similar training contexts or generating search queries to obtain relevant documents from external sources (Borgeaud et al., 2022). These methodologies, along with a mentioned post-processing method to utilize retrieved knowledge without additional training or fine-tuning, represent critical pathways for attributing LLM responses or generated text to external knowledge, aiming to make the outputs of LLMs verifiable external knowledge sources (Izacard and Grave, 2021; He et al., 2023).

4 Datasets for Attribution

As an information-seeking task, datasets for attribution are often built in the form of Question Answering (QA) or summarization (see Table 1). Several benchmarks are proposed based on existing QA datasets by proposing methods to evaluate the performance of attribution, as the golden citation annotation is not a necessity. Nakano et al. (2021) built a long-form QA dataset with web search results. After that Qin et al. (2023) built a similar Chinese dataset for the same purpose. However, these datasets are not directly built for verifying citations, but for factual accuracy. More recently, several works (Qian et al., 2023; Gao et al., 2023b; Kamalloo et al., 2023; Malaviya et al., 2023; Li et al., 2023b) focus on measuring and improving the accuracy of citations in generated text based on a given set of quotes, varying on question domain and citation granularity.

Question Domain. Most recent attribution datasets are designed for open-domain. However, ExpertQA (Malaviya et al., 2023) choose 32 domain-specific scenarios, some of which are high-stakes fields, and bring domain experts in the loop. BioKaLMA (Li et al., 2023b) focuses on biography domain for its practical application and convenient evaluation.

Attribution Granularity. There are two kinds of citation granularity in recent works: entity and sentence. The entity level attribution is more fine-grained, sentence level attribution requires citation for every completed sentence.

Among them, SEMQA (Schuster et al., 2023), ExpertQA (Malaviya et al., 2023) and BioKaLMA (Li et al., 2023b) make attribution at entity level, whereas other methods make attribution at sentence level.

5 Approaches to Attribution

5.1 Direct Generated Attribution

Attribution from parametric knowledge can help reduce hallucination and improve the truthfulness of generated text. By asking models to do self-detection and self-attribution, some works indicate that the generated texts are more grounded on facts and additionally improve performance on downstream tasks (Sun et al., 2023).

Recently, researchers found that large language models can not provide knowledge sources or evidence clearly when answering domain-specific knowledge-based questions (Peskoff and Stewart, 2023; Zuccon et al., 2023; Gravel et al., 2023). In most cases, models can only provide a knowledge source that is loosely related to the keywords in questions or irrelevant to current topics. Even if the model answered the question correctly, the evidence it provided is still likely to have mistakes. Weller et al. (2023) tries to ground model's generated text to its pre-training data by proposing according-to prompting, who finds the method can affect model's groundedness and therefore affect performance on information-seeking tasks. Anonymous (2023) introduces an intermediate planning module, asking the model to generate a series of questions as blueprints to the current question. The model first proposes a blueprint and then combines the texts which are generated based on the blueprint questions as the final answer. The blueprint models allow for different forms of attribution during each question answering step, which can be expected to be more explainable.

5.2 Post-retrieval Answering

Numerous studies have delved into the post-retrieval answering strategy for attribution (Chen et al., 2017; Lee et al., 2019; Khattab and Zaharia, 2020). Reddy et al. (2023) introduces the SmartBook framework, which aims to generate structured situation reports incorporating factual evidence through rich links. The framework autonomously identifies crucial questions for situation analysis and extracts pertinent information to compose the report. Each question is addressed with concise summaries containing tactical details of pertinent claims, supported by reliable and trustworthy factual evidence. To tackle the issue of misalignment between user queries and stored knowledge, where LLMs struggle to correlate questions with the appropriate grounding, MixAlign (Zhang et al., 2023a) presents a framework that combines automatic question-knowledge alignment with user clarifications. This approach effectively mitigates language model hallucination. To assess the adequacy of document support for an answer, LLatrieval (Li et al., 2023a) updates the retrieval results until it confirms that the retrieved documents can sufficiently support the answer to the question. This iterative verification process significantly enhances the accuracy of the attribution by ensuring that the generated response is supported by verifiable evidence. Similarly, Self-RAG (Asai et al., 2023) trains an arbitrary language model to generate reflection-specific tokens after knowledge retrieval, thereby augmenting the attribution of retrieved passages. Furthermore, Search-in-the-chain (SearChain) (Xu et al., 2023) introduces a method to address the challenges posed by incorrect knowledge retrieved by information retrieval systems, which can mislead LLMs or disrupt their reasoning chains. It verifies and corrects answers within the global reasoning chain, known as Chain-of-Query (CoQ), while also identifying missing knowledge in CoQ. These operations significantly improve the attribution accuracy of LLMs in complex knowledge-intensive tasks, improving their reasoning ability and knowledge utilization.

5.3 Post-Generation Attribution

In order to facilitate accurate attribution without compromising the robust benefits offered by recent generation models, some research aims at attribution after generation, which employ search engines or document retrieval systems to search the evidence base on the input questions and generated answers. This approach allows researchers to assess or improve the factuality of answers without needing to access the model's parameters directly. The post-generation attribution workflow is illustrated in Figure 4. RARR (Gao et al., 2023a) autonomously identifies the attribution of the output of any text generation model. It progressively verifies the factual consistency between the output and its source, and performs post-editing to rectify unsupported content, whilst striving to retain the original output to the greatest extent feasible. In the

Dataset	Domain	Source	Structure	Granularity	Response Source	#Questions
WebGPT (Nakano et al., 2021)	Open-domain	Web Pages	Unstructured	Sentence	GPT-3	19,578
WebBrain (Qian et al., 2023)	Open-domain	Wikipedia	Unstructured	Sentence	GPT-3	2.74M
WebCPM (Qin et al., 2023)	Open-domain	Web Pages	Unstructured	Sentence	Human	5,500
HAGRID (Kamalloo et al., 2023)	Open-domain	Wikipedia	Unstructured	Sentence	GPT-3.5, Human	1,922
ALCE (Gao et al., 2023b)	Open-domain	Wikipedia+Sphere	Unstructured	Sentence	Human	2,984
SEMQA (Schuster et al., 2023)	Open-domain	Wikipedia	Unstructured	Entity	Human	1,376
BioKaLMA (Li et al., 2023b)	Biography	Wikipedia	Structured	Entity	GPT-3.5,GPT4,LLaMA	1,085
ExpertQA (Malaviya et al., 2023)	Specific domains	Wikipedia	Unstructured	Entity	GPT-4, Human	2,507

Table 1: Comparsion between different datasets for attribution.

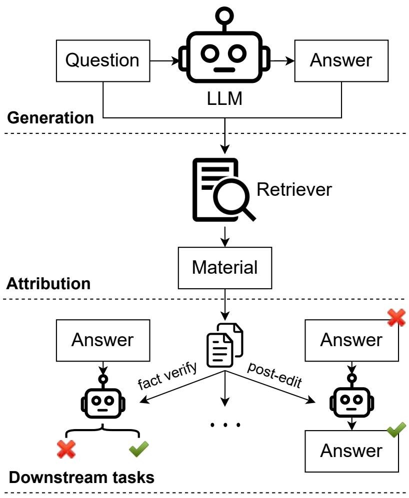

Figure 4: Workflow of post-generation attribution. Retrieval is performed after an answer being generated. The retrieved documents are used to perform citation and attribution, subsequently used to do fact verification and post-editing.

work of Huo et al. (2023), materials are retrieved from the corpus based on coarse-grained sentences or fine-grained factual statements. These retrieved materials are then utilized to prompt the LLM to verify the consistency between the generated responses and the retrieved material, and to make necessary edits to reduce the hallucinations. Chen et al. (2023b) introduces a fully automatic pipeline designed to verify complex political claims, which is achieved by retrieving evidence from the web. It breaks down each claim into subquestions and retrieves specific evidence for each, creating focused summaries and using them for claim verification. During training, the system evaluates its individual components based on the comprehensiveness and faithfulness.

6 Other Attribution Systems

Thoppilan et al. (2022) introduce LaMDA, a dialogue-focused language model. While enlarging the model improves its quality, it does not necessarily enhance safety and accuracy. By fine-tuning LaMDA with annotated data and enabling it to access external knowledge, they significantly improve its safety and factual grounding. The grounding challenge of this study aims to generate responses based on credible external sources instead of merely plausible ones. The WebGPT model (Nakano et al., 2021) based on GPT-3 is trained to search and navigate the web and is fine-tuned for answering long-form questions in a web-browsing environment. For human evaluation of its factual accuracy, the model is required to gather references while browsing Microsoft Bing to support its answers. This ensures that the answers provided have a basis or attribution from credible web sources. Similarly, GopherCite (Menick et al., 2022) trained with reinforcement learning references evidence from multiple documents or a single user-provided document and refrains from answering when uncertain. Human evaluations show that GopherCite produces high-quality responses 80% at most. Nonetheless, citation alone is not a complete solution for ensuring safety and trustworthiness, as evidence-backed claims can still be false. Sparrow (Glaese et al., 2022) is trained to search the internet using Google Search to provide more accurate answers, allowing it to reference the latest information. In the user interface, evidence used by the model is displayed alongside its response, offering raters a means to validate the correctness of answer. To train the model in searching and using evidence, a preference model is used based on human judgments. Through human evaluation, it was found that responses with evidence were deemed plausible and supported 78% of the time. Comparisons between different systems are shown in Table 2.

7 Attribution Evaluation

Human Evaluation. To detect attribution errors, current attributed LLMs predominantly depend on human evaluation, a process that is both costly and time-intensive (Nakano et al., 2021; Kazemi

System	Model Training	Evidence Type	Citation Type	Integration
LaMDA (Thoppilan et al., 2022)	Multi-task SFT	Snippets	URLs	Appended
WebGPT (Nakano et al., 2021)	SFT + RL	Well-curated documents	Documents	Embedded
GopherCite (Menick et al., 2022)	SFT + RL	Long documents	Documents	Embedded
Sparrow (Glaese et al., 2022)	RL	Well-curated documents	Documents	Appended

Table 2: Features of different attribution systems. `SFT` means supervised fine-tuning, while `RL` means reinforcement learning optimization.

Evaluation Metrics	Evaluation Method	Description
Recall, Precision	Automatic, Statistics, Model-based	binary categorization based on NLI models
EM, BLEU, ROUGE	Automatic, Statistics	metrics for downstream tasks
QUIP-Score (Weller et al., 2023)	Automatic, Statistics	character-level n-gram metrics
Liu et al. (2023)	Human	fluency, perceived utility
AttrScore (Yue et al., 2023)	Human	attributability, extrapolatory, contradiction

Table 3: Comparison between different evaluation metrics for attribution.

et al., 2023; Chen et al., 2023a). For example, the typical cost of annotating a single (query, answer, reference) example stands at around $1 (Liu et al., 2023). In practical applications of attributed LLMs, the responsibility falls on users to be cautious of attributions and to undertake manual verification, imposing a significant responsibility on them.

Categorization-Based Evaluation. For the sake of clarity, earlier research mainly employed binary categorization by repurposing other NLP tasks (e.g., natural language inference) to determine whether an answer is supported by a reference or not (attributable or not) (Rashkin et al., 2021; Bohnet et al., 2022; Gao et al., 2023b; Muller et al., 2023). Liu et al. (2023) carry out a human assessment to evaluate the veracity of responses from generative search engines, categorizing the degree of reference support into full, partial, or no support. Building on this, Yue et al. (2023) introduce a refined categorization of attribution: 1) attributable–where the reference entirely backs the generated statement; 2) extrapolatory–where the reference offers insufficient backing for the statement; and 3) contradictory–where the statement directly opposes the referenced citation.

Quantitative Evaluation Metrics. Assessment of attribution quality is approached from three distinct angles (Li et al., 2023b): 1) Correctness–evaluating the alignment of generated text with the provided sources; 2) Precision–measuring the percentage of generated attributions pertinent to the question at hand; and 3) Recall–assessing the scope to which generated attributions capture crucial knowledge. Moreover, the F1-Score is derived from the Precision and Recall metrics. Thoppilan et al. (2022) introduces citation accuracy as the frequency with which the model refers to web sources for its assertions, excluding widely recognized truths. The QUIP-Score (Weller et al., 2023), an n-gram overlap metric, is founded on swift membership inquiries and evaluates the extent to which a section is comprised of exact spans within a text corpus.

As shown in Table 3, while human evaluations provide in-depth insights, their costly and time-consuming nature emphasizes the growing appeal for automated methods. Future research is expected to refine these methods, ensuring their practicality and reliability in real-world applications.

8 Discussion

8.1 Attribution Error Analysis

Attribution error has several forms. In this study, we systematically categorize these errors into three primary types, as outlined in Table 4, while acknowledging the possibility of other error types.

- **Granularity Error.** For ambiguous questions, the answer may involve multiple aspects. In this case, the retrieved multi-document may contain complex and diverse information. Thus the answer is complex and hybrid, leading to insufficient citation.

- **Mistaken Synthesis.** Models may mix up relationships between entities and events when several complex documents are provided. The citation should be faithful to the generated text and cite all the references.

- **Hallucinated Generation.** The reference documents may be irrelevant or not relevant to the question, or the model has conflicts between external documents and parameter knowledge.

The answer will be hallucinated and the citation is inaccurate.

8.2 Limitations of Attributions

Attribution in LLMs is fraught with inherent difficulties. One primary challenge is discerning when and how to attribute. Differentiating between general knowledge, which may not require citations, and specialized knowledge, which should ideally be attributed, is a nuanced task. This gray area can lead to inconsistencies in attribution (Huang and Chang, 2023). And LLMs now do not have ability to attribute parameter knowledge of itself (Litschko et al., 2023). Another limitation is the potential inaccuracy in attributions (Liu et al., 2023). LLMs might link content to irrelevant or incorrect sources. This misattribution can confuse users, leading them wrong and affecting the reliability of the information presented. For example, an LLM in the medical field could wrongly associate faulty medical guidance with a trustworthy reference, which might guide users towards detrimental health choices. Furthermore, the fluidity of knowledge means that while some information remains static, other data evolves and changes over time (Min et al., 2023). Consequently, some attributions made by LLMs may quickly become outdated, especially in rapidly advancing domains, such as computer science and clinical medicine. Additionally, we recommend readers refer to §4.1 in Menick et al. (2022).

8.3 Challenges for Attributions

Despite the potential solutions on the horizon, implementing these improvements for attributions is laden with challenges.

One such challenge is excessive attribution or over attribution (Huang and Chang, 2023; Liu et al., 2023). If LLMs give credit too often, users might get overwhelmed with too much information, confusing them and making it difficult to tell what is important and relevant from what is not.

At the same time, there is a real chance of LLMs accidentally revealing private information. Finding a balance between clear attribution and protecting private details is a tricky task.

Bias is another big challenge. LLMs might unintentionally lean towards some sources or kinds of information, pushing certain views while ignoring others. To tackle this bias, we need to use varied training data and improve the methods used for giving credit (Gunasekar et al., 2023).

Lastly, the shadow of incorrect information is ever-present. Without solid validation measures, LLMs could potentially spread wrong or misleading details, undermining the reliability of the information landscape. Future models should recognize ambiguous references and refrain from making statements when the evidence is not clear, instead of presenting unfounded claims.

Overall, though LLMs seem to be on a positive path, they face many obstacles and doubts. Proper credit is not just a side aspect; it is vital to the growth, approval, and effectiveness of LLMs. Guaranteeing correct and reliable credits, while promoting new ideas, will definitely influence the future of LLMs.

8.4 Future Directions for Attributions

Continuous Refreshment of LLMs. A promising direction for upcoming advancements is to create a system that consistently refreshes the information of LLMs (Thoppilan et al., 2022; Nakano et al., 2021), akin to how search engines update their databases. This approach not only ensures up-to-date content for attribution but also offers a platform for continuous learning and adaptation.

Enhancing the Reliability of LLM Outputs. Another pivotal direction entails boosting the trustworthiness of LLM outputs. This can be achieved by incorporating rigorous systems that assess the credibility and precision of the sources to which they attribute information (Min et al., 2023). Ensuring reliable and consistent sources will instill greater confidence in users about the content generated. As the adoption of LLMs expands across various domains, the reliability of their output becomes critical for informed decision making in various sectors.

Balancing Creativity with Proper Credit Attribution. LLMs are recognized for their creative content generation. Striking a balance between this inventive ability and proper credit-giving is a delicate act that needs investigation. While creativity is one of the significant strengths of LLMs, it is vital to ensure that the generated content remains trustworthy and rooted in factual bases. The aim is to make sure LLMs acknowledge sources without hindering their creative potential. Balancing these two aspects can foster an environment where users both benefit from the model and trust its outputs.

Limitation

While language models have the capability to cite their sources, undeniably enhance their utility, several limitations arise that need careful consideration (cf. Section 8.2). Our paper, in its current form, does not provide a solution to navigate such complex territory. It is important to address these limitations in future works and to continually educate users about the potential pitfalls of relying solely on machine-generated text.

References

Ayush Agrawal, Lester Mackey, and Adam Tauman Kalai. 2023. Do language models know when they're hallucinating references? *CoRR*, abs/2305.18248.

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Tachard Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Z. Chen, Eric Chu, J. Clark, Laurent El Shafey, Yanping Huang, Kathleen S. Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan A. Botha, James Bradbury, Siddhartha Brahma, Kevin Michael Brooks, Michele Catasta, Yongzhou Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, C Crépy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, M. C. D'iaz, Nan Du, Ethan Dyer, Vladimir Feinberg, Fan Feng, Vlad Fienber, Markus Freitag, Xavier García, Sebastian Gehrmann, Lucas González, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, An Ren Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wen Hao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Mu-Li Li, Wei Li, Yaguang Li, Jun Yu Li, Hyeontaek Lim, Han Lin, Zhong-Zhong Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alexandra Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Marie Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniela Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Ke Xu, Yunhan Xu, Lin Wu Xue, Pengcheng Yin, Jiahui Yu, Qiaoling Zhang, Steven Zheng, Ce Zheng, Wei Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. 2023. Palm 2 technical report. *ArXiv*, abs/2305.10403.

Anonymous. 2023. Learning to plan and generate text with citations. In *Submitted to The Twelfth International Conference on Learning Representations*. Under review.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *CoRR*, abs/2310.11511.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, T. J. Henighan, Nicholas Joseph, Saurav Kadavath, John Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom B. Brown, Jack Clark, Sam McCandlish, Christopher Olah, Benjamin Mann, and Jared Kaplan. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *ArXiv*, abs/2204.05862.

Bernd Bohnet, Vinh Q. Tran, Pat Verga, Roee Aharoni, Daniel Andor, Livio Baldini Soares, Jacob Eisenstein, Kuzman Ganchev, Jonathan Herzig, Kai Hui, Tom Kwiatkowski, Ji Ma, Jianmo Ni, Tal Schuster, William W. Cohen, Michael Collins, Dipanjan Das, Donald Metzler, Slav Petrov, and Kellie Webster. 2022. Attributed question answering: Evaluation and modeling for attributed large language models. *CoRR*, abs/2212.08037.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W. Rae, Erich Elsen, and Laurent Sifre. 2022. Improving language models by retrieving from trillions of tokens. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 2206–2240. PMLR.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 632–642. The Association for Computational Linguistics.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1870–1879. Association for Computational Linguistics.

Hung-Ting Chen, Fangyuan Xu, Shane A. Arora, and Eunsol Choi. 2023a. Understanding retrieval augmentation for long-form question answering.

Jifan Chen, Grace Kim, Aniruddh Sriram, Greg Durrett, and Eunsol Choi. 2023b. Complex claim verification with evidence retrieved in the wild. *CoRR*, abs/2305.11859.

I-Chun Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, and Pengfei Liu. 2023. Factool: Factuality detection in generative AI - A tool augmented framework for multi-task and multi-domain scenarios. *CoRR*, abs/2307.13528.

Li Du, Yequan Wang, Xingrun Xing, Yiqun Ya, Xiang Li, Xin Jiang, and Xuezhi Fang. 2023. Quantifying and attributing the hallucination of large language models via association analysis. *CoRR*, abs/2309.05217.

Besnik Fetahu, Katja Markert, Wolfgang Nejdl, and Avishek Anand. 2016. Finding news citations for wikipedia. In *Proceedings of the 25th ACM International Conference on Information and Knowledge Management, CIKM 2016, Indianapolis, IN, USA, October 24-28, 2016*, pages 337–346. ACM.

Martin Funkquist, Ilia Kuznetsov, Yufang Hou, and Iryna Gurevych. 2022. Citebench: A benchmark for scientific citation text generation.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Y. Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. 2023a. RARR: researching and revising what language models say, using language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 16477–16508. Association for Computational Linguistics.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023b. Enabling large language models to generate text with citations. *CoRR*, abs/2305.14627.

Amelia Glaese, Nat McAleese, Maja Trebacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin J. Chadwick, Phoebe Thacker, Lucy Campbell-Gillingham, Jonathan Uesato, Po-Sen Huang, Ramona Comanescu, Fan Yang, Abigail See, Sumanth Dathathri, Rory Greig, Charlie Chen, Doug Fritz, Jaume Sanchez Elias, Richard Green, Sona Mokrá, Nicholas Fernando, Boxi Wu, Rachel Foley, Susannah Young, Iason Gabriel, William Isaac, John Mellor, Demis Hassabis, Koray Kavukcuoglu, Lisa Anne Hendricks, and Geoffrey Irving. 2022. Improving alignment of dialogue agents via targeted human judgements. *CoRR*, abs/2209.14375.

Jocelyn Gravel, Madeleine D'Amours-Gravel, and Esli Osmanlliu. 2023. Learning to fake it: Limited responses and fabricated references provided by chatgpt for medical questions. *Mayo Clinic Proceedings: Digital Health*, 1(3):226–234.

Nianlong Gu and Richard H. R. Hahnloser. 2022. Controllable citation text generation. *CoRR*, abs/2211.07066.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuanzhi Li. 2023. Textbooks are all you need. *CoRR*, abs/2306.11644.

Zhijiang Guo, Michael Sejr Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Trans. Assoc. Comput. Linguistics*, 10:178–206.

Xiaochuang Han, Daniel Simig, Todor Mihaylov, Yulia Tsvetkov, Asli Celikyilmaz, and Tianlu Wang. 2023. Understanding in-context learning via supportive pretraining data. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 12660–12673. Association for Computational Linguistics.

Xiaochuang Han and Yulia Tsvetkov. 2022. ORCA: interpreting prompted language models via locating supporting data evidence in the ocean of pretraining data. *CoRR*, abs/2205.12600.

Hangfeng He, Hongming Zhang, and Dan Roth. 2023. Rethinking with retrieval: Faithful large language model inference. *CoRR*, abs/2301.00303.

Jie Huang and Kevin Chen-Chuan Chang. 2023. Citation: A key to building responsible and accountable large language models. *CoRR*, abs/2307.02185.

Siqing Huo, Negar Arabzadeh, and Charles L. A. Clarke. 2023. Retrieving supporting evidence for generative question answering. *CoRR*, abs/2309.11392.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 874–880. Association for Computational Linguistics.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online. Association for Computational Linguistics.

Palak Jain, Livio Baldini Soares, and Tom Kwiatkowski. 2023. 1-pager: One pass answer generation and evidence retrieval. *CoRR*, abs/2310.16568.

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.

Ehsan Kamalloo, Aref Jafari, Xinyu Zhang, Nandan Thakur, and Jimmy Lin. 2023. HAGRID: A human-llm collaborative dataset for generative information-seeking with attribution. *arXiv:2307.16883*.

Ryo Kamoi, Tanya Goyal, Juan Diego Rodriguez, and Greg Durrett. 2023. Wice: Real-world entailment for claims in wikipedia. *CoRR*, abs/2303.01432.

Mehran Kazemi, Najoung Kim, Deepti Bhatia, Xin Xu, and Deepak Ramachandran. 2023. LAMBADA: Backward chaining for automated reasoning in natural language. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6547–6568, Toronto, Canada. Association for Computational Linguistics.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, pages 39–48. ACM.

Himabindu Lakkaraju, Dylan Slack, Yuxin Chen, Chenhao Tan, and Sameer Singh. 2022. Rethinking explainability as a dialogue: A practitioner's perspective. *CoRR*, abs/2202.01875.

Dongyub Lee, Taesun Whang, Chanhee Lee, and Heuiseok Lim. 2023. Towards reliable and fluent large language models: Incorporating feedback learning loops in qa systems. *arXiv preprint arXiv:2309.06384*.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 6086–6096. Association for Computational Linguistics.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Dongfang Li, Baotian Hu, Qingcai Chen, Tujie Xu, Jingcong Tao, and Yunan Zhang. 2022. Unifying model explainability and robustness for joint text classification and rationale extraction. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 10947–10955. AAAI Press.

Xiaonan Li, Changtai Zhu, Linyang Li, Zhangyue Yin, Tianxiang Sun, and Xipeng Qiu. 2023a. Llatrieval: Llm-verified retrieval for verifiable generation. *arXiv preprint arXiv:2311.07838*.

Xinze Li, Yixin Cao2, Liangming Pan, Yubo Ma, and Aixin Sun. 2023b. Towards verifiable generation: A benchmark for knowledge-aware language model attribution.

Robert Litschko, Max Müller-Eberstein, Rob van der Goot, Leon Weber, and Barbara Plank. 2023. Establishing trustworthiness: Rethinking tasks and model evaluation.

Frederick Liu and Besim Avci. 2019. Incorporating priors with feature attribution on text classification. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 6274–6283. Association for Computational Linguistics.

Nelson F. Liu, Tianyi Zhang, and Percy Liang. 2023. Evaluating verifiability in generative search engines. *ArXiv*, abs/2304.09848.

Chaitanya Malaviya, Subin Lee, Sihao Chen, Elizabeth Sieber, Mark Yatskar, and Dan Roth. 2023. Expertqa: Expert-curated questions and attributed answers. *ArXiv*, abs/2309.07852.

Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, Francis Song, Martin Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, and Nathan McAleese. 2022. Teaching language models to support answers with verified quotes. *ArXiv*, abs/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. *CoRR*, abs/2305.14251.

Benjamin Muller, John Wieting, Jonathan H. Clark, Tom Kwiatkowski, Sebastian Ruder, Livio Baldini Soares, Roee Aharoni, Jonathan Herzig, and Xinyi Wang. 2023. Evaluating and modeling attribution for cross-lingual question answering. *CoRR*, abs/2305.14332.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. 2021. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke E. Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Francis Christiano, Jan Leike, and Ryan J. Lowe. 2022. Training language models to follow instructions with human feedback. *ArXiv*, abs/2203.02155.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The pagerank citation ranking : Bringing order to the web. In *The Web Conference*.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra-Aimée Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedweb dataset for falcon llm: Outperforming curated corpora with web data, and web data only. *ArXiv*, abs/2306.01116.

Denis Peskoff and Brandon Stewart. 2023. Credible without credit: Domain experts assess generative language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 427–438. Association for Computational Linguistics.

Fabio Petroni, Samuel Broscheit, Aleksandra Piktus, Patrick S. H. Lewis, Gautier Izacard, Lucas Hosseini, Jane Dwivedi-Yu, Maria Lomeli, Timo Schick, Pierre-Emmanuel Mazaré, Armand Joulin, Edouard Grave, and Sebastian Riedel. 2022. Improving wikipedia verifiability with AI. *CoRR*, abs/2207.06220.

Aleksandra Piktus, Christopher Akiki, Paulo Villegas, Hugo Laurençon, Gérard Dupont, Sasha Luccioni, Yacine Jernite, and Anna Rogers. 2023. The ROOTS search tool: Data transparency for llms. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics: System Demonstrations, ACL 2023, Toronto, Canada, July 10-12, 2023*, pages 304–314. Association for Computational Linguistics.

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. *CoRR*, abs/2304.04358.

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.

Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. 2021. Measuring attribution in natural language generation models. *CoRR*, abs/2112.12870.

Vipula Rawte, A. Sheth, and Amitava Das. 2023. A survey of hallucination in large foundation models. *ArXiv*, abs/2309.05922.

Revanth Gangi Reddy, Yi R. Fung, Qi Zeng, Manling Li, Ziqi Wang, Paul Sullivan, and Heng Ji. 2023. Smartbook: Ai-assisted situation report generation. *CoRR*, abs/2303.14337.

Tal Schuster, Ádám D. Lelkes, Haitian Sun, Jai Gupta, Jonathan Berant, William W. Cohen, and Donald Metzler. 2023. SEMQA: semi-extractive multi-source question answering. *CoRR*, abs/2311.04886.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 3784–3803. Association for Computational Linguistics.

Zhiqing Sun, Xuezhi Wang, Yi Tay, Yiming Yang, and Denny Zhou. 2023. Recitation-augmented language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Yi Tay, Vinh Tran, Mostafa Dehghani, Jianmo Ni, Dara Bahri, Harsh Mehta, Zhen Qin, Kai Hui, Zhe Zhao, Jai Prakash Gupta, Tal Schuster, William W. Cohen, and Donald Metzler. 2022. Transformer memory as a differentiable search index. In *NeurIPS*.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam M. Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, Yaguang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Yanqi Zhou, Chung-Ching Chang, I. A. Krivokon, Willard James Rusch, Marc Pickett, Kathleen S. Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Hartz Søraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Díaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravindran Rajakumar, Alena Butryna, Matthew Lamm, V. O. Kuzmina, Joseph Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Rogers Croak, Ed Huai hsin Chi, and Quoc Le. 2022. Lamda: Language models for dialog applications. *ArXiv*, abs/2201.08239.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and verification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 809–819. Association for Computational Linguistics.

Haoran Wang and Kai Shu. 2023. Explainable claim verification via knowledge-grounded reasoning with large language models.

Orion Weller, Marc Marone, Nathaniel Weir, Dawn Lawrie, Daniel Khashabi, and Benjamin Van Durme. 2023. " according to..." prompting language models improves quoting from pre-training data. *arXiv preprint arXiv:2305.13252*.

Jia-Yan Wu, Alexander Te-Wei Shieh, Shih-Ju Hsu, and Yun-Nung Chen. 2021. Towards generating citation sentences for multiple references with intent control. *CoRR*, abs/2112.01332.

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

Xinyu Xing, Xiaosheng Fan, and Xiaojun Wan. 2020. Automatic generation of citation texts in scholarly papers: A pilot study. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 6181–6190. Association for Computational Linguistics.

Shicheng Xu, Liang Pang, Huawei Shen, Xueqi Cheng, and Tat-Seng Chua. 2023. Search-in-the-chain: Towards the accurate, credible and traceable content generation for complex knowledge-intensive tasks. *CoRR*, abs/2304.14732.

Hongbin Ye, Tong Liu, Aijia Zhang, Wei Hua, and Weiqiang Jia. 2023a. Cognitive mirage: A review of hallucinations in large language models. *ArXiv*, abs/2309.06794.

Xi Ye, Ruoxi Sun, Sercan Ö Arik, and Tomas Pfister. 2023b. Effective large language model adaptation for improved grounding. *arXiv preprint arXiv:2311.09533*.

Xiang Yue, Boshi Wang, Kai Zhang, Ziru Chen, Yu Su, and Huan Sun. 2023. Automatic evaluation of attribution by large language models. *CoRR*, abs/2305.06311.

Shuo Zhang, Liangming Pan, Junzhou Zhao, and William Yang Wang. 2023a. Mitigating language model hallucination with interactive question-knowledge alignment. *CoRR*, abs/2305.13669.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023b. Siren's song in the ai ocean: A survey on hallucination in large language models. *ArXiv*, abs/2309.01219.

Guido Zuccon, Bevan Koopman, and Razia Shaik. 2023. Chatgpt hallucinates when attributing answers.

A Attribution Before the Era of Large Language Model

A.1 Related Natural Language Processing Tasks

The relationship between attribution tasks and other Natural Language Processing (NLP) tasks manifests in the overarching goal of understanding, evaluating, and leveraging the content retrieved or generated in response to particular stimuli such as questions or claims. Here is an exploration of how attribution tasks are intertwined with other NLP tasks, anchored on the retrieval of related content:

Open-domain question answering: Both tasks hinge on retrieving pertinent documents or information to address a posed question or claim. While open-domain QA zeroes in on the accuracy and relevance of the answer, attribution tasks scrutinize whether the answer or generated text can be accurately traced back to the retrieved documents (Chen et al., 2017; Bohnet et al., 2022).

Fact-checking & Claim verification (subtask of fact checking): Fact-checking and attribution tasks both necessitate the retrieval of external evidence to validate a claim or generated text. The emphasis in fact-checking is on verifying the truthfulness of a claim, whereas attribution tasks focus on the correct attribution of generated text to the sourced evidence (Thorne et al., 2018). On the other hand, attribution tasks and claim verification both center around validating information against reference or sourced material, yet they serve different purposes. Attribution ensures that generated text or answers accurately reflect the provided references, while claim verification assesses the truthfulness of a claim based on evidence or source material (Guo et al., 2022). Both tasks necessitate the retrieval of related content for verification (Wang and Shu, 2023), making them inherently reliant on the accuracy and relevance of the retrieved material. claim verification pivotal in fact-checking and misinformation detection, they share the fundamental objective of endorsing the accuracy and trustworthiness of information by juxtaposing it against a reference.

Natural Language Inference (NLI): Both tasks engage in evaluating the relationship between two snippets of text; however, NLI concentrates on logical entailment, contradiction, or neutrality, while attribution evaluates the substantiation provided by references for generated text (Bowman et al., 2015).

Summarization: Summarization and attribution tasks both generate condensed or altered text and necessitate a check on the fidelity of the generated text to the original or sourced content. Attribution in summarization is pivotal to averting hallucinations (generation of false or unsupported information) and ensuring the summary accurately mirrors the input text (Ji et al., 2023).

The commonality among these tasks lies in the requisite to retrieve, analyze, and validate content against some form of reference material, be it external evidence, retrieved documents, or a different segment of text. The capacity to retrieve related content forms a cornerstone for these tasks, enabling the necessary comparisons and evaluations to ascertain accuracy, relevance, and correct attribution.

A.2 Interpretability of NLP Models

Interpretability (e.g., feature attribution) dives into understanding which parts of the input (e.g., words or phrases) are crucial for a model's decision or output (Liu and Avci, 2019; Li et al., 2022). It helps in identifying the importance of different features in the input data concerning the model's performance. Compared to feature attribution, explicit attribution for LLMs serves as a conduit to trace the sources of the information they generate, which is pivotal for accountability, especially in critical domains like healthcare or finance. It enables verifiability, allowing users or other systems to check the accuracy and reliability of the information provided. Trustworthiness is also fostered through explicit attribution, as users are more likely to trust the model if they know where the information is coming from. Additionally, it plays a role in interpretability, aiding users in understanding how the model arrives at certain conclusions by revealing the sources of information.‘ This alignment with interpretability objectives helps in making the model's decision-making process more transparent and comprehensible.

Attribution and interpretability, though interconnected, serve distinct purposes. Attribution specifically refers to the process of tracing back the generated information or decisions of a model to its source material or input features, providing a clear reference or basis for the output. On the other hand, interpretability is a broader concept encompassing the understanding of how a model processes input data to arrive at a particular output (Lakkaraju et al., 2022), making the inner workings of the model transparent and comprehensible to users. While attribution can be seen as a component or a specific form of interpretability, aiding in understanding and trusting the model's outputs by providing source references, interpretability dives deeper into elucidating the model's behavior, decision-making process, and the significance of different input features in those decisions, thus fostering a comprehensive understanding and trust in the model. Note that the evaluation of attribution focuses on the verifiability of the answer based on the reference, similar to *faithfulness* evaluation of explanations (Jacovi and Goldberg, 2020). It does not measure the relevance about whether the answer correctly responds to the query, similar to *plausibility* in Jacovi and Goldberg (2020).

Granularity Error	**Question:** Where was the biggest gold nugget ever found? **References:** [doc1] The largest gold nugget ever found in California was found in August 1869 in Sierra Buttes by five partners. [doc2] Considered by most authorities to be the biggest gold nugget ever found, the Welcome Stranger was found at Moliagul, Victoria, Australia in 1869. [doc3] Pepita Canaã, was found on September 13, 1983 by miners at the Serra Pelada Mine in the State of Para, Brazil. **Answer:** Most authorities consider the Welcome Stranger to be the biggest gold nugget ever found, found in 1869 in Moliagul, Victoria, Australia[2]. The biggest gold nugget found in Brazil was found in the State of Para and in United States was found in Sierra Buttes [3]. **Explanation:** The second sentence in generated answer has two facets(the biggest nugget in Barzil, and the biggest nugget in United States). However, only document 3 is cited.
Mistaken Synthesis	**Question:** What's the difference in the portrayal of Ares in Roman mythology and Greek mythology? **References:** [doc1] In Roman mythology, he is called Mars. The Latin name for Mars derives from his Roman name; "Tuesday" comes from the corresponding deity Tyr in Norse mythology...[doc2] Ares is the god of war in ancient Greek mythology, one of the twelve Olympian gods. He is the son of Zeus and Hera, or alternatively, he was born after Juno (Hera's Roman name) swallowed a violent-eyed giant snake... **Answer:** In Roman mythology, Ares is called Mars, one of the twelve main gods of Rome, representing the symbols of power and war[1]. In Greek mythology, Ares is also one of the twelve Roman gods, representing the symbols of power and war[2]. **Explanation:** The answer mistakenly mixes up Olympian gods and Roman mythology thus both document 1 and document 2 should be treat as referenced documents.
Hallucinated Generation	**Question:** How many times have the lakers won the finals? **References:** [doc1] ... won six times in the 1960s, three times in the 1980s, in 2008, and in 2010... [doc2] ...and Lakers met in the Finals for the first time since 1987, with the Celtics winning the series 4-2. They faced off once again in the 2010 NBA Finals, which the Lakers won in 7 games... **Answer:** The Los Angeles Lakers have won the NBA Finals five times, in 1980, 1982, 1985, 1987, and 2010[1]. **Explanation:** The answer is incorrect as the generation is not grounded on reference documents. Hallucination may happen because of the conflict between reference documents and parameter knowledge, thus the citation will result in failure.

Table 4: List of different attribution errors types and example instance.