

Factuality of Large Language Models: A Survey

Yuxia Wang¹, Minghan Wang², Muhammad Arslan Manzoor¹, Fei Liu³, Georgi Georgieva¹, Rocktim Jyoti¹

Abstract

Large language models (LLMs), especially when instruction-tuned for chat, have become part of our daily lives, freeing people from the process of searching, extracting, and, integrating information from multiple sources by offering a straightforward answer to a variety of questions in a single place. Unfortunately, in many cases, LLM responses are factually incorrect, which limits their applicability in real-world scenarios. As a result, research on evaluating and improving the factuality of LLMs has attracted a lot of attention recently. In this survey, we critically analyze existing work with the aim to identify the major challenges and their associated causes, pointing out to potential solutions for improving the factuality of LLMs, and analyzing the obstacles to automated factuality evaluation for open-ended text generation. We further offer an outlook on where future research should go.

1 Introduction

Large language models (LLMs) have become an integral part of our daily lives. When instruction-tuned for chat, they have enabled digital assistants that can free people from the need to search, extract, and integrate information from multiple sources by offering straightforward answers in a single chat. While people naturally expect LLMs to always present reliable information that is consistent with real-world knowledge, LLMs tend to fabricate ungrounded statements, resulting in misinformation (Tonmoy et al., 2024), which limits their utility. Thus, assessing and improving the factuality of the text generated by LLMs has become an emerging and crucial research area, aiming to identify potential errors and to advance the development of more reliable LLMs (Chen

et al., 2023).

To this end, researchers have collected multiple datasets, introduced a variety of measures to evaluate the factuality of LLMs, and proposed numerous strategies leveraging external knowledge through retrieval, self-reflection, and early refinement in model generation to mitigate factual errors (Tonmoy et al., 2024). Numerous surveys (Tonmoy et al., 2024; Huang et al., 2023a; Wang et al., 2023b) have explored factuality or hallucinations in large language models across various modalities. While they either lack in-depth discussion or are too specific to grasp the fundamental challenges, promising solutions in factuality evaluation and enhancement, and some ambiguous concepts in LLM factuality. We summarized these surveys in Table ??.

Our survey aims to bridge this gap by providing an in-depth analysis of LLM factuality, with an emphasis on recent studies to reflect the rapidly evolving nature of the field. We offer a comprehensive overview of different categorizations, evaluation methods, and mitigation techniques for LLM factuality in both language and vision modalities. Additionally, we explore a novel research avenue that seeks to improve LLM calibration. This includes making models aware of their knowledge limitations and enhancing the reliability of their output confidence.

2 Background

Hallucination and factuality, while conceptually distinct, often occur in similar contexts and are sometimes used interchangeably, rendering them intricately intertwined, posing a challenge in discerning their distinct boundaries, and causing a considerable amount of misconception. In this section, we seek to disambiguate and refine

our understanding of these two closely aligned concepts, thereby preventing misinterpretation and reducing potential confusion. Additionally, we further include two closely-related axes : relevance and trustworthiness for LLM evaluation to illustrate their nuance in relation to factuality.

2.1 Hallucination vs. Factuality

The concept of hallucination in the context of traditional natural language generation tasks is typically referred to as the phenomenon in which the generated content appears nonsensical or unfaithful to the provided source content (Ji et al., 2023). One concrete example is made-up information in an abstractive summary with additional insights beyond the scope of the original source document.

In the age of LLMs, the term hallucination has been reimagined, encompassing any deviation from factual reality or the inclusion of fabricated elements within generated texts (Tonmoy et al., 2024; Rawte et al., 2023b). (Zhane et al., 2023c) define hallucination as the characteristic of LLMs to generate content that diverges from the user input, contradicts previously generated context, or mis-aligns with established world knowledge. (Huang et al., 2023b) merge the input- and context-conflicting types of hallucinations and further take logical inconsistency into account to form faithfulness hallucination. Another category is factuality hallucination, referring to the discrepancy between generated content and verifiable real-world facts, manifesting as (1) factual inconsistency and (2) factual fabrication.

Factuality, on the other hand, is concerned with a model's ability to learn, acquire, and utilize factual knowledge. (Wang et al., 2023b) characterize factuality issues as the probability of LLMs producing content inconsistent with established facts. It is important to note that hallucination content may not always involve factual missteps. Though a piece of generated text may exhibit divergence from the initial prompt's specifics, it falls into hallucinations, not necessarily a factual issue if the content is accurate.

It is crucial to distinguish between factual er-

rors and instances of hallucination. The former involves inaccurate information whereas the latter can present unanticipated and yet factually substantiated content (Wang et al., 2023b).

Summary: Factuality is the ability of LLMs to generate content consistent with factual information and world knowledge. Although both hallucinations and factuality may impact the credibility of LLMs in the context of content generation, they present distinct challenges. Hallucinations occur when LLMs produce baseless or untruthful content, not grounded in the given source. In contrast, factuality errors arise when the model fails to accurately learn and utilize factual knowledge. It is possible for a model to be factually correct yet still produce hallucinations by generating content that is either off-topic or more detailed than what is requested.

2.2 Trustworthiness/Reliability vs. Factuality

In the context of LLMs, factuality (Wang et al., 2023b) refers to a model's capability of generating contents of factual information, grounded in reliable sources (e.g., dictionaries, Wikipedia or textbooks), with commonsense, world and domain-specific knowledge taken into account. In contrast, „trustworthiness“ (Sun et al., 2024) extends beyond mere factual accuracy and is measured on eight dimensions: truthfulness, safety, fairness, robustness, privacy, ethics, transparency, and accountability.

3 Evaluating Factuality

Evaluating LLM factuality on open-ended generations presents a non-trivial challenge, discerning the degree to which a generated textual statement aligns with objective reality. Studies employ vari-

3.1 Datasets and Metrics

While (Zhang et al., 2023c) outlined tasks and measures for hallucination evaluation, there is no comparative analysis of existing datasets to assess various aspects in regards to model factuality (e.g., knowledge grounding, fast-changing

facts, snowballing hallucinations, robustness to false premises, and uncertainty awareness). We categorize the datasets in the format of discrimination or generation, and highlights the challenges in automatic evaluation for long-form open-ended generations. Current benchmarks largely assess the factuality in LLMs based on two capabilities: proficiency in distinguishing factual accuracy in a context and ability to generate factually sound content.

The former typically comes in the form of a multi-choice question, with the expected response being a label of one of A, B, C, and D. For instance, HotpotQA, StrategyQA, MMLU. This form of evaluation has been widely used to measure the general knowledge proficiency and factual accuracy of LLMs, largely thanks to its automation-friendly nature. Under this evaluation formulation, model responses are easily parsed and compared with gold standard labels, enabling the calculation of accuracy or F1 scores against established benchmarks.

Precisely assessing the factuality of free-form LLM outputs remains a significant challenge due to the inherent limitations of automatic methods in the face of open-ended generation and the absence of definitive gold standard responses within an expansive output space. To make automatic evaluation feasible, many studies constrain the generation space to (1) Yes/No; (2) short-form phrase; and (3) a list of entities through controlling the categories of questions and generation length.

Perhaps the most demanding, yet inherently realistic scenario is free-form long text generation, such as biography generation. For this, the most commonly used and reliable methods rely on human experts following specific guidelines, and automatic fact-checkers based on retrieved information, such as FactScore, Factool and Factcheck-GPT, to facilitate efficient and consistent evaluation.

These automatic fact-checkers generally first decompose a document into a set of atomic claims, and then verify one by one whether the claim is true or false based on the retrieved evidence, either from offline Wikipedia or online Web pages. The percentage of true claims

over all statements in a document is used to reflect the factual status of a response (refer to FactScore). The averaged FactScore over a dataset is in turn used to assess a model’s factuality accuracy. However, there is no guarantee that automatic fact-checkers are 100% accurate. (Rawte et al., 2023c) show that even the state-of-the-art verifier equipped with GPT-4 and supporting evidence retrieved with Google search, can only achieve an F1 score of 0.63 in identifying false claims and F1=0.53 using PerplexityAI (compared with human-annotated labels for claims: true or false).

Summary: We categorize datasets that evaluate LLM factuality into four types, depending on the answer space and the difficulty degree on which accurate automatic quantification can be performed (see Table ??). They are: (I) open-domain, free-form, long-term responses (FactScore: the percentage of the correct claims verified by human or automated fact-checker); (II) Yes/No answer w/ w/o explanation (extract Yes/No metrics for binary classification); (III) short-form answer (Exact match the answer with gold labels and calculate accuracy) or the listing answer (recall@K); and (IV) multi-choice QA (metrics for multi-class classification).

3.2 Other Metrics

In addition to evaluating the methods discussed above, (Lee et al., 2022) quantified the hallucinations using two metrics, both requiring document-level groundtruth: (1) hallucinated named entities error measures the percentage of named entities in the generations that do not appear in the ground-truth document; (2) entailment ratio evaluates the number of generations that can be entailed by the ground-truth reference, over all generations.

(Rawte et al., 2023a) defined the hallucination vulnerability index (HVI), which takes a spectrum of factors into account, to evaluate and rank LLMs.

Some factuality measurement tasks, such as claim extraction and evidence retrieval are non-trivial to automate. (Rawte et al., 2023a) curated publicly available LLM hallucination mitigation benchmark, where LLM generations are scored

by humans when automated extemal knowledge retrieval fails to resolve a claim clearly. While widely used for factuality evaluation, this hybrid approach may suffer from human annotation bias.

4 Improving Factuality

Improving the factuality of an LLM often requires updating its internal knowledge, editing fake, out- dated and biased elements, thereby making its out- put reflect a revised collection of facts, maximiz- ing the probability of $P(\text{truth}|\text{prompt})$. One op- tion is to adopt gradient-based methods to update model parameters to encourage desired model out- put. This includes pre-training, supervised fine- tuning and RLXF. We can also explore injecting a new fact into LLMs or overwriting the false knowl- edge stored in LLM memory by in-context learning (ICL). When models store factually correct knowl- edge but produce errors, they can in some cases rectify them through self- reasoning, reflection, and multi-agent debates.

We discuss these methods throughout the life- cycle of an LLM, ranging from pre-training, to inference, to post-processing. Another important element is retrieval augmentation, which enhances the generation capabilities of LLMs by anchoring them in external knowledge that may not be stored or contradict the information in LLM parametric memory. It can be incorporated at various stages throughout model training and the subsequent in- ference process (Gao et al., 2023b), and is therefore not discussed individu- ally.

4.1 Pre-training

LLMs store a vast amount of world knowledge in their parameters through the process of pre- training. The quality of the pre-training data plays a crucial role and misinformation could potentially cause LLMs to generate false responses, motivat- ing the utilization of high-quality textual corpora. However, the prohibitively massive amount of pre- training data, typically consisting of trillions of tokens, renders manual filtering and editing imprac- tically laborious. To this end, automated filtering methods have been proposed.

For instance, (Brown et a1.,2020) introduce a method to only focus on a small portion of the CommonCrawl dataset that ex- hibits similarity to high-quality reference corpora. (Touvron et al., 2023) prop se to enhance factual robustness of mixed corpora by up-sampling docu- ments from the most reliable sources, thereby am- plifying knowledge accuracy and mitigating hal- lucinations. During the pre-training phase of phi- 1.5, (Li and et a1.,2023b) synthesize "textbook- like" data, consists of and rich in high-quality commonsense reasoning and world knowledge. While careful copus curation remains the cornerstone of pre-training for enhanced factuality, the task be- comes increasingly challenging with the expansion of dataset scale and the growing demand for linguis- tic diversity. It is therefore crucial to develop novel strategies that guarantee the consistency of factual knowledge across diverse cultural landscapes.

(Borgeaud et a1.,2021) propose RETRO, a re- trieval augmented pre-training approach. An auto- regressive LLM is trained from scratch with a re- trieval module that is practically scalable to large- scale pre-training by retrieving billions oftokens. RETRO shows better accuracy and is less prone to hallucinate compared to GPT (Wang et al.,2O23a). While limitations lie in that RETRO performance could be compromised if the retrieval database con- tains inaccurate, biased or outdated information. -25ya additional computation is required for the pre-training of LLMs with retrieval.

4.2 Tlrning and RLXF

Continued domain-specific SFT has shown to be ef- fective for enhancing factuality, particu- larly in the absence of such knowledge during pre-training. For instance, (Elaraby eta1.,2023) enhance the factual accuracy of LLMs through knowledge injection (KI). Knowledge, in the form of entity summaries or entity triplets, is incorpo- rated through SFT by either intermediate tuning, i.e. first on knowledge and then on instruction data; or combined tuning, i.e. on the mixture of both. While some improve- ments are exhibited, the method alone can be insuf- ficient to fully

mitigate factual errors.

For general-purpose LLMs, SFT is typically employed to improve the instruction-following capabilities as opposed to factual knowledge which is mostly learned in pre-training. However, this process may inadvertently reveal areas of knowledge not covered in the pre-training, causing the risk of behavior cloning, where a model feigns understanding and responds with hallucinations to questions it has little knowledge of (Torabi et al., 2018). R-tuning (Zhang et al., 2023a) is proposed to address this issue with two pivotal steps: first, assessing the knowledge gap between the model’s parametric knowledge and the instruction tuning data, and second, creating a refusal-aware dataset for SFT. It enables LLMs to abstain from answering queries beyond their parametric knowledge scope. On the other hand, BelInfo (Razumovskaya et al., 2023) improves factual alignment through the form of behavioral fine-tuning. The creation of the behavioral tuning dataset emphasizes two goals: selectivity (choosing correct information from the knowledge source) and response adequacy (informing the user when no relevant information is available or asking for clarification). Both methods effectively control LLMs on non-parametric questions but require extra effort in dataset curation and might hinder the models’ retention of parametric knowledge.

Sycophancy (Sharma et al., 2023), another source of factuality errors, often arises from misalignments during SFT and RLHF (Ouyang et al., 2022). This is partially attributed to human annotators’ tendency to award higher scores to responses they like rather than those that are factually accurate. (Wei et al., 2023) explore the correlation of sycophancy with model scaling and instruction tuning. They propose a synthetic-data intervention method, using various NLP tasks to teach models that truthfulness is independent of user opinions. However, one limitation is that the generalizability of their approach remains unclear for varied prompt formats and diverse user opinions.

(Tian et al., 2023) utilize direct preference optimization (DPO) (Rafailov et al., 2023) with the feedback of factuality score either from automatic

fact-checkers or LLMs predictive confidence. In-domain evaluation shows promising results on biographies and medical queries, but generalization performance across domains and unseen domains is under-explored. (Kdksal et al., 2023) propose hallucination-augmented recitations (HAR). It encourages the model to attribute to the contexts rather than its parametric knowledge, by tuning the model on the counterfactual dataset created leveraging LLM hallucinations. This approach offers a novel way to enhance LLM attribution and grounding in open-book QA. However, challenges lie in refining counterfactual generation for consistency and expanding its application to broader contexts.

Retrieval-Augmentation Incorporating retrieval mechanisms during fine-tuning has been shown to enhance the LLM factuality on downstream tasks, particularly in open-domain QA. DPR (Karpukhin et al., 2020) refines a dual-encoder framework, consisting of two BERT models. It employs a contrastive loss to align the hidden representations of questions and their corresponding answers, obtained through the respective encoder models. RAG (Lewis et al., 2020) and FiD (Izacard and Grave, 2020) study a fine-tuning recipe for retrieval-augmented generation models, focusing on open-domain QA tasks. WebGPT (Nakano et al., 2021) fine-tunes GPT-3 (Brown et al., 2020) by RLHF, providing questions with factually correct long-form reference generation. The implementation in a text-based web-browsing environment allows the model to search and navigate the web.

4.3 Inference

We categorize approaches to improve factuality during inference into two: (1) optimizing decoding strategies to strengthen model factuality; and (2) empowering LLM learned ability by either in-context learning (ICL) or self-reasoning.

4.3.1 Decoding Strategy

Sampling from the top subword candidates with a cumulative probability of p , known as nucleus sampling (top-p) (Holtzman et al., 2020), sees a

decrease in factuality performance compared to greedy decoding, despite higher diversity. This is likely due to its over-encouragement of randomness. Building on the hypothesis that sampling randomness may damage factuality when generating the latter part of a sentence than the beginning, (Lee et al., 2022) introduce factual-nucleus sampling, which dynamically reduces the nucleus-p value as generation progresses to limit diversity and improve factuality, modulating factual integrity and textual diversity.

Apart from randomness, some errors arise when knowledge conflicts, where context contradicts information present in the model’s prior knowledge. Context-aware decoding (CAD) (Shi et al., 2023) prioritizes current context over prior knowledge, and employs contrastive ensemble logits, adjusting the weight of the probability distribution when predicting the next token with or without context. Despite the factuality boost, CAD is a better fit for tasks involving knowledge conflicts and heavily reliant on high-quality context.

In contrast, DoLa (Chuang et al., 2023) takes into account both upper and lower (earlier) layers, as opposed to only the final (mature) layer. This method dynamically selects intermediate layers at each decoding step, in which an appropriate premature layer contains less factual information with maximum divergence among the subset of the early layers. This method effectively harnesses the distinct contributions of each layer to factual generations. However, DoLa increases the decoding time by 1.0lx to 1.08x and does not utilize external knowledge, which limits its ability to correct misinformation learned during training.

4.3.2 ICL and Self-reasoning

In context learning (ICL) allows an LLM to leverage and learn from demonstration examples in its context to perform a particular task without the need to update model parameters. (Zheng et al., 2023) present that it is possible to perform knowledge editing via ICL through facts included in our demonstration examples, thereby correcting fake, or outdated facts. The objective

of demonstration examples is to teach LLMs how to: (1) identify and copy an answer; (2) generalize using in-context facts; (3) ignore irrelevant facts in context.

While it is rather easy for LLMs to copy answers from contexts, changing predictions of questions related to the new facts accordingly, and keeping the original predictions if the question is irrelevant to the modified facts, remains tough.

Another line of research leverages the self-reasoning capability of LLMs. (Du et al., 2023) improve LLM factuality through multi-agent debate. This approach first instantiates a number of agents and then makes them debate over answers returned by other agents until a consensus is reached. One interesting finding is that more agents and longer debates tend to lead to better results. This approach is orthogonal and can be applied in addition to many other generation methods, such as complex prompting strategy (e.g., CoT (Wei et al., 2022), ReAct (Yao et al., 2023), Reflexion (Shinn et al., 2023)) and retrieval augmentation.

Take-away: Zheng et al. (2023) evaluate the effectiveness of knowledge editing on subject-relation-object triplets, an unrealistic setting compared to open-ended free-form text assessment. Previous methods (Mitchell et al., 2021; Meng et al., 2022) use finetuning over texts containing specific text to improve factuality. The relationship between SFT and ICL may also be an interesting avenue to explore. More specifically, we seek answers to two research questions: (1) What types of facts and to what extent can facts be edited effectively, learned by LLMs through ICL? (2) Would SFT do a better job at learning from examples that are difficult for ICL? More broadly, what is the best way to insert new facts or edit false knowledge stored in LLMs. The community may also benefit from an in-depth comparative analysis of the effectiveness of improving factuality between SFT and ICL (perhaps also RLXF).

Retrieval Augmentation can be applied before, during, and after model generation. one commonly used option is to apply `retrievalaugmentationprioritoresponse`. For questions requiring up-to-date world knowledge to answer, retrieved information and demonstrate the effectiveness on im-

curacy on FreshQA, where ChatGPT and GpT-4 struggle to detect their falsehoods due to a coarse granularity, to-date information. (Gao et al., 2023a) place all credibility at the granular level. Compared to fine model to cite supporting evidence, instructing LLMs to understand trained classification guidelines, or strict rules, thereby improving

Pre-generation retrieval augmentation is beneficial as the generation process is conditioned on the retrieval results, implicitly constraining the output space. While improving factual accuracy, this comes at the cost of spontaneous and creative responses, largely limiting the capabilities of LLMs.

An alternative method is to verify and rectify factual errors after the model generates all content. However, LLMs have been shown to be susceptible to hallucination snowballing (Zhang et al., 2023b), a common issue where a model attempts to make its response consistent with previously generated content even if it is factually incorrect.

Striking a balance between preserving creative elements and avoiding error propagation, EVER (Kang et al., 2023) and "a stitch in time saves nine" (Varshney et al., 2023) actively detect and correct factual errors during generation sentence by sentence. The former leverages retrieved evidence for verification, and the latter incorporates the probability of dominant concepts in detection. Their findings suggest that timely correcting errors during generation can prevent snowballing and further improve factuality. Nonetheless, the primary concern for this iterative process of generate-verify-correct in real-time systems is latency, making it difficult to meet the high-throughput and responsiveness demand (Kang et al., 2023).

4.4 Automatic Fact Checkers

An automatic fact-checking framework typically consists of three components: claim processor, retriever, and verifier as shown in Figure ??, though the implementation of verification pipelines may differ. For example, FACTOR (Muhlgay et al., 2023) and Factscore (Min et al., 2023) only detect falsehoods without correction. While RA.RR depends on web-retrieved information (Gao et al., 2022), and CoVe (Dhuliawala et al., 2023) only relies on LLM parametric knowledge (Dhuliawala et al., 2023) to perform both detection

spot false spans precisely and tends to result in poor preservation of the original input. FactBol (Chern et al., 2023) and Factcheck-GPZ (Wang et al., 2023c) edit atomic claims. While the former breaks a document down to independent checkworthy claims with three steps: decomposition, decontextualization and checkworthiness identification, the latter employs GPT-4 to extract verifiable claims directly. Evaluating the effectiveness of fact-checkers remains challenging, making the improvement of such systems a difficult task.

Engineering and Practical Considerations Automatic fact-checking involve tasks of extracting atomic check-worthy claims, collecting evidence either by leveraging the knowledge stored in the model parameters or retrieved externally, and verification. While straightforward to implement, this pipeline may be susceptible to error propagation. Major bottleneck lies in the absence of automatic evaluation measures to assess the quality of intermediate steps, in particular, the claim processor and evidence retriever as there is no gold standard.

The input to a claim processor is a document and the expected output is a list of atomic check-worthy claims or atomic verifiable facts. There is no consensus on the granularity of "atomic claims", making consistent decomposition difficult. Additionally, the concept of check-worthy and verifiable claims are subjective. Consequently, the definition of an atomic check-worthy claim remains a highly debatable topic. This naturally leads to different "gold" human-annotated atomic claims annotated following various guidelines and distinct implementation approaches to decompose a document.

Given a document, even if assuming a ground-truth list of atomic claims, it is an open question how to assess the quality of automatically derived decomposition results. (Wang et al., 2023c) assess the agreement in the number of claims between ground truth and predictions, followed by examining the semantic similarity between two claims at the same index when the claim count aligns.

IMAGE NOT PROVIDED

Figure 1: Fact-checker framework: claim processor, retriever, and verifier, with optional step of summarizing and explaining in gray.

Entailment ratio presented in Section 3.2 is also applicable (Lee et al., 2022).

While it is much simpler when the evidence is constrained (e.g., to Wikipedia documents as is the case for FEVER (Thorne et al., 2018)), accurate retrieval of evidence from the Internet and subsequently quantifying the quality of such retrieval results remain challenging. Similar to the assessment of atomic claims, gold-labeled evidence is

unavailable and infeasible to obtain in the expansive open search space.

The only step where we can confidently evaluate its quality is the accuracy of verification, a simple binary true/false label given a document/claim. In conclusion, perhaps the most significant hurdle for the development and improvement of automatic fact-checkers lies in the automated assessment and quantification of the quality at intermediate stages.

5 Factuality of Multimodal LLMs

Factuality or hallucination in Multimodal Large Language Models refers to the phenomenon of generated responses being inconsistent with the image content. Current research on multimodal factuality can be further categorized into three types:

1. Existence Factuality: incorrectly claiming the existence of certain objects in the image.
2. Attribute Factuality: describing the attributes of certain objects in a wrong way, e.g. identifying the colour of a car incorrectly.
3. Relationship Factuality: false descriptions of relationships between objects, such as relative positions and interactions.

Evaluation

CHAIR (Rohrbach et al., 2018) is the first benchmark for assessing the accuracy of object existence within captions, focusing on a pre-defined set of objects in the COCO dataset (Lin et al., 2014). However, this approach can be misleading since the COCO dataset is frequently used in training sets, providing a limited perspective when used as the sole basis for evaluation. In contrast, POPE (Li et al., 2023) evaluates object hallucination with multiple binary choice prompts, both positive and negative, querying if a specific object exists in the image. More recently, (Li et al., 2023) proposed GPT4 -Assisted Visual Instruction Evaluation (GAVIE) to evaluate the visual hallucination. Additionally, (Gunjal et al., 2023) demonstrated the use of human evaluation to avoid inaccuracies and systematic biases.

Mitigation

The methods for improving factuality in MLLMs can be broadly categorized into the categories: finetuning-based method, inference time correction and representation learning.

Fine-tuning methods such as LRV-Instruction (Liu et al., 2023) and LLaVA-RLHF (Sun et al., 2023) follow an intuitive and straightforward solution of collecting specialized data such as positive and negative instructions or human preference pairs. This data is used for finetuning the model, thus resulting in models with fewer hallucinated responses. Whereas inference time approaches mitigate factuality by correcting output generation. Woodpecker (Yin et al., 2023a) and LURE (Zhou et al., 2023) use specialized models to rectify model generation. There are other works such as HallE-Switch (Zhai et al., 2023), VCD (Leng et al., 2023), and HACL (Jiang et al., 2023) that analyse and improve feature representation to improve factuality.

6 Challenges and Future Directions

We first identify three major challenges for improving the factuality of LLMs, and then we point to

several promising directions for future work.

Language models learn a language distribution, not facts. The training objective of language modeling is to maximize the probability of a sentence, as opposed to that of a factual statement. While capable of generating seemingly coherent and fluent outputs upon convergence, models are not guaranteed to always return a factual response.

Automatic evaluation of the factual accuracy of open-ended generations remains challenging. Existing studies on factuality enhancement use different benchmarks and evaluation measures, making fair comparisons difficult, which motivates the need for a unified automated evaluation framework that uses the same collection of datasets and metrics. Current approaches rely on either human evaluation or results of automated fact-checkers such as FactScore and FacTool (Min et al., 2023; Chem et al., 2023). However, automatically quantifying the quality of automated fact-checkers is itself an open question, resulting in a chicken and egg situation.

Latency and multi-hop reasoning could be the bottleneck of RAG systems. Retrievers serve as the core component in RAG systems, and the effectiveness of RAGs is largely influenced by the quality (coverage and relevance) of the retrieved documents. Latency and difficulties in gathering the most pertinent evidence are the primary challenges in retrieval. While this is partly due to the inability of ranking algorithms to retrieve such documents, certain facts require information gathered from various sources and multi-hop reasoning.

Potential Future Directions

Mitigation in inference: We observe that models can often generate a correct answer in multiple trials even if some attempts are wrong (Tian et al., 2023). This motivates us to ask how to provide an anchor that can guide LLM decoding to the factually correct path? Iteratively detecting, correcting, and generating during generation has been demonstrated to be effective to mitigate hallucinations. If simply correcting the first one or two sentences, how much improvements

can we expect for subsequent generations? Can factually correct and relevant sentences, phrases or concepts serve as anchors?

Development of better retrieval algorithms: Integrating Retrieval-Augmented Generation (RAG) into Large Language Models (LLMs) is challenging due to the prevalence of unreliable information, such as fake news, on the internet. This compromises the accuracy of the knowledge retrieved, resulting in LLMs generating responses based on incorrect input. Consequently, future research should focus on improving retrieval techniques to enhance the factuality of LLM-generated responses.

Improving the efficiency and the accuracy of automated fact-checkers : The key breakthrough in effectively evaluating the factual accuracy of LLMs lies in establishing accurate and efficient fact-checkers. This requires improvement of the quality of the evidence used for making veracity decisions. Moreover, many recent methods rely on the factuality of stronger models such as GPT-4 for claim verification. Not only is this computationally expensive, but it also tends to be highly sensitive to minor prompt changes and LLM updates. A small task-specific and well fine-tuned NLI model can be a more viable, robust, and cost-efficient option.

7 Conclusion

We presented an overview on the factuality of LLMs, surveying a number of studies covering topics such as evaluation and improvement methods (applicable at various stages: pre-training, SFT, inference and post-processing) along with their respective challenges. We also identified three major issues and pointed out to promising future research directions.

Limitations

Despite conducting an extensive literature review to encompass all existing research on the factuality of LLMs, some studies may have been omitted due to the rapidly evolving nature of this research area. We endeavored to include all pertinent stud-

ies and references wherever feasible. This survey only briefly touches upon the factuality issues associated with vision language models. However, there is room for a more in-depth exploration of mitigation techniques specific to vision-language models. Additionally, comprehensive discussions are also necessary for language models that incorporate other modalities, such as video and speech.

References

- [1] Sebastian Borgeaud, Arthur Mensch, and Jordan Hoffmann et al. 2021. Improving language models by retrieving from trillions of tokens. In ICML.
- [2] Tom B. Brown, Benjamin Mann, Nick Ryder, and Melanie Subbiah et al. 2020. Language models are few-shot learners. In NeurIPS 2020.
- [3] Shiqi Chen, Yiran Zhao, Jinghan Zhang, and, et al. 2023. Felm: Benchmarking factuality evaluation of large language models. arXiv preprint arXiv : 2310.00741 .
- [4] I-Chun Chern, Steffi Chem, and Shiqi Chen et al. 2023. Factool: Factuality detection in generative AI - A tool augmented framework for multi-task and multi-domain scenarios. CoRR, abs/2307.13528.
- [5] Yung-Sung Chuang, Yujia Xie, and Hongyin Luo et al. 2023. Dola: Decoding by contrasting layers improves factuality in large language models. CoRR, abs/2309.03883.
- [6] Shehzaad Dhuliawala, Mojtaba Komeili, and et al. 2023. Chain-of-verification reduces hallucination in large language models. arXiv preprint arXiv : 2309.11495 .
- [7] Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. CoRR, abs/2305.14325.
- [8] Mohamed Elaraby, Mengyin Lu, and Jacob Dunn et al. 2023. Halo: Estimation and reduction of hallucinations in open-source weak large language models. CoRR, abs/2308.11764.
- [9] Luyu Gao, Zhuyun Dai, and Panupong Paupat. 2022. Attributed text generation via post-hoc research and revision. arXiv preprint arXiv:2210.08726.
- [10] Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023a. Enabling large language models to generate text with citations. In EMNLP, pages 6465-6488.
- [11] Yunfan Gao, Yun Xiong, and et al. 2023b. Retrieval-augmented generation for large language models: A survey. CoRR, abs/2312.10997 .
- [12] Mor Geva, Daniel Khashabi, and et al. 2021. Did Aristotle use a laptop? A question answering benchmark with implicit reasoning strategies. TACL, 9:346-361 .
- [13] Anish Gunjal, Jihan Yn, and Erhan Bas. 2023. Detecting and preventing hallucinations in large vision language models. In AAAI Conference on Artificial Intelligence.
- [14] Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. TACL, 10:178-206.
- [15] Dan Hendrycks, Collin Burns, and et al. 2022L. Measuring massive multitask language understanding. In ICLR 2021 .
- [16] Junyi Li and Xiaoxue Cheng et al. 2023a. Halueval: A large-scale hallucination evaluation benchmark for large language models. CoRR, abs/2305.1174i.
- [17] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Jirong Wen. 2023. Evaluating object hallucination in large vision-language models. In Conference on Empirical Methods in Natural Language Processing, 88-99.
- [18] Yuanzhi Li and Sébastien Bubeck et al. 2023b. Textbooks are all you need II: phi-1.5 technical report. CoRR, abs/2309.05463.

- [19] Stephanie Lin, Jacob Hilton, and Owain Evans.2022. Truthfulqa: Measuring how models mimic human falsehoods. In ACL, pages 3214-3252.

[20] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick.2014. Microsoft coco: Common objects in context. In European Conference on Computer Vision.

[21] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, yaser Yacoob, and Lijuan Wan.2023. Mitigating hallucination in large multimodal models via robust instruction tuning.

[22] Kevin Meng, David Bau, Alex Andonian, and yonatan Belinkov. 2022. Locating and editing factual associations in gpt. In NeurIPS Information processing Systems.

[23] Sewon Min and Kalpesh Krishna et al.2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. CoRR, abs/2305.14251 .

[24] Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D. Manning. 2021. Fast model editing at scale. ArXiv [B4] abs/2110.11309.

[25] Dor Muhlgay, Ori Ram, and Inbal Magar et al.2023. Generating benchmarks for factuality evaluation of language models. CoR[B5] abs/2307.06908.

[26] Reiichiro Nakano, Jacob Hilton, and et al. 2021. Webgpt: Browser-assisted question answering with human feedback . ArXiv, abs/2112.09332.

[27] Long Ouyang, Jeff Wu, and Xu Jiang [B7] aL.2022. Training language models to follow instructions with human feedback . ArXiv, abs/2203.02155 .

[28] Rafael Rafailov, Archit Sharma, and et al. 2023. Direct preference optimization: Your language model is secretly a reward model. CoRR, abs/2305.18290.

[29] Vipula Rawte, Swagata Chakraborty, and Agnibh et al. Pathak. 2023a. The troubling emergence of hallucination in large language models - an extensive definition, quantification, and prescriptive remediations. In EMNLP 2023, pages 2541-2513.

[30] Vipula Rawte, Amit P. Sheth, and Amitava Das. 2023b. A survey of hallucination in large foundation models. CoRR, abs/2309.05922.

[31] Evgenia Razumovskaya, Ivan Vulic, and Pavle Markovic et al. 2023. Dial beinfo for faithfulness: Improving factuality of information-seeking dialogue via behavioural fine-tuning. CoRR, abs/2311.09800.

[32] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In Conference on Empirical Methods in Natural Language processing.

[33] Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, and Amanda Askell et al.2023. Towards understanding sycophancy in language models. CoRR, abs/2305.14251 .

[34] Weijia Shi, Xiaochuang Han, and et al.2023. Trusting your evidence: Hallucinate less with context-aware decoding. arXiv preprint arXiv : 2305.14739.

[35] Noah Shinn, Federico Cassano, and Gopinath et al.2023. Reflexion: Language agents with verbal reinforcement learning . In NeuralPS.

[36] Lichao Sun, Yue Huang, and Haoran Wang et al.2024. Trustilm: Trustworthiness in large language models. ArXiv, abs/2401.05561 .

[37] Boxin Wang, Wei Ping, and et al. 2023a. Shall we pretrain autoregressive language models with retrieval? a comprehensive study. ArXiv, abs/2305.14251 .

[38] Cunxiang Wang, Xiaoze Liu, and et al. 2023b. Survey on factuality in large language models : Knowledge - edge, retrieval and domain - specificity. ArXiv, abs/2310.07521.

- [39] Yuxia Wang, Revanth Gangi Reddy, and [51] Yi11ng Zhou, Chenhang Cui, Jaehong yoon, Lin-
a].2023c. Factcheck-gpt: End-to-end fine-grained document- level fact-checking and correction of LLM output. CoRR, abs/2311.09000.
- [40] Jason Wei, Xuezhi Wang, and Dale et al.2022. Chain of [52] Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian thoughtpromptingelicitsreasoninginlarge language models. ArXiv, abs/2310.00754.
- [41] Jerry W Wei, Da Huang, and yifeng L et a].2023. Simple synthetic data reduces sycophancy in large language models. Co RR, abs/2308.0395g.
- [42] Zhilin Yang and Peng Qi et al. 201g. Hot- [53] James Thorne, Andreas Vlachos, and et al. 2018. potqa: A dataset for diverse, explainable multi- FEVER: a large-scale dataset for fact extraction hopquestion answering. In EMNLP 20 I 8, pages and VERification. In NAACL, pages 809-819.
- [43] Shunyu Yao, Jeffrey Zhao, andDian et al. 2023. React: Synergizing reasoning and acting in lan- [54] Katherine Tian, Eric Mitchell, and et al. 2023. guage models. In ICLR. Fine- tuning language models for factuality. arXiv preprint arXiv:231 1.08401 .
- [44] Shukang Yin, Chaoyou Fu, and et al. 2023a. Woodpecker : [55] S. M. Towhidul Islam Tonmoy, S. M. Mehedi Hallucinationcorrectionformultimodal largelanguagmodels. GRR, abs/2310.16045.01313. language models. CoRR, abs/2401.01313.
- [45] Zhangye Yin, Qiushi Sun, and eipeng Guo et al. [56] Faraz Torabi, Garrett Warnell, and Peter Stone. 2023b. Do large language models know what they 2018. Behavioral cloning from observation. In don,t know? In ACL,pages 8653-8665. IJCAI, pages 4950-4957. ijcai.org.
- [46] Bohan Zhai, Str.lia Yang, Xiangchen Zhao, Chen- [57] Hugo Touvron, Louis Martin, and et a].2023. feng Xu, Sheng Shen, Dongdi Zhao, Kurt Keutzer, Llama2: Open foundation and fine-tuned chat ManjingLi, TanYan, andXiangunFan.2023.Halle- models. CoRR, abs/2307,09288.
- [47] Hanning Zhang, Shizhe Diao, and et al. 2OZ3a. R- tuning: Teaching large language models to [58] Neeraj Varshney, Wenlin Yao, and et air.2023. refuse unknown questions. CoRR, abs/2311.0967 j . A stitch in time saves nine: Detecting and mitigating hallucinations of llms by validating low- confidence genera- tion. CoRR, abs/2307.03987.
- [48] Muru Zhang, Ofir Press, and et al. 2OZ3b. How [59] Tu Vu, Mohit Iyyer, and et al. 2023. Freshllms: lan- guage model hallucinations can snowball . Re- freshing large language models with search engine augmentation . arXiv p rep r int arXiv : CoRR, abs/2305.13534.
- [49] Yue Zhang, Yafu Li, and et al. 2023c. Siren,s song in the AI ocean: A survey on hallucination in large language models. CoRR, abs/230901219.
- [50] CeZheng, Lei Li, and et al. 2023. Can we edit factual knowledge by in-context learning? In EMNLP, pages 48624876.

Survey	Date	Pages	Eval	Improve	Multimodal	<u>Contributions and limitations</u>
[MISSING]	[MISSING]	/	/	/	Discusses ambiguous concepts in LLM factuality, compaes and analyzes evaluation and enhancement approach from academic and practical perspectives, outlining major challenges and promising avenues to explore. (Tonmoy et al., 2024)	Our work 15-June-2024
49	/	/	[MISSING]	Summrizes 08-lan-	recent work in tems of niri-gating LLM hatiu-cina-tioDs, bur lacks com. parison be-tween differ-ent ap-proaches and discus-sions to identify	44 2024

Type	Dataset (Reference)	Topic	Size	ER	Factscore-Bio (Min and et al., 2023)
549	42.6	Human annotation md automated fact-checkers	4 I	I	Factcheck- Open-ended questions GPT (Waet al., 2023c)
64.9	Hummm annotation	1 I	FacTool- QA (Chern et a.,2023)	Knowledge50 based QA	Knowledge50 based QA
Human mono- tation md auto- mated factcheck- ers	2 I	FE bV - W K (Cher et. il., 2023)	Knowledge184 based QA	46.2	
1 I	HaluEval (Li and, et al.,2023a)	Open-ended questions	5000	t2.3	Human annotation, AUROC + LLMjudg PARENT
FreshQA	Open-ended questions (Y! et a1.,2023)	499	ut:	Human annotation	2 I
Open- ended ques- tions	3369	[MISSING]	Evaluate the LLM awae- ness of un- known bv Fl- score	1 II	Snowball (Zhmg et al., 2023b)
1500	9,1	Exact match Accuracy,/Fl-score	+ [MISSING]Wki- cateBory	Multispan QA	
[MISSING]Precision/recall@5		[MISSING] III	[MISSING] III	(Dhuliawala et al., some 2023) [Mex- icon flms]	428
- Ex- act match + F1 score	[MISSING] III	(Dhuliawala et al., 2023)	Shon'tem Answer	[MISSING][MISSING]	
[MISSING]TruthfulQA	(Lin et al., 2022)	False belief or miscon- ception	817	[MISSING]	Accuracy
IV	HotpotQA	Multi-step reasoning (Yang and et al., 2018)	113k	[MISSING]- Exact match + F1 score	11 IV
Multi- step rea- soning	2780	[MISSING]	15	- Re- caU@10	MMLU (Hendrycks et al.,2021)