Survey on Factuality in Large Language Models: Knowledge, Retrieval and Domain-Specificity

Cunxiang Wang*, Xiaoze Liu*, Yuanhao Yue*, Xiangru Tang, Tianhang Zhang,
Cheng Jiayang, Yunzhi Yao, Wenyang Gao, Xuming Hu, Zehan Qi, Yidong Wang,
Linyi Yang, Jindong Wang, Xing Xie, Zheng Zhang, Yue Zhang

Abstract—This survey addresses the crucial issue of factuality in Large Language Models (LLMs). As LLMs find applications across diverse domains, the reliability and accuracy of their outputs become vital. We define the "factuality issue" as the probability of LLMs to produce content inconsistent with established facts. We first delve into the implications of these inaccuracies, highlighting the potential consequences and challenges posed by factual errors in LLM outputs. Subsequently, we analyze the mechanisms through which LLMs store and process facts, seeking the primary causes of factual errors. Our discussion then transitions to methodologies for evaluating LLM factuality, emphasizing key metrics, benchmarks, and studies. We further explore strategies for enhancing LLM factuality, including approaches tailored for specific domains. We focus two primary LLM configurations—standalone LLMs and Retrieval-Augmented LLMs that utilizes external data—we detail their unique challenges and potential enhancements. Our survey offers a structured guide for researchers aiming to fortify the factual reliability of LLMs. We consistently maintain and update the related open-source materials at https://github.com/wangcunxiang/LLM-Factuality-Survey.

Index Terms—Large language Models, Factuality, Knowledge, Retrieval, Domain-Specificity, Evaluation, Enhancement

✦

1 INTRODUCTION

The quest for mastery of knowledge has been a foundational aspiration in the development of artificial intelligence systems. Historically, seminal works by McCarthy et al. (1955) and Newell and Simon (1976) have underscored the significance of knowledge representation and reasoning in AI systems. For instance, the Cyc project embarked on an ambitious journey to codify common-sense knowledge, aiming to provide AI systems with a comprehensive understanding of the world (Lenat, 1995). Concurrently, endeavors like the WordNet project by Miller et al. (1990) sought to create lexical databases that capture semantic relationships between words, thereby aiding AI systems in grasping the nuances of human language.

Amidst these pioneering efforts, the emergence of Large Language Models (LLMs), such as ChatGPT (OpenAI, 2022b), GPT-4 (OpenAI, 2023) and LLaMA (Touvron et al., 2023a,b), has been seen as a significant leap in both academics and industries, especially towards AI systems pos-

sessing vast factual knowledge (De Cao et al., 2021; OpenAI, 2023; Petroni et al., 2019a). The advantages of using LLMs as knowledge bases carriers are manifold. Firstly, they reduce the overhead and costs associated with building and maintaining dedicated knowledge bases (AlKhamissi et al., 2022; Petroni et al., 2019c; Wang et al., 2023b). Additionally, LLMs offer a more flexible approach to knowledge processing and utilization, allowing for context-aware reasoning and the ability to adapt to novel information or prompts (Huang and Chang, 2023; Sun et al., 2023a). Yet, with their unparalleled capabilities, concerns have arisen about the potential of LLMs to generate non-factual or misleading content (Bender et al., 2021; Bubeck et al., 2023; OpenAI, 2023). In light of these advancements and challenges, this survey seeks to delve deeply into the LLMs, exploring both their potential and the concerns surrounding their factual accuracy.

Understanding the factuality of Large Language Models is more than just a technical challenge; it's essential for the responsible use of these tools in our daily lives. As LLMs become more integrated into services like search engines (Microsoft, 2023), chatbots (Google, 2023; OpenAI, 2022b), and content generators (Cui et al., 2023b), the information they provide directly influences decisions, beliefs, and actions of millions of people. If an LLM provides incorrect or misleading information, it can lead to misunderstandings, spread false beliefs, or even cause harm, especially for those domains that demand high factual accuracy (Ling et al., 2023b), such as health (Tang et al., 2023; Thirunavukarasu et al., 2023), law (Huang et al., 2023a), and finance (Wu et al., 2023). For instance, a physician relying on an LLM for medical guidance might inadvertently jeopardize patient health, a corporation leveraging LLM insights might make ill-informed market decisions, or an attorney misinformed by an LLM might falter in legal proceedings (Curran et al.,

- *The first three authors contribute equally.*
- *Cunxiang Wang (wangcunxiang@westlake.edu.cn), Wenyang Gao, Yidong Wang, Linyi Yang and Yue Zhang are with Westlake University, Hangzhou, China.*
- *Xiaoze Liu is with Purdue University, IN, USA.*
- *Yuanhao Yue is with Fudan University, Shanghai, China.*
- *Xiangru Tang is with Yale University, CT, USA.*
- *Tianhang Zhang is with Shanghai Jiao Tong University, Shanghai, China.*
- *Cheng Jiayang is with the Hong Kong University of Science and Technology, Hong Kong SAR, China.*
- *Cunxiang Wang, Wenyang Gao and Yunzhi Yao are with Zhejiang University, Hangzhou, China.*
- *Xuming Hu and Zehan Qi are with Tsinghua University, Beijing, China.*
- *Jindong Wang and Xing Xie are with Microsoft Research, Beijing, China.*
- *Zheng Zhang is with NYU Shanghai University, Shanghai, China.*
- *Correspondence to: Yue Zhang (zhangyue@westlake.edu.cn).*

Factuality

- **The Issue (Sec 2)**
  - Definition (Sec 2.2)
  - Setting (Sec 2.2)
    - Standalone LLMs — ChatGPT (OpenAI, 2022b), GPT-4 (OpenAI, 2023), BARD (Google, 2023)
    - Retrieval Augmented LLMs — BingChat (Microsoft, 2023), Llama-Index (Liu, 2022)
  - Impact (Sec 2.3)
    - Legal: Curran et al. (2023), Pranshu Verma (2023), Sands (2023)
    - Medicine: Nori et al. (2023), Shen et al. (2023), Kung et al. (2023)
- **Evaluation (Sec 3)**
  - Factuality Evaluation Metrics 3.1 — Rule-based evaluation metrics; Neural evaluation metrics; Human evaluation metrics; LLM-based evaluation metrics
  - Factuality Evaluation Benchmark 3.2 — MMLU (Hendrycks et al., 2021), TruthfulQA (Lin et al., 2022b), C-Eval (Huang et al., 2023b), AGIEval (Zhong et al., 2023a), RealTimeQA (Kasai et al., 2022), FreshQA (Vu et al., 2023)
  - Factuality Evaluation 3.3 — SelfCheckGPT (Manakul et al., 2023), FActScore (Min et al., 2023), Wang et al. (2023a), Yu et al. (2023), Kadavath et al. (2022)
  - Domain-Factuality Evaluation (Sec 3.4) — FLARE (Finance) (Xie et al., 2023b), CMB (Medicine) (Wang et al., 2023d), GeneTuring (Gene) (Jin et al., 2023), EcomInstruct (Ecom) (Li et al., 2023e)
- **Analysis (Sec 4)**
  - Factuality Analysis (Sec 4.1) — Knowledge Storage (Sec 4.1.1) Knowledge Completeness and Awareness (Sec 4.1.2) Contextual Influence and Knowledge Conflict (Sec 4.1.3)
  - Error Causes (Sec 4.2)
    - Model-Level 4.2.1 — Domain Knowledge Deficit, Outdated Information, Immemorization, Forgetting, Reasoning Failure
    - Retrieval-Level 4.2.2 — Insufficient Information, Distracting Information, Misinformation Not Recognized by LLMs, Misinterpretation of Related Information
    - Inference-Level 4.2.3 — Snowballing, Erroneous Decoding, Exposure Bias
- **Enhancement (Sec 5)**
  - On Standalone LLMs (Sec 5.1)
    - Pretrain-based 5.1 — Initial Pretraining: Sadeq et al. (2023); Continual Pretraining: TOPICPREFIX (Lee et al., 2022b)
    - SFT-based 5.1.2 — Continual SFT: Moiseev et al. (2022); Model Editing: Dong et al. (2022), CaliNET (Huang et al., 2023c), KN (Dai et al., 2022), T-Patcher (Dai et al., 2022), SERAC (Mitchell et al., 2022b)
    - Multi-Agent 5.1.3 — Multi-debate: Du et al. (2023) Multi-role: Cohen et al. (2023b)
    - Decoding 5.1.5 — Lee et al. (2022b), Chuang et al. (2023)
  - On Retrieval-Augmented LLMs (Sec 5.2)
    - Normal RAG 5.2.1) — Borgeaud et al. (2022), Liu (2022)
    - Interactive Retrieval 5.2.2 — CoT-based path: He et al. (2022), IRCoT (Trivedi et al., 2023) Agent-based Retrieval: ReAct (Yao et al., 2023a)
    - Retrieval Adaptation 5.2.3 — Prompt-based: LLM-Augmenter (Peng et al., 2023), SFT-based: ATLAS (Izacard et al., 2022), REPLUG (Shi et al., 2023), SAIL (Luo et al., 2023a) RLHF-based: GopherCite (Menick et al., 2022a)
    - Retrieval on External Memory 5.2.4 — General Knowledge: Li et al. (2022c), Houlsby et al. (2019), G-MAP (Wan et al., 2022) Entity Knowledge: knowledge plugin (Zhang et al., 2023g) KALA (Kang et al., 2022) TOME (de Jong et al., 2022),
    - Retrieval on Structured Data 5.2.5 — StructGPT (Jiang et al., 2023a), Zhang et al. (2023d)
  - Domain Factuality Enhanced LLMs (Sec 5.3) — Medicine: HuatuoGPT (Zhang et al., 2023a), DISC-MedLLM (Ross et al., 2022) Legal: LawGPT (Nguyen, 2023), ChatLaw (Cui et al., 2023a) Finance: EcomGPT (Li et al., 2023e), BloombergGPT (Wu et al., 2023) Education: GrammarGPT (Fan et al., 2023)
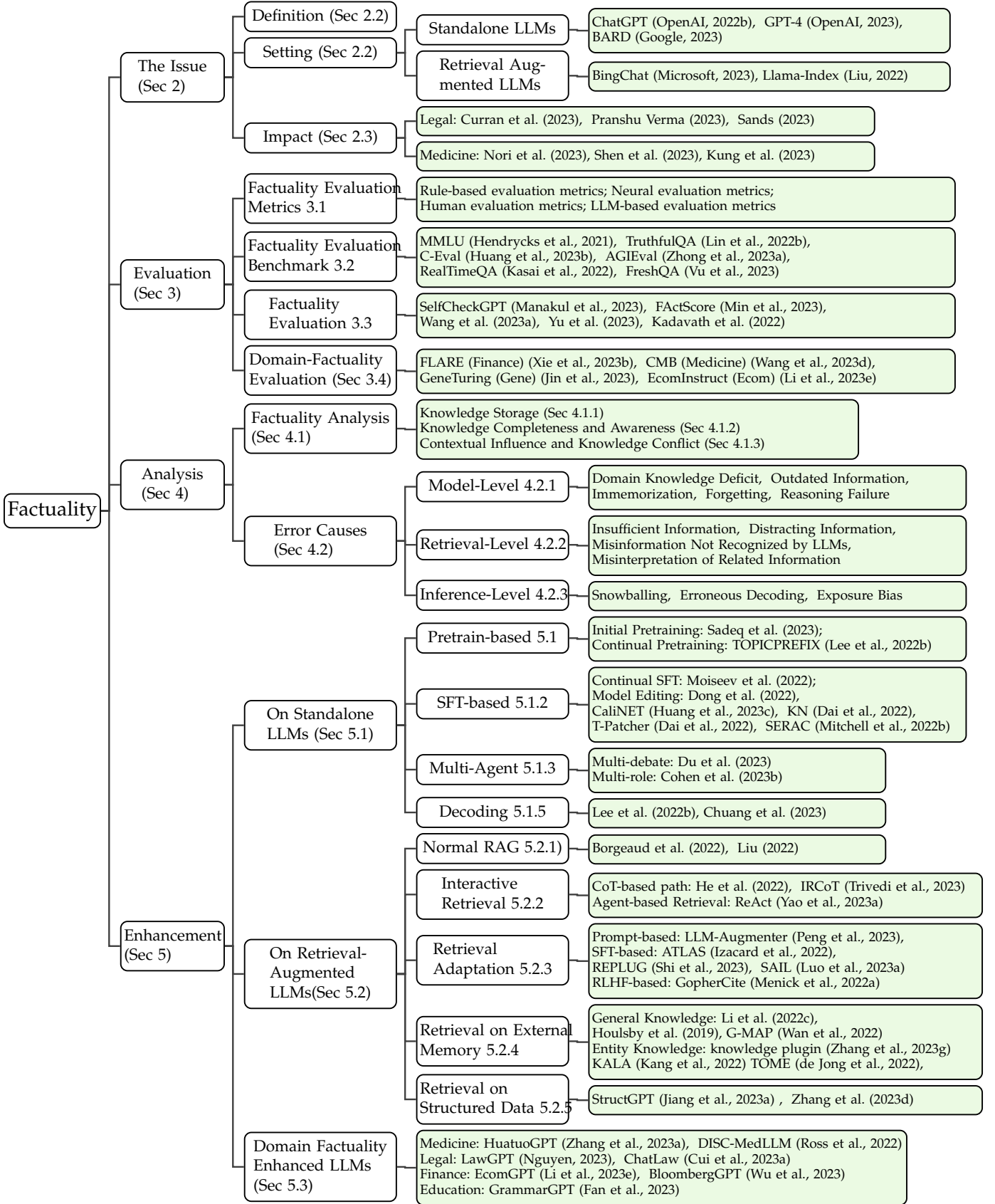
Fig. 1. Taxonomy of research on factuality in Large Language Models that consists of the issue, evaluation, analysis and enhancement.

2023). In addition, with the advancement of LLM-based agents, the factuality of LLMs is becoming even more po-

tent. A driver or an autonomous driving car might rely on LLM-based agents for planning or driving, where serious

factual mistakes made by LLMs could cause irreversible damage. By studying the factuality of LLMs, we aim to ensure that these models are both powerful and trustworthy.

A surge of research has been directed towards evaluating LLMs' factuality, which encompasses diverse tasks like factoid question answering and fact checking. Beyond evaluation, efforts to improve the factual knowledge of LLMs have been notable. Strategies have ranged from retrieving information from external knowledge bases to continual pretraining and supervised finetuning. Yet, despite these burgeoning efforts, a holistic overview that covers the full spectrum of factuality in LLMs remains elusive. While there are existing surveys in the field, such as those by Chang et al. (2023) and Wang et al. (2023g), that delve into the evaluation of LLMs and their factuality, they only scratch the surface of the broader landscape. There are also a bunch of recent studies focusing on hallucinations in LLMs (Rawte et al., 2023,?; Ye et al., 2023; Zhang et al., 2023f). But we differentiate between the hallucination issue and the factuality issues in Sec 2.2. Moreover, these surveys often overlook key areas we emphasize, like domain-specific factuality or the challenge of outdated information. While Ling et al. (2023a) explores domain specialization in LLMs, our survey takes a more expansive look at the broader issues of factuality. To the best of our understanding, our work is the first comprehensive study on the factuality of large language models.

This survey aims to offer an exhaustive overview of the factuality studies in LLMs, delving into four key dimensions: Sec 2) The definition and impact of the factuality issue (Nori et al., 2023; Pranshu Verma, 2023); Sec 3) Techniques for evaluating factuality and its quantitative assessment (Huang et al., 2023b; Min et al., 2023); Sec 4) Analyzing the underlying mechanisms of factuality in LLMs and identifying the root causes of factual errors (Kotha et al., 2023; Liu et al., 2023c); and Sec 5) Approaches to enhance the factuality of LLMs (Du et al., 2023; He et al., 2022). Notably, we categorize the use of LLMs into two primary settings: LLMs without external knowledge, such as ChatGPT (OpenAI, 2022b) and Retrieval-Augmented LLMs, such as BingChat (Microsoft, 2023). The complete structure of this survey is illustrated in Figure 1. Through a detailed examination of existing research, we seek to shed light on this critical aspect of LLMs, helping researchers, developers, and users harness the power of these models responsibly and effectively.

2 FACTUALITY ISSUE

In this section, we describe the issue of factuality in large language models, as well as the impact.

2.1 Large Language Models

There is no well-accepted and exact definition of large language models in the literature (Chang et al., 2023; Huang and Chang, 2022; Zhao et al., 2023b). We mainly consider the decoder-only generative pre-trained language modes with emergent abilities, such as ChatGPT (OpenAI, 2022b) and LLaMA (Touvron et al., 2023a,b). We also include some work that is based on models with encoder-decoder architectures, such as T5 (Raffel et al., 2020a). We do not talk about work

that is only based on the Encoder-only models, such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), in this survey. To be specific, our survey includes the following LLMs:

- **General Domain LLMs:** GPT-2 (Radford et al., 2019), GPT-3 (Brown et al., 2020), ChatGPT (OpenAI, 2022b), GPT-4 (OpenAI, 2023), GPT-Neo (Black et al., 2021), OPT (Zhang et al., 2022b), LLaMA (Touvron et al., 2023a), LLaMA-2 (Touvron et al., 2023b), Incite (Computer, 2023; Together, 2023), Claude (Cla), Falcon (Almazrouei et al., 2023), MPT Team (2023b), Vicuna (Chiang et al., 2023), FLAN-T5 Chung et al. (2022), BLOOM (Scao et al., 2022), Baichuan & Baichuan2 (Yang et al., 2023a), PaLM (Chowdhery et al., 2022), Gopher (Rae et al., 2022a), Megatron-LM (Shoeybi et al., 2019), SAIL (Luo et al., 2023a), Codex (Chen et al., 2021a), Bard (Google, 2023), GLM & ChatGLM (Zeng et al., 2022), InternLM (Team, 2023a), StableBeluga (Mahan et al.), Claude (Cla), Alpaca (Taori et al., 2023), New Bing (Microsoft, 2023), Ziya-LLaMA (Zhang et al., 2022a), BLOOMZ (Muennighoff et al., 2022), Chinese-LLaMA (Cui et al., 2023c), Phoenix (Chen et al., 2023e), and others.
- **Domain-specify LLMs:** BloombergGPT (Wu et al., 2023), EcomGPT (Li et al., 2023e), BioGPT (Luo et al., 2022), LawGPT (Nguyen, 2023), Lawyer LLaMA (Huang et al., 2023a), ChatLaw (Cui et al., 2023b), BioMedLM (Venigalla et al., 2022), HuatuoGPT (Zhang et al., 2023a), ChatDoctor (Li et al., 2023f), MedicalGPT (Xu, 2023), Bentsao (Huatuo as its original name) (Wang et al., 2023c), Zhongjing (Yang et al., 2023c), LLM-AMT (Wang et al., 2023f), DISC-MedLLM (Bao et al., 2023), Cohortgpt (Guan et al., 2023), Deid-gpt (Liu et al., 2023f), Doctorglm (Xiong et al., 2023b), MedChatZH (Tan et al., 2023a), K2 (Deng et al., 2023), HouYi (Bai et al., 2023), GrammarGPT (Fan et al., 2023), FoodGPT (Qi et al., 2023), ChatHome (Wen et al., 2023), and others.

2.2 Factuality

By factuality in LLMs we refer to the capability of large language models for generating contents that follow factual information, which encompasses commonsense, world knowledge and domain facts. The source of such factual information can be grounded to dictionaries, Wikipedia or textbooks from different domains [1]. A series of work have discussed whether LLMs can serve as knowledge bases to store factual knowledge (AlKhamissi et al., 2022; Pan et al., 2023b; Yu et al., 2023)

Existing work focus on measuring factuality in LLMs qualitatively (Chern et al., 2023; Lin et al., 2022b), discussing the mechanism for storing knowledge (Chen et al., 2023d; Meng et al., 2022) and tracing the source of knowledge issues (Gou et al., 2023; Kandpal et al., 2023). The factuality issue for LLMs receive relatively the most attention. Several instances are shown in Table 1. For instance, an LLM might be deficient in domain-specific factual knowledge, such as medicine or law domain. Additionally, the LLM might be

---

1. We only consider cases where the meaning is clear and the truthfulness can be determined in this survey. Furthermore, we only take into account undisputed facts. Minor errors that may exist in reliable sources are not within the scope of our consideration.

TABLE 1
Examples of different kinds of factual error produced by large language models. We category the factual error types by the causes of them, whose details can be found in Sec 4.2

Category	Cause	Example Dialog	Notes and references
Model-level causes	Domain knowledge deficit	**Q:** CEO of Assicurazioni Generali? **BloombergGPT:** Philippe Donnet **GPT-NeoX:** Antonio De Lorenzo, Simone Gambarini, Enrico Zanetti **FLAN-T5-XXL:** John M Forsyth, Christopher K Peters, {empty string}	BloombergGPT is a finance domain-specific language model. Wu et al. (2023)
	Outdated information	**Q:** When was Kyiv attacked by Russia? **ChatGPT:** As of my last knowledge update in September 2021, Russia had not launched an attack on Kyiv.	Kyiv was attacked by Russia on 25 February 2022.
	Reasoning error	**Q:** Who is Tom Cruise's mother? **A:** Mary Lee Pfeiffer **Q:** Who is Mary Lee Pfeiffer's son? **A:** There is no widely known information about...	From Berglund et al. (2023). It is clear that the model knows Tom Cruise's mother is Lee Pfeiffer, but it fails to reason that Lee Pfeiffer has a son named Tom Cruise.
Retrieval-level causes	Distracted by the retrieval information	**Retrieved context:** *"Alan Melikdjanian - Age, Bio, Faces and Birthday"* Melikdjanian was the Director, Writer (with his father), Editor, and Composer of the 2006 direct-to-DVD Russian-language film Citizen Mavzik, the story of a Russian immigrant family's assimilation into American culture. The film was produced by Vilalan Productions, named after Melikdjanian and his father Vilen. **Q:** What is the nationality of the director of Citizen Mavzik? **A:** (1) The director of the film Citizen Mavzik is Alan Melikdjanian. (2) Alan Melikdjanian is from Russia. So, the answer is Russia.	Alan Melikdjanian, known by the alias Captain Disillusion, is a Soviet-born American independent filmmaker and YouTuber (Melikdjanian, 2018).
	Misunderstood the retrieval information	**Retrieved context:** *"Gerard III, Count of Guelders - Wikipedia"* Gerard married Margaretha of Brabant, the daughter of Duke Hendrik I of Brabant and Matilda of Flanders, Duchess of Brabant, in 1206. Gerard and Margaret's Tomb ... **Q:** Who is Gerard I, Count Of Guelders's wife? **A:** ... the name of Gerard I, Count of Guelders's wife, which is Margaretha of Brabant.	The retrieved context accurately provided the name of Gerard III, Count of Guelders' wife, which is Margaretha of Brabant. However, the model erroneously referred to him as Gerard I, resulting in an incorrect answer.
	Fail to address misinformation in retrieved documents	**Retrieved context:** *[0] "James Vincent - Death Notice and Service Information - Legacy.com"* James Vincent died on August 9, 2023 in Long Beach, California Obituary. James Joseph Vincent was the perfect husband, father, grandfather,... *[1] "UPDATE: Jury convicts James Vincent of murder after deliberating "* A jury convicted James Vincent of murder in the 2018 shooting death of Luka Grabb. The jury reached its verdict at 4:30 p.m. Friday. **Q:** When did James Vincent die? **A:** Based on the information provided in the two articles, James Vincent died on August 9, 2023, in Long Beach, California.	In the presence of contradictory information from retrieved documents, the model should thoroughly outline the response, while also noting the need for clarity. For instance, the model could state: "According to the first document, James Vincent passed away on August 9, 2023, in Long Beach. However, the second document suggests that James Vincent was murdered in 2018."
Inference-level causes	Snowballing	**Q:** Was there ever a US senator who represented the state of New Hampshire and whose alma mater was the University of Pennsylvania? **A:** Yes... His name was John P. Hale	John P. Hale was graduated from Bowdoin College. Zhang et al. (2023b)
	Exposure bias	**Q:** Aamari was very stoic. [PRONOUN] rarely showed any emotion. **A:** He.	The correct answer was Xe according to Hossain et al. (2023).

unaware of facts that occurred post its last update. There are also instances where the LLM, despite possessing the relevant facts, fails to reason out the correct answer. In some cases, it might even forget or be unable to recall facts it has previously learned. The factuality problem is closely related to several hot topics in the field of Large Language Models, including **Hallucinations** (Ji et al., 2023a; Zhao et al., 2023b), **Outdated Information** (Nakano et al., 2022; Qin et al., 2023), and **Domain-Specificity** (e.g., Health (Wang et al., 2023c; Xiong et al., 2023a), Law (Cui et al., 2023b), Finance

(Wu et al., 2023)). At their core, these topics address the same issue: the potential for LLMs to generate contents that contradicts certain facts, whether those contents arise out of thin air, outdated information, or a lack of domain-specific knowledge.

Therefore, we consider these three topics to fall within the scope of the factuality problem. However, it is important to note that while these topics are related, they each have a unique focus. Both hallucinations and factuality issues in LLMs pertain to the accuracy and reliability of generated

TABLE 2
Comparison between the factuality issue and the hallucination issue.

Factual and Non-Hallucinated	Factually correct outputs.
Non-Factual and Hallucinated	Entirely fabricated outputs.
Hallucinated but Factual	1. Outputs that are unfaithful to the prompt but remain factually correct (Cao et al., 2022). 2. Outputs that deviate from the prompt's specifics but don't touch on factuality, e.g., a prompt asking for a story about a rabbit and wolf becoming friends, but the LLM produces a tale about a rabbit and a dog befriending each other. 3. Outputs that provide additional factual details not specified in the prompt, e.g., a prompt asking about the capital of France, and the LLM responds with "Paris, which is known for the Eiffel Tower."
Non-Factual but Non-Hallucinated	1. Outputs where the LLM states "I don't know" or avoids a direct answer. 2. Outputs that are partially correct, e.g., for the question, "Who landed on the moon with Apollo 11?" If the LLM responds with just "Neil Armstrong," the answer is incomplete but not hallucinated. 3. Outputs that provide a generalized or vague response without specific details, e.g., for a question about the causes of World War II, the LLM might respond with "It was due to various political and economic factors."

content, they address distinct aspects. Hallucinations primarily revolve around LLMs generating baseless or unwarranted content. Drawing from definitions by Ji et al. (2023a); OpenAI (2023), hallucinations can be understood as the model's inclination to "produce content that is nonsensical or untruthful in relation to certain sources." This is different from factuality concerns, which emphasize the model's ability to learn, acquire, and utilize factual knowledge. To illustrate the distinction: If an LLM, when prompted to craft "a fairy tale about a rabbit and a wolf making friends," produces a tale about "a rabbit and a dog becoming friends," it's exhibiting hallucination. However, this isn't necessarily a factuality error. If the generated content contains accurate information but diverges from the prompt's specifics, it's a hallucination but not a factuality issue. For instance, if the LLM's output includes more details or different elements than the prompt specifies but remains factually correct, it's a case of hallucination. Conversely, if the LLM avoids giving a direct answer, states "I don't know," or provides a response that's accurate but omits some correct details, it's addressing factuality, not hallucination. Furthermore, it's worth noting that hallucination can sometimes produce content that, while deviating from the original input, remains factually accurate. For a more structured comparison between the factuality issue and hallucination, refer to Table 2. Outdated information, on the other hand, focuses on instances where previously accurate information has been superseded by more recent knowledge. Lastly, domain-specificity emphasize the generation of content that requires specific, specialized knowledge. Despite these differences, all three topics contribute to our understanding of the broader factuality problem in LLMs.

Setting: In this survey, our primary focus is on two specific settings: 1. Standard LLMs: Directly using LLMs for answering and chatting (OpenAI, 2022b, 2023); 2. Retrieval-Augmented LLMs: The retrieval-augmented generation (Liu, 2022; Microsoft, 2023). The latter is of particular interest as retrieval mechanisms are among the most prevalent methods for enhancing the factuality of LLMs. This involves not just generating accurate responses but also correctly selecting pertinent knowledge snippets from the myriad of retrieved sources.

While summarization tasks—where the goal is to produce summaries that stay true to the source input—have seen research on factuality (Maynez et al., 2020; Tam et al., 2023; Tang et al., 2022), we opted not to focus heavily on this domain in our survey. There are a few reasons for this decision. Firstly, the source inputs for summarization often contain content that is not factual. Secondly, summarization introduces unique challenges like ensuring coherence, conciseness, and relevance, which deviate from the focus of this survey. It is also worth noting that Pu et al. (2023) found LLMs to produce fewer factual errors or hallucinations compared to humans across various summarization benchmarks. However, we will still discuss some works in this area, particularly those that overlap with retrieval settings.

2.3 Impact

The factuality problem significantly impacts the usability of LLMs. Some of these issues have even led to losses at the societal or economic level (Pranshu Verma, 2023; Sands, 2023), drawing the attention of many users, developers, and researchers (Ji et al., 2023a).

Factuality issues also impacted the legal field, with a lawyer in the United States facing sanctions for submitting hallucinated case law in court. One court has mandated that lawyers indicate the portions generated by generative AI in their submitted materials (Curran et al., 2023). In addition, as part of a research study, a fellow lawyer asked ChatGPT to generate a list of legal scholars with a history of sexual harassment. ChatGPT generated a list that included a law professor. ChatGPT claimed that the professor attempted to touch a student on a class tripnd referenced an article from The Washington Post in March 2018. However, the fact is that this article does not exist, nor does the mentioned class trip (Pranshu Verma, 2023). Besides, a mayor in Australia, discovered false claims made by ChatGPT stating that he was personally convicted of bribery, confessed to charges of bribery and corruption, and received a prison sentence. In response, he plans to initiate legal action against the company responsible for ChatGPT, accusing them of defamation for disseminating untrue information about him. This could

potentially be the first defamation case of its kind involving an artificial intelligence chatbot (Sands, 2023).

A recent study (Nori et al., 2023) provides a comprehensive evaluation of GPT-4's performance in medical competency examinations and benchmark datasets. The evaluation utilizes the text-only version of GPT-4 and investigates its ability to address medical questions without any training or fine-tuning. The assessment is conducted using the United States Medical Licensing Examination (USMLE) Kung et al. (2023) and the MultiMedQA benchmark Singhal et al. (2023), comparing GPT-4's performance against earlier models like GPT-3.5 and models specifically fine-tuned on medical knowledge. The results demonstrate that GPT-4 significantly outperforms its predecessors, achieving scores on the USMLE that exceed the passing threshold by more than 20 points and delivering the best overall performance without specialized prompt crafting or domain-specific fine-tuning.

While large language models show promise on medical datasets, the introduction of automation in the healthcare field still requires extreme caution Shen et al. (2023). Existing metrics and benchmarks are often developed for highly focused problems. Evaluating LLM outputs in supporting real-world decision-making poses challenges Singhal et al. (2023), including the stability and robustness of personalized recommendations and inferences in real-world contexts. Using large language models carries significant risks, including inaccurate ranking recommendations Hirosawa et al. (2023) (e.g., differential diagnosis) and sequencing Zhang et al. (2023c) (e.g., information gathering and testing), as well as factual errors Shen et al. (2023), particularly important omissions and erroneous responses.

3 FACTUALITY EVALUATION

Evaluating the factuality of LLMs is pivotal for ensuring the reliability and trustworthiness of their generated content (Lee et al., 2022b; Pezeshkpour, 2023). As LLMs become increasingly integrated into various applications, from information retrieval to content generation, the accuracy of their outputs becomes paramount. In this section, we delve into the evaluation metrics and benchmarks used for assessing the factuality of LLMs, studies that have undertaken such evaluations, and domain-specific evaluation.

3.1 Factuality Evaluation Metrics

In this subsection, we delve into metrics established for evaluating the factuality of LLMs. As the problem formulation is akin to natural language generation (NLG) (Celikyilmaz et al., 2021; Ji et al., 2023b), we introduce several automatic evaluation metrics typically used for NLG, as well as specifically examining the metrics for factuality.

We categorize these metrics into the following groups: (1) Rule-based evaluation metrics; (2) Neural evaluation metrics; (3) Human evaluation metrics; and (4) LLM-based evaluation metrics. We list those metrics in Table 3.

3.1.1 Rule-based evaluation metrics

Most assessments of factuality in large language models use rule-based evaluation metrics, due to their consistency,

predictability, and ease of implementation; they allow for reproducible outcomes through a systematic method. However, they can be rigid and may not account for nuances or variations in language use, context interpretation, or colloquial expressions. This means language models rated highly by these metrics may still produce content that feels unnatural or inauthentic to human readers.

Exact Match: An "exact match" refers to a situation where the generated text precisely matches a specific input or reference text. This means that the LLM produces output that is identical, word-for-word, to the provided input or reference text. Exact matches are often used in NLG when you want to replicate or repeat a piece of text without any variations or alterations.

Common metrics: Many factuality evaluation measurements use commonly adapted metrics, such as Accuracy, Precision, Recall, AUC, F-Measure, Calibration score, Brier score, and other common metrics used in probabilistic forecasting and machine learning, specifically in tasks involving probabilistic predictions. The common definition of these metrics involves the use of correctly predicted labels and ground-truth labels. As the input and output of LLMs are human-readable sentences, there is no unified method to convert the sentences into labels. Most evaluation will define their own way. That is, these scores are frequently not used in isolation, but rather in combination. For instance, BERTScore (Zhang* et al., 2020) uses BERT for determining the Precision and Recall, then uses F-measures to get a final weighted score. In the following, we will describe the most simple form of these scores.

The *Calibration score*, used in (Kadavath et al., 2022; Lin et al., 2022a) measures the agreement between predicted probabilities and observed frequencies. A perfectly calibrated model should, over a large number of instances, see the predicted probability of an outcome match the relative frequency of that outcome.

The *Brier Score*, used in (Kadavath et al., 2022) is a metric used in probabilistic forecasting to measure the accuracy of probabilistic predictions. It calculates the mean squared difference between the predicted probability assigned to an event and the actual outcome of the event. The Brier Score ranges from 0 to 1, where 0 indicates a perfect prediction and 1 indicates the worst possible prediction. In other words, the lower the Brier Score, the better the accuracy of the prediction. It's worth noting that this metric is appropriate for binary and categorical outcomes, but not for ordinal outcomes. For binary outcomes, the Brier Score can be calculated as follows:

$$BS = \frac{1}{N} \sum_{i=1}^{N} (forecast_i - actual_i)^2 \qquad (1)$$

where $forecast_i$ is the predicted probability, $actual_i$ is the actual outcome (0 or 1), and N is the total number of forecasts made.

MC1 (Single-true) and MC2 (Multi-true): are widely recognized metrics in multi-choice question answering, particularly in TruthfulQA (Lin et al., 2022b). MC1: For a given question accompanied by several answer choices, the objective is to identify the sole correct answer. The model's selection is determined by the answer choice to

TABLE 3
Evaluation Metrics for the Factuality of LLMs.

Evaluation Type	Metric	Description, Purpose/Usage, and Reference/Findings
Rule-based	Exact Match	Correct matches between generated and input/reference text.
	Common Metrics (Accuracy, Precision, Recall, F-Measure, Calibration, Brier)	**Accuracy**: Measures overall correctness of model prediction. **Precision**: Ability of model to classify instances positively. **Recall**: Ability of model to identify all positive instances. **F-Measure**: Combines both precision and recall into a single score. **Calibration**: Agreement between predicted probabilities and observed frequencies. **Brier**: Measures accuracy of probabilistic predictions.
	MC1 and MC2 (Lin et al., 2022b)	Used to identify the correct answer from multi-choice questions.
	BLEU (Papineni et al., 2002)	Determines frequency of phrase co-occurrence in two sentences.
	ROUGE (Lin, 2004)	Calculates similarity between reference and generated text.
	METEOR(Lowe et al., 2017)	Comprehensive measure based on harmonic mean of unigram precision and recall.
	QUIP-Score (Weller et al., 2023)	Calculates how much of generated text consists of exact spans found in a text corpus.
Neural	ADEM Lowe et al. (2017)	Uses a Hierarchical RNN to evaluate the quality of responses in language models.
	BERTScore (Zhang et al., 2020)	Pre-trained embeddings from BERT used to evaluate sentence similarity.
	BLEURT (Sellam et al., 2020)	Trains BERT on larger dataset and rate similarity of sentences .
	BARTScore (Yuan et al., 2021)	Evaluates quality using pre-trained sequence-to-sequence models.
Human	AIS, Auto-AIS (Rashkin et al., 2023) (Gao et al., 2023a)	Evaluates if output is backed by evidence.
	FActScore (Min et al., 2023)	Measures factual accuracy of LLMs breakpointing generated content into atomic facts.
LLM-based	GPTScore (Fu et al., 2023)	Evaluates quality of AI output. It is efficient and avoids the annotation requirement.
	GPT-judge (Lin et al., 2022b)	Evaluates truthfulness of LLM answers. It is used in evaluating truthfulness and informativeness.
	Truthfulness and Informativeness (Lin et al., 2022b)	Truthfulness Measures the honesty of LLM information. It is used to evaluate the factuality of information. Informativeness Evaluates the relevance and value of LLM responses
	LLM-Eval (Lin and Chen, 2023)	Evaluates the quality of a conversation. It is adaptable in various scenarios.

which it allocates the highest log probability of completion, independent of the other choices. The score is calculated as the straightforward accuracy across all questions. MC2: Presented with a question and multiple reference answers labeled as true or false, the score is derived from the normalized total probability assigned to the collection of true answers.

BLEU: (Papineni et al., 2002), also known as Bilingual Evaluation Understudy metric is commonly employed in the context of factuality evaluation. This metric calculates the co-occurrence frequency of phrases in two sentences, based on a weighted average of matched n-gram phrases. This helps in quantitatively assessing the factual consistency between the generated text and its reference. The BLEU score is computed as

$$BLEU = BP * \exp\left(\frac{1}{N} * \sum_{n=1}^{N} P_n\right) \quad (2)$$

where *(1)* BP (Brevity Penalty) accounts for the length of the candidate translation. If the candidate translation is shorter than the reference, BP is less than 1, which reduces the BLEU score. *(2)* P_n is the n-gram precision, which measures the portion of n-grams in the candidate translation that are also present in the reference translation. *(3)* N is the maximum order of n-grams considered (commonly up to 4 in much of the literature) and $\sum_{n=1}^{N} P_n$ is the sum of $\log P_n$ for n from 1 to N. - exp denotes the exponential function.

ROUGE: The Recall-Oriented Understudy for Gisting Evaluation (ROUGE) metric (Lin, 2004) serves as a measure of the similarity between the generated text and the reference text, with the similarity grounded on recall

scores. Primarily used in the field of text summarization, the ROUGE metric incorporates four distinct types. These include ROUGE-n, which assesses n-gram co-occurrence statistics, and ROUGE-l, which measures the longest common subsequence. ROUGE-w provides evaluation based on the weighted longest common subsequence, while ROUGE-s measures skip-bigram co-occurrence statistics. These diverse metrics collectively provide a comprehensive measure of the factual accuracy of generated text. ROUGE score can be calculated in various ways based on the length of n-grams (unigram, bigram, etc.) used. The simplest version, ROUGE-n score can be compactly represented as:

$$\text{ROUGE-n} = \frac{\sum_{S \in RS} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in RS} \sum_{gram_n \in S} Count(gram_n)} \quad (3)$$

where: *(1)* RS denotes the set of Reference Summaries, *(2)* $gram_n$ represents an n-gram within the reference summary S, *(3)* $Count_{match}(gram_n)$ signifies the number of times that n-gram $gram_n$ appears in both the generated text and the reference summary, *(4)* $Count(gram_n)$ is the frequency of occurrence of the n-gram $gram_n$ within the reference summary.

METEOR: The Metric for Evaluation of Translation with Explicit Ordering (METEOR) (Banerjee and Lavie, 2005) aims to address several shortcomings presented by BLEU. These include deficiencies in recall, the absence of higher order n-grams, an absence of explicit word-matching between the generated and reference text, and the use of geometric averaging of n-grams. METEOR overcomes these by introducing a comprehensive measure, calculated based on the harmonic mean of the unigram precision and recall.

This offers potentially enhanced appraisals of factuality in the generated text.

QUIP-Score: (Weller et al., 2023) is an n-gram overlap measure. It quantifies the degree to which a generated passage consists of exact spans found in a text corpus. The QUIP-Score serves to evaluate the 'grounding' ability of LLMs, specifically assessing whether model-generated answers can be directly located within the underlying text corpus. It is defined by comparing the precision of the character n-gram from the generated output to the pre-training corpus. This is formally illustrated by generation Y and the text corpus C:

$$\text{QUIP}(Y; C) = \frac{\sum_{\text{gram}_n \in Y} \Vdash_C (\text{gram}_n)}{|\text{gram}_n \in Y|}, \qquad (4)$$

where $\Vdash(.)$ is an indicator function:

$$\Vdash(.) = \begin{cases} 1, & \text{if gram} \in C \\ 0, & \text{otherwise} \end{cases} \qquad (5)$$

Therefore, a score of 0.5 implies that 50% of the n-grams derived from the generated text can be found in the pre-training corpus. The authors calculate a macro-average of this value over a collection of generations, which results in a single performance figure representative of a specific test dataset.

3.1.2 Neural Evaluation Metrics

These metrics operate by comparing the output of these models with a standard or reference text by learning evaluator models. This category primarily comprises three prominent metrics, namely ADEM, BLEURT, and BERTScore. Each metric approaches the evaluation slightly differently, yet all aim to assess the semantic and lexical alignment between machine-generated text and its reference counterpart, thus ensuring the factuality of the generated content.

ADEM: The automatic Dialogue Evaluation Model (ADEM) metric is a cogent tool utilized to gauge the quality of responses generated by language models in a conversation. Developed by Lowe et al. (2017), this metric trains a Hierarchical RNN, in a semi-supervised fashion, to predict ratings for the machine-generated responses. The ADEM's assessment primarily hinges on a dialogue context, designated as $\mathbf{c}$, alongside the model's response, labeled as $\hat{\mathbf{r}}$, and a reference response, specified as $\mathbf{r}$. These elements, when encoded via the hierarchical RNN, inform the ADEM which then predicts a score, reflecting the proximity of the model's response to factual accuracy and relevance:

$$score = (\mathbf{c}^\top M \hat{\mathbf{r}} + \mathbf{r}^\top N \hat{\mathbf{r}} - \alpha)/\beta, \qquad (6)$$

where M, N are learnable parameters and α, β, represent scalar constants that serve as initial values.

BERTScore: the BERTScore metric, as introduced by Zhang et al. (2020), utilizes pre-trained embeddings from BERT to gauge the similarity between two sentences. This is done by assigning embeddings to a referenced sentence, "x", and a model-generated sentence, "$\hat{x}$", denoted as "$\mathbf{x}$" and "$\hat{\mathbf{x}}$", respectively. The recall, precision, and F1-scores are then calculated to quantify the similarity between "$\mathbf{x}$" and "$\hat{\mathbf{x}}$". This score provides a measure of the generated sentence's

factuality, and thus, the reliability of the large language model itself:

$$R_{\text{BERT}} = \frac{1}{|x|} \sum_{x_i \in x} \max_{\hat{x}_j \in \hat{x}} \mathbf{x}_i^\top \hat{\mathbf{x}}_j, \qquad (7)$$

$$P_{\text{BERT}} = \frac{1}{|\hat{x}|} \sum_{\hat{x}_j \in \hat{x}} \max_{x_i \in x} \mathbf{x}_i^\top \hat{\mathbf{x}}_j, \qquad (8)$$

$$F_{\text{BERT}} = 2 \frac{P_{\text{BERT}} \cdot R_{\text{BERT}}}{P_{\text{BERT}} + R_{\text{BERT}}}, \qquad (9)$$

where the recall and precision elements are calculated based on a token-match approach. The recall score stems from comparing each token in the reference sentence "x" to the corresponding token in the generated sentence "$\hat{x}$". Conversely, the precision score is derived by comparing each token in "$\hat{x}$" to a token in "x". A greedy matching strategy is employed to pair tokens that demonstrate the highest degree of similarity. This method provides a comprehensive analysis of how precise and accurate the language model's output aligns with the factual reference sentence.

BLEURT: the Bilingual Evaluation Understudy with Representations from Transformers (BLEURT) metric (Sellam et al., 2020) employs a unique pre-training scheme, where BERT is initially trained on a significant corpus of synthetic sentence pairs. This is reinforced by multiple lexical and semantic-level supervision signals, used concurrently. Following this pre-training stage, BERT is further fine-tuned on rating data, and its objective is to estimate human rating scores accurately. The initial stage of pre-training is essential for this metric because it significantly enhances the model's robustness, thereby effectively transforming it into a secure bulwark against quality drifts inherent in generative systems.

BARTScore: (Yuan et al., 2021) is a metric proposed to evaluate the quality of the generated text, such as in applications like machine translation and summarization. This metric conceives this evaluation as a problem of text generation, modeled using pre-trained sequence-to-sequence models. BARTScore uses BART, an encoder-decoder-based pre-trained model, to translate generated text to and from a reference point, earning higher scores when the text is more accurate and fluent. This metric offers several variants that can be flexibly applied in an unsupervised manner for different perspectives of text evaluation, such as fluency or factuality. Tests have shown that BARTScore can outperform existing top-scoring metrics in the majority of test settings across multiple datasets and perspectives.

3.1.3 Human Evaluation Metrics

Human evaluation in factuality assessment is crucial due to its sensitivity to nuanced elements of language and context that may elude automated systems. Human evaluators excel at interpreting abstract concepts and emotional subtleties that can significantly inform the accuracy of evaluation. However, they are subject to limitations such as subjectivity, inconsistency, and potential for error. On the other hand, automated evaluations offer consistent results, and efficient processing of large data sets, and are ideal for tasks needing quantitative measurements. They also provide an objective benchmark for model performance comparison. Overall, an

ideal evaluation framework might blend automated evaluation's scalability and consistency with human evaluation's ability to interpret complex linguistic concepts.

Attribution: is a metric to verify that the output of LLMs is sharing only verifiable information about the external world. As proposed by (Rashkin et al., 2023), Attributable to Identified Sources (AIS) is a human evaluation framework that adopts a binary concept of attribution. A text passage y is deemed attributable to a set A of evidence if, and only if, an arbitrary listener would agree with the statement "According to A, y" within the context of y. The AIS framework awards a full score (1.0) if every element of content in passage y can be linked to the evidence set A. Conversely, it gives a score of zero (0.0) if this condition is not met.

Based on AIS, Gao et al. (2023a) propose a more fine-grained, sentence-level extension of AIS called Auto-AIS, where annotators assign AIS scores to each sentence, and an average score across all sentences is reported. This procedure effectively measures the percentage of sentences that are fully attributable to the evidence. Context, such as surrounding sentences and the question the text answers, is provided to annotators for more informed judgment. A limit is also set for the number of evidence snippets in the attribution report to maintain conciseness.

During model development, an automated AIS metric is defined to approximate human AIS evaluations, using a natural language inference model, which correlates well with AIS scores. Before computing the scores, they improve accuracy by decontextualizing each sentence in context.

FActScore: (Min et al., 2023) is a novel evaluation metric designed to assess the factual precision of long-form text generated by LLMs. The challenge of evaluating the factuality of such text arises from two main issues: (1) the generated content often contains a mix of supported and unsupported information, making binary judgments insufficient, and (2) human evaluation is both time-consuming and expensive. To address these challenges, FActScore breaks down a generated text into a series of atomic facts—short statements that each convey a single piece of information. Each atomic fact is then evaluated based on its support from a reliable knowledge source. The overall score represents the percentage of atomic facts that are supported by the knowledge source. The paper conducted extensive human evaluations to compute FActScores for biographies generated by several state-of-the-art commercial LLMs, including InstructGPT, ChatGPT, and the retrieval-augmented PerplexityAI. The results revealed that these LLMs often contain factual inaccuracies, with FActScores ranging from 42% to 71%. Notably, the factual precision of these models tends to decrease as the rarity of the entities in the biographies increases.

3.1.4 LLM-based Metrics

Using LLMs for evaluation offers efficiency, versatility, reduced reliance on human annotation, and the capability of evaluating multiple dimensions of conversation quality in a single model call, which improves scalability. However, potential issues include a lack of established validation, which can lead to bias or accuracy problems if the LLM used for evaluation is not thoroughly vetted. The decision process to identify suitable LLMs and decoding strategies can be complex and pivotal to obtaining accurate evaluations. The range of evaluation may also be limited, as the focus is often on open-domain conversations, possibly leaving out assessments in specific or narrow domains. While reducing human input can be beneficial, it can also miss out on crucial interaction quality aspects better evaluated by human judges, such as emotional resonance or nuanced understandings.

GPTScore: (Fu et al., 2023) is a new evaluation framework designed to assess the quality of output from generative AI models. To provide these evaluations, GPTScore taps into the emergent capabilities of 19 different pre-trained models, such as zero-shot instruction, and uses them to judge the generated texts. These models vary in scale from 80M to 175B. Testing across four text generation tasks, 22 aspects of evaluation, and 37 related datasets, has demonstrated that GPTScore can effectively evaluate text per instructions in natural language. This attribute allows it to sidestep challenges traditionally encountered in text evaluation, like the need for sample annotations and achieving custom, multi-faceted evaluations.

GPT-judge: (Lin et al., 2022b) is a finetuned model based on the GPT-3-6.7B, which is trained to evaluate the truthfulness of answers to questions in the TruthfulQA dataset. The training set consists of triples in the form of question-answer-label combinations, where the label can be either true or false. The model's training set includes examples from the benchmark and answers generated by other models assessed by human evaluation. In its final form, the GPT-judge uses examples from all models to evaluate the truthfulness of responses. This training includes all questions from the dataset, with the goal being to evaluate truth, not generalize new questions.

The study conducted by the authors (Lin et al., 2022b) focuses on the application of GPT-judge in assessing *Truthfulness* and *Informativeness* using the TruthfulQA Dataset. The authors undertook the fine-tuning of two distinct GPT-3 models to evaluate two essential aspects: Truthfulness, which pertains to the accuracy and honesty of information provided by the LLM, and Informativeness, which measures how effectively the LLM conveys relevant and valuable information in its responses. From these two fundamental concepts, the authors derived a combined metric denoted as *truth * info*. This metric represents the product of scalar scores for both truthfulness and informativeness. It not only quantifies the extent to which questions are answered truthfully but also incorporates the assessment of informativeness for each response. This comprehensive approach prevents the model from generating generic responses like "I have no comment" and ensures that responses are not only truthful but also valuable. These metrics have found widespread deployment in evaluating the factuality of information generated by LLMs (Chuang et al., 2023; Li et al., 2023d).

LLM-Eval: (Lin and Chen, 2023) is a novel evaluation methodology for open-domain dialogues with LLMs. Unlike conventional evaluation methods which rely on human annotations, ground-truth responses, or multiple LLM prompts, LLM-Eval uses a unique prompt-based evaluation

process employing a unified schema to assess various elements of a conversation's quality during a single model function. Extensive evaluations of LLM-Eval's performance using multiple benchmark datasets indicate it is effective, efficient, and adaptable compared to traditional evaluation practices. Further, the authors stress the necessity of selecting appropriate LLMs and decoding strategies for precise evaluation outcomes, underscoring LLM-Eval's versatility and dependability in assessing open-domain conversation systems across a variety of circumstances.

3.2 Benchmarks for Factuality Evaluation

In this section, we delve into the benchmarks that are prominently employed to assess the factuality of LLMs. Specific benchmarks tailored for evaluating factuality in LLMs are tabulated in Table 4.

MMLU (Hendrycks et al., 2021) and TruthfulQA (Lin et al., 2022b) stand as two pivotal benchmarks in the realm of evaluating the factuality of LLMs (OpenAI, 2023; Park, 2023; Touvron et al., 2023b) . The MMLU benchmark is proposed to measure a text model's multitask accuracy across a diverse set of 57 tasks. These tasks span a wide range of subjects, from elementary mathematics to US history, computer science, law, and more. The benchmark is designed to test both the world knowledge and problem-solving ability of models. The findings from the paper suggest that while most recent models perform at near random-chance accuracy, the largest GPT-3 model showcased a significant improvement. However, even the best models still have a long way to go before achieving expert-level accuracy across all tasks (OpenAI, 2023). TruthfulQA is a benchmark designed to assess the truthfulness of a language model's generated answers. The benchmark consists of 817 questions spanning 38 categories, including health, law, finance, and politics. These questions were crafted in such a way that some humans might answer them falsely due to misconceptions or false beliefs. The goal for models is to avoid generating these false answers that they might have learned from imitating human texts. The TruthfulQA benchmark serves as a tool to highlight the potential pitfalls of relying solely on LLMs for accurate information and emphasizes the need for continued research in this area.

HaluEval (Li et al., 2023c) is a benchmark designed to understand and evaluate the propensity of LLMs like ChatGPT to generate hallucinations. A hallucination, in this context, refers to content that either conflicts with the source or cannot be verified based on factual knowledge. The HaluEval benchmark offers a vast collection of generated and human-annotated hallucinated samples, aiming to evaluate the performance of LLMs in recognizing such hallucinations. The benchmark utilizes a two-step framework, termed "sampling-then-filtering", based on ChatGPT to generate these samples. Additionally, human labelers were employed to annotate hallucinations in ChatGPT responses. The HaluEval benchmark is a comprehensive tool that not only evaluates the hallucination tendencies of LLMs but also provides insights into the types of content and the extent to which these models are prone to hallucinate.

BigBench (Srivastava et al., 2023) focuses on the capabilities and limitations of LLMs. It comprises 204 tasks

from diverse domains such as linguistics, childhood development, math, common-sense reasoning, biology, physics, social bias, software development, and more. The benchmark is designed to evaluate tasks believed to be beyond the capabilities of current language models. The study evaluates the performance of various models, including OpenAI's GPT models, on BIG-bench and compares them to human expert raters. Key findings suggest that model performance and calibration improve with scale but are still suboptimal when compared to human performance. Tasks that involve a significant knowledge or memorization component show predictable improvement, while tasks that exhibit "breakthrough" behavior at a certain scale often involve multiple steps or components.

Huang et al. (2023b) propose C-Eval, the first comprehensive Chinese evaluation suite. It can be used to evaluate the advanced knowledge and reasoning abilities of foundational models within a Chinese context. The evaluation suite comprises multiple-choice questions spanning 52 diverse disciplines, with four levels of difficulty: middle school, high school, college, and professional. Additionally, C-Eval Hard is introduced for very challenging subjects within the C-Eval suite, which demands advanced reasoning abilities to solve. Their evaluations of state-of-the-art LLMs, including both English and Chinese-oriented models, have shows that there is significant room for improvement as only GPT-4 managed to achieve an average accuracy of over 60%. The authors focus on assessing LLMs' advanced abilities within a Chinese context. The researchers assert that LLMs intended for a Chinese environment should be evaluated based on their knowledge of Chinese users' primary interests, such as Chinese culture, history, and laws. With C-Eval, the authors aim to guide developers in understanding the abilities of their models from multiple dimensions to facilitate the development and growth of foundational models for Chinese users. Simultaneously, C-Eval has not just introduced a whole suite but subsets that can serve as individual benchmarks, thereby assessing certain model abilities and analyzing key strengths and limitations of foundational models. Experiment results have shown that although GPT-4, ChatGPT, and Claude were not exclusively tailored for Chinese data, they emerged as the top performers on C-Eval.

SelfAware (Yin et al., 2023) aims to investigate whether models recognize what they don't know. This dataset encompasses two types of questions: unanswerable and answerable. The dataset comprises 2,858 unanswerable questions gathered from various websites and 2,337 answerable questions extracted from sources such as SQuAD, HotpotQA, and TriviaQA. Each unanswerable question is confirmed as such by three human evaluators. In the conducted experiments, GPT-4 achieves the highest F1 score of 75.5, compared to a human score of 85.0. Larger models tend to perform better and in-context learning can enhance performance.

The Pinocchio benchmark (Hu et al., 2023) serves as an extensive evaluation platform, emphasizing factuality and reasoning for LLMs. This benchmark encompasses 20,000 varied factual queries from diverse sources, timeframes, fields, regions, and languages. It tests an LLM's capability to discern combined facts, process both organized and scat-

TABLE 4
Benchmarks for Evaluating Factuality in LLMs. This table details each benchmark's task type, associated sub-datasets, evaluation metrics, and the performance of select representative LLMs.

Reference	Task Type	Dataset	Metrics	Performance of Representative LLMs
MMLU (Hendrycks et al., 2021)	Multi-Choice QA	Humanities, Social, Sciences, STEM...	ACC	(ACC, 5-shot) GPT-4: 86.4 GPT-3.5: 70 LLaMA2-70B: 68.9
TruthfulQA (Lin et al., 2022b)	QA	Health, Law, Conspiracies, Fiction...	Human Score, GPT-judge, ROUGE, BLEU, MC1,MC2...	(zero-shot) GPT-4: ∼29 (MC1) GPT-3.5: ∼28 (MC1), 79.92(%true) LLaMA2-70B: 53.37 (%true)
C-Eval (Huang et al., 2023b)	Multi-Choice QA	STEM, Social Science, Humanities...	ACC	(zero-shot, average ACC) GPT-4: 68.7 GPT-3.5: 54.4 LLaMA2-70B: 50.13
AGIEval (Huang et al., 2023b)	Multi-Choice QA	Gaokao, (geometry, Bio, history...), SAT, Law...	ACC	(zero-shot, average ACC) GPT-4: 56.4 GPT-3.5: 42.9 LLaMA2-70B: 40.02
HaluEval (Li et al., 2023c)	Hallucination Evaluation	HaluEval	ACC	(general ACC) GPT-3.5: 86.22
BigBench (Srivastava et al., 2023)	Multi-tasks (QA, NLI, Fact Checking, Reasoning...)	BigBench	Metric to each type of task	(Big-Bench Hard) GPT-3.5: 49.6 LLaMA-65B: 42.6
ALCE (Gao et al., 2023b)	Citation Generation	ASQA, ELI5, QAMPARI	MAUVE, Exact Match, ROUGE-L...	(ASQA, 3-psg, citation prec) GPT-3.5: 73.9 LLaMA-33B: 23.0
QUIP (Weller et al., 2023)	Generative QA	TriviaQA, NQ, ELI5, HotpotQA	QUIP-Score Exact match	(ELI5, QUIP, null prompt) GPT-4: 21.0 GPT-3.5: 27.7
PopQA (Mallen et al., 2023)	Multi-Choice QA	PopQA, EntityQuestions	ACC	(overall ACC) GPT-3.5: ∼37.0
UniLC (Zhang et al., 2023e)	Fact Checking	Climate, Health, MGFN	ACC F1	(zero-shot, fact tasks, average F1) GPT-3.5: 51.62
Pinocchio (Hu et al., 2023)	Fact Checking QA, Reasoning	Pinocchio	ACC F1	GPT-3.5: (Zero-shot ACC: 46.8, F1:44.4) GPT-3.5: (Few-shot ACC: 47.1, F1:45.7)
SelfAware (Yin et al., 2023)	Self-evaluation	SelfAware	ACC	(instruction input, F1) GPT-4: 75.47 GPT-3.5: 51.43 LLaMA-65B: 46.89
RealTimeQA (Kasai et al., 2022)	Multi-Choice QA, Generative QA	RealTimeQA	ACC, F1	(original setting, GCS retrieval) GPT-3: 69.3 (ACC for MC) GPT-3: 39.4 (F1 for generation)
FreshQA (Vu et al., 2023)	Generative QA	FRESHQA	ACC (Human)	(strict ACC, null prompt) GPT-4: 28.6 GPT-3.5: 26.0

tered evidence, recognize the temporal evolution of facts, pinpoint minute factual disparities, and withstand adversarial inputs. Each reasoning challenge within the benchmark is calibrated for difficulty to allow for detailed analysis.

Kasai et al. (2022) introduce a dynamic QA platform named REALTIMEQA. This platform is unique in that it announces questions and evaluates systems on a regular basis, specifically weekly. The questions posed by REALTIMEQA pertain to current events or novel information, challenging the static nature of traditional open domain QA datasets. The platform aims to address instantaneous information needs, pushing QA systems to provide answers about recent events or developments. Their preliminary findings indicate that while GPT-3 can often update its generation results based on newly-retrieved documents, it sometimes returns outdated answers when the retrieved documents lack sufficient information.

Vu et al. (2023) present a dynamic benchmark called FreshQA designed to evaluate up-to-date world knowledge of LLMs. Its questions range from those that are never-changing to those that are fast-changing, as well as questions based on false premises. The aim is to challenge the

TABLE 5
Factuality Evaluation Studies of LLMs. This table shows studies focused on factuality evaluation, and those providing valuable insights into such evaluations. The column 'Human Eval' denotes if the evaluation incorporated human evaluation. The 'Granularity' column specifies the level of evaluation, with token-level (T) and sentence-level (S) distinctions.

Reference	Task	Dataset	Metrics	Human Eval	Evaluated LLMs	Granularity
FActScore (Min et al., 2023)	Biography Generation	183 people entities	F1	✓	GPT-3.5 ChatGPT...	T
SelfCheckGPT (Manakul et al., 2023)	Bio Generation	WikiBio	AUC-PR, Human Score	✓	GPT-3, LLaMA, OPT, GPT-J...	S
(Wang et al., 2023a)	Open QA	NQ, TQ	ACC, EM	✓	GPT-3.5 ChatGPT GPT-4 Bing Chat	S
(Pezeshkpour, 2023)	Knowledge Probing	T-REx, LAMA	ACC		GPT3.5	T
(De Cao et al., 2021)	QA, Fact Checking	KILT, FEVER, zsRE	ACC		GPT-3 FLAN-T5	S/ T
(Varshney et al., 2023)	Article Generation	Unnamed Dataset	ACC, AUC		GPT3.5 Vicuna	S
FactTool (Chern et al., 2023)	KB-based QA	RoSE	ACC F1...		GPT-4 ChatGPT FLAN-T5	S
(Kadavath et al., 2022)	Self-evaluation	BIG Bench, MMLU, LogiQA, TruthfulQA, QuALITY TriviaQA Lambada	ACC, Brier Score, RMS Calibration Error...		Claude	T
(Chen et al., 2023c)	QA	NQ, WoW	Factuality, Relevance, Coherence, Informativeness, Helpfulness, and Validity		LLaMA, FLAN-T5, ...	T

static nature of LLMs and test their adaptability to ever-changing knowledge through human evaluations. They develop a reliable evaluation protocol that uses a two-mode system: RELAXED and STRICT for a comprehensive understanding of model performance, ensuring that answers are confident, definitive, and accurate. They also provide a strong baseline named FRESHPROMPT, which seeks to enhance LLM performance by integrating real-time data from search engines. Initial experiments reveal that, outdated training data weakens the performance of LLMs and the FRESHPROMPT method can significantly enhance it. The research underscores the need for LLMs to be refreshed with current information to ensure their relevance and accuracy in a constantly evolving world.

FreshQA (Vu et al., 2023) is a dynamic benchmark designed to evaluate up-to-date world knowledge of LLMs. Its questions range from those that are never-changing to those that are fast-changing, as well as questions based on false premises. The aim is to challenge the static nature of LLMs and test their adaptability to ever-changing knowledge through human evaluations. They develop a reliable evaluation protocol that uses a two-mode system: RELAXED and STRICT for a comprehensive understanding of model performance, ensuring that answers are confident, definitive, and accurate. They also provide a strong baseline named FRESHPROMPT, which seeks to enhance LLM performance by integrating real-time data from search engines. Initial experiments reveal that, outdated training data weakens the performance of LLMs and the FRESHPROMPT method can significantly enhance it. The research underscores the need for LLMs to be refreshed with current information to ensure their relevance and accuracy in a constantly evolving world.

Several benchmarks, such as BigBench (Srivastava et al., 2023) and C-Eval (Huang et al., 2023b), encompass subsets that extend beyond the realm of factual knowledge or factuality. In this work, we specifically emphasize and focus on

those subsets related to factuality.

There are benchmarks primarily designed for Pretrained Language Models (PLMs) that can also be adapted for LLMs. Some studies use them for evaluating LLM's factuality, but they are not so widely-used, so we have chosen to exclude them from the table for clarity. These benchmarks predominantly encompass knowledge-intensive tasks, as highlighted by (Petroni et al., 2021). Those benchmarks include NaturalQuestions (NQ) (Kwiatkowski et al., 2019), TriviaQA (TQ) (Joshi et al., 2017), OTT-QA (Chen et al., 2021b), AmbigQA Min et al. (2020) and WebQuestion (WQ) (Berant et al., 2013) of Open-domain Question Answering (QA) task, the HotpotQA (Yang et al., 2018), 2WikiMultihopQA (Ho et al., 2020), IIRC (Ferguson et al., 2020), MuSiQue (Trivedi et al., 2022) of multi-step QA, the ELI5 (Fan et al., 2019) of the Long-form QA task, the FEVER (Thorne et al., 2018), FM2 (Eisenschlos et al., 2021), HOVER (Jiang et al., 2020), FEVEROUS (Aly et al., 2021) of the Fact Checking task, the T-REx (Elsahar et al., 2018), zsRE (Levy et al., 2017) and LAMA (Petroni et al., 2019a) to examine factual knowledge contained in pretrained language models, the WikiBio (Lebret et al., 2016) of biography generation task, the RoSE (Liu et al., 2023d), WikiAsp (Hayashi et al., 2021) of summarization task, the KILT (Petroni et al., 2021) of comprehensive knowledge intensive tasks, the MassiveText (Rae et al., 2022b), Curation Corpus (Curation, 2020), Wikitext103 (Merity et al., 2016), Lambada (Paperno et al., 2016), C4 (Raffel et al., 2020b), Pile (Gao et al., 2020) of langauge modeling task, the WoW (Dinan et al., 2019), DSTC7 track2 (Galley et al., 2019a), DSTC11 track5 (Zhao et al., 2023a) of dialogue task, the RealToxicityPrompts Gehman et al. (2020) of toxicity reduction, the CommaQA (Khot et al., 2022), StrategyQA (Geva et al., 2021a), TempQuestions (Jia et al., 2018), IN-FOTABS (Gupta et al., 2020a) of diverse reasoning tasks.

Some studies (Manakul et al., 2023; Min et al., 2023) also provide a small dataset, but they mainly concentrate on the evaluation metrics or methods for factuality, we choose to discuss them in the next subsection.

3.3 Factuality Evaluation Studies

In this section, we delve into studies that evaluate factuality in LLMs without introducing a specific benchmark, focusing primarily on those whose main contribution lies in the evaluation methodology. We spotlight works that have pioneered evaluation techniques, metrics, or have offered distinctive insights into the factuality evaluation of LLMs.

Manakul et al. (2023) use an evaluation process that encompasses several key steps. Initially, synthetic Wikipedia articles are generated using GPT-3, focusing on individuals from the Wikibio dataset. Subsequently, manual annotation is performed at the sentence level, classifying sentences as "Major Inaccurate" , "Minor Inaccurate", or "Accurate", with "Major Inaccurate" denoting sentences unrelated to the topic. Passage-level scores are derived by averaging sentence-level labels, and identifying cases of total inaccuracies through score distribution analysis. Inter-annotator agreement is assessed using Cohen's κ Cohen (1960). Evaluation metrics primarily employ precision-recall curves (PR-Curves), distinguishing between "Non-Factual Sentences," "Non-Factual* Sentences" (a specific subset), and "Factual Sentences." These PR-Curves elucidate the trade-off between precision and recall for different detection methods.

Wang et al. (2023a) ask several LLMs, including ChatGPT, GPT-4 (OpenAI, 2023), BingChat (Microsoft, 2023) to answer open questions from NaturalQuestions (Kwiatkowski et al., 2019) and TriviaQA (Joshi et al., 2017). They manually estimate the accuracy of those LLMs on open question answering, and find that though LLMs can achieve nice performance but still far away from perfect.

Pezeshkpour (2023) propose a new metric for measuring whether a certain type of knowledge is present in LLM. This metric is based on information theory and measures knowledge by analyzing the probability distribution of predictions made by LLM before and after injecting target knowledge. The accuracy of GPT-3.5 in the Knowledge Probing task is tested on the T-REx (Elsahar et al., 2018) and LAMA (Petroni et al., 2019a) datasets.

Varshney et al. (2023) point out a common real-world occurrence where users often ask questions based on false premises. These questions are challenging for state-of-the-art models. This necessitated the creation of a new evaluation dataset. To this end, the authors have conducted a case study and compiled a set of 50 such adversarial questions, all of which the GPT-3.5 model answered incorrectly. The aim is to create a challenging experimental setup to assess the performance of models faced with such questions. In order to enhance the evaluation, corresponding true premise questions were created for each of the false premise questions. This allowe for a holistic evaluation of the model's performances, taking into consideration both correct and incorrect premises. The authors make sure to evaluate the complete answers given by the model for correct and incorrect questions - in this context, it's not enough for an answer to be partially correct, the entire answer needs to be accurate to be marked as correct. For example: For the false premise question "Why does Helium have an atomic number of 1?", the corresponding true premise question is "Why does Hydrogen have an atomic number of 1?".

FACTOOL (Chern et al., 2023) is a tool designed to function as a factuality detector, with the primary purpose of auditing generative chatbots and assessing the reliability of their outputs. This tool is employed to evaluate several contemporary chatbots, including GPT-4, ChatGPT, Claude, Bard (Google, 2023), and Vicuna (Chiang et al., 2023). Notably, FACTOOL itself leverages the capabilities of GPT-4. For the evaluation process, the researchers have curated a diverse set of prompts: 30 from knowledge-based question answering (KB-QA), 10 each from code, math, and scientific domains. The KB-QA prompts were sourced from a prior study, code prompts were taken from HumanEval, math prompts from another distinct study, while the scientific prompts were crafted by the authors themselves. The evaluation metrics included both claim-level and response-level accuracies for each chatbot. To offer a more comprehensive and equitable evaluation, a weighted claim-level accuracy is used. The weighting is determined based on the proportion of prompts from each category. The findings are illuminating. GPT-4 emerge as the top performer in terms of both weighted claim-level factual accuracy and response-level accuracy among all the chatbots assessed. Another intrigu-

TABLE 6
Factuality Evaluation of LLMs. This table lists the evaluation metrics used by studies aiming to improve the factuality of LLMs. The column 'Human Eval' denotes if the evaluation incorporated human evaluation. The 'Granularity' column specifies the level of evaluation, with token-level (T) and sentence-level (S) distinctions.

Reference	Task	Dataset	Metrics	Human Eval	Evaluated LLMs	Granularity
Retro (Borgeaud et al., 2022)	QA, Language Modeling	MassiveText, Curation Corpus, Wikitext103, Lambada, C4,Pile, NQ	PPL, ACC, Exact Match	✓	Retro	T
GenRead (Yu et al., 2023)	QA, Dialogue, Fact Checking	NQ, TQ, WebQ, FEVER, FM2, WoW	EM, ACC, F1, Rouge-L		GPT3.5, Codex GPT-3, Gopher FLAN, GLaM PaLM	S
GopherCite (Menick et al., 2022b)	Self-supported QA	NQ, ELI5, TruthfulQA (Health Law, Fiction Conspiracies)	Human Score	✓	GopherCite	S
(Trivedi et al., 2023)	QA	HotpotQA, IIRC 2WikiMultihopQA, MuSiQue(music)	Retrieval recall, Answer F1		GPT-3 FLAN-T5	S/ T
(Peng et al., 2023)	QA, Dialogue	DSTC7 track2, DSTC11 track5, OTT-QA	ROUGE, chrF, BERTScore, Usefulness, Humanness...	✓	ChatGPT	S/ T
CRITIC (Gou et al., 2023)	QA Toxicity Reduction	AmbigNQ, TriviaQA, HotpotQA, RealToxicityPrompts	Exact Match, maximum toxicity, perplexity, n-gram diversity, AUROC...,		GPT-3.5 ChatGPT	T
(Khot et al., 2023)	QA, long-context QA	CommaQA-E, 2WikiMultihopQA, MuSiQue, HotpotQA	Exact Match, Answer F1		GPT-3 FLAN-T5	T
ReAct (Yao et al., 2023a)	QA Fact Verification	HotpotQA, FEVER	Exact Match, ACC		PaLM GPT-3	S/T
(Jiang et al., 2023b)	QA, Commonsense Reasoning, long-form QA...	2WikiMultihopQA, StrategyQA, ASQA, WikiAsp	Exact Match, Disambig-F1, ROUGE, entity F1...		GPT-3.5	T
(Lee et al., 2022b)	Open-ended Generation	FEVER	Entity score, Entailment Ratio, ppl...		Megatron-LM	T
SAIL (Luo et al., 2023a)	QA Fact Checking	UniLC	ACC F1		LLaMA Vicuna SAIL	T
(He et al., 2022)	Commonsense Reasoning, Temporal Reasoning, Tabular Reasoning	StrategyQA, TempQuestions, IN-FOTABS	ACC		GPT-3	T
(Pan et al., 2023a)	Fact Checking	HOVER FEVEROUS-S	Macro-F1		Codex FLAN-T5	S
(Du et al., 2023)	Biography MMLU	Unnamed Biography Dataset, MMLU	ChatGPT Evaluator, ACC		Bard ChatGPT	S

ing observation is that chatbots that underwent supervised fine-tuning, such as Vicuna-13B, exhibited commendable performance in standard scenarios like KB-QA. However, their performance dip in more intricate scenarios, including those involving math, code, and scientific queries.

Kadavath et al. (2022) investigate whether language models can evaluate the accuracy of their own assertions and predict which questions they can answer correctly. It is found that larger models are well-calibrated on diverse multiple-choice and true/false questions if given in the appropriate format. The approach to self-evaluation on open-ended tasks is to ask the models to initially suggest answers, and then evaluate the probability (P[True]) that their answers are correct. This resulted in compelling performance, calibration, and scaling on a diverse range of tasks. Furthermore, self-evaluation performance improved when the models are allowed to consider many of their own suggestions before predicting the validity of a specific one.

Yu et al. (2023) explore whether the internal knowledge of LLMs can replace the retrieved documents on knowledge intensive tasks. They ask LLMs, such as InstructGPT (OpenAI, 2022a), to directly generate contexts given a question rather than retrieving from database. They find the generated documents contain the golden answers more often than the top retrieved documents. Then they feed the generated docs and retrieved docs to the Fusion-in-Decoder model (Izacard and Grave, 2021b) for knowledge-intensive tasks such as Open-domain QA (Kwiatkowski et al., 2019) and find the generated docs are more effective than the retrieved docs, suggesting that the LLMs contain enough knowledge for knowledge-intensive tasks.

Menick et al. (2022b) propose a task named *Self-supported QA* to evaluate LLMs' ability in also producing citations when generating answers. Authors ask humans to evaluate whether the responses of their proposed model GopherCite are plausible and whether they are supported by the accompanying quote evidence on datasets such as NQ, ELI5, TruthfulQA.

Chen et al. (2023c) propose CONNER, a framework that evaluates LLMs as generators of knowledge. It focuses on six areas: Factuality, Relevance, Coherence, Informativeness, Helpfulness, and Validity. It evaluates whether the generated information can be backed by external proof (Factuality), is relevant to the user's query (Relevance), and is logically consistent (Coherence). It also checks if the knowledge provided is novel or surprising (Informativeness). The Extrinsic evaluation measures whether the knowledge enhances downstream tasks (Helpfulness) and its results are factually accurate (Validity).

In the realm of factuality evaluation, the Model Editing task holds a unique position, focusing on refining the internal knowledge of models. This task comes with its own set of specialized evaluation metrics. Predominantly, the Zero-Shot Relation Extraction (zsRE) Levy et al. (2017) and CounterFact Meng et al. (2022) serve as the primary benchmarks for assessing Model Editing techniques. When evaluating these methods, several key criteria emerge: Reliability: Post-editing, the model should consistently generate the intended output. Generalization: The model should be adept at producing the target output even when presented with paraphrased inputs. Locality: Edits should be localized, ensuring that facts not related to the specific edit remain intact. However, given the intricate web of interconnected facts, recent studies Cohen et al. (2023a); Yao et al. (2023b); Zhong et al. (2023b) have advocated for a more holistic evaluation approach. They introduce broader criteria for fact updates, encompassing aspects like portability, logical generalization, among others.

Some work's main contributions lie at the methods to improve the factuality of LLMs, and their evaluation part may also be informative and important when people reach the related study. So we choose to list evaluation parts in the Table 6 but not to discuss their evaluation in detail.

3.4 Evaluating Domain-specific Factuality

To assess the performance and factuality of these specialized LLMs, a plethora of datasets and benchmarks have been proposed across various domains. These resources not only serve as critical tools for evaluating the capabilities of LLMs but also facilitate advancements in specialized applications. We summarize them in Table 7. The distinction between this subsection and the previous two lies in its focus. This subsection delves deeper into factuality evaluation tailored to specific domains, while Sec 3.2 and 3.3 primarily concentrate on general factuality evaluation, with only a portion of their content dedicated to datasets evaluating factuality within specific domains.

Finance: Xie et al. (2023b) designed a financial natural language understanding and prediction evaluation benchmark dubbed FLARE, based on their collected financial instruction tuning dataset FIT. This benchmark is used to evaluate their FinMA model. It randomly selects validation sets from FIT to choose the best model checkpoint and utilizes distinct test sets for evaluation. FLARE is a broader variant compared to the existing FLUE benchmark Shah et al. (2022) as it also encapsulates financial prediction tasks like stock movement prediction in addition to standard NLP tasks. The FLARE dataset includes several subtasks, such as sentiment analysis (FPB, FiQA-SA), news headline classification (Headline), named entity recognition (NER), question answering (FinQA, ConvFinQA), and stock movement prediction (BigData22, ACL18, CIKM18). Performance is gauged via a variety of metrics for each task, such as the accuracy and weighted F1 Score for sentiment analysis, entity-level F1 score for named entity recognition, and accuracy and Matthews correlation coefficient for stock movement prediction. Several methods, including their own FIT-fine-tuned FinMA and other LLMs (BloombergGPT, GPT-4, ChatGPT, BLOOM, GPT-NeoX, OPT-66B, Vicuna-13B) are dedicated to their comparison. BloombergGPT's performance is assessed in various shot scenarios, while the zero-shot performance is reported for the remaining results. Some of the baselines depend on human evaluations given that LLMs without fine-tuning fail to generate instruction-defined answers. Conversely, FinMA's results are conducted on a zero-shot basis and can be evaluated automatically. To enable direct comparison between the performance of FinMA and BloombergGPT, despite the former not releasing their test datasets, test datasets were constructed with the same data distribution.

Li et al. (2023e) propose EcomInstruct benchmark, which experiment investigates the performance of their EcomGPT

TABLE 7
Benchmarks for domain-specific factuality evaluation. The table presents the domain, tasks, datasets, and the LLMs evaluated in the respective study.

Reference	Domain	Task	Datasets	Metrics	Evaluated LLMs
Xie et al. (2023b)	Finance	Sentiment analysis, News headline classification Named entity recognition Question answering Stock movement prediction	FLARE	F1, Acc, Avg F1, Entity F1, EM, MCC	GPT-4 , BloombergGPT, FinMA-(7B, 30B, 7B-full), Vicuna-7B
Li et al. (2023e)	Finance	134 E-com tasks	EcomInstruct	Micro-F1, Macro-F1, ROUGE	BLOOM, BLOOMZ, ChatGPT, EcomGPT
Wang et al. (2023d)	Medicine	Multi-Choice QA	CMB	Acc	GPT-4, ChatGLM2-6B, ChatGPT, DoctorGLM, Baichuan-13B-chat, HuatuoGPT, MedicalGPT, ChatMed-Consult, ChatGLM-Med , Bentsao, BianQue-2
Li et al. (2023b)	Medicine	Generative-QA	Huatuo-26M	BLEU, ROUGE, GLEU	T5, GPT2
Jin et al. (2023)	Medicine	Nomenclature, Genomic location, Functional analysis, Sequence alignment	GeneTuring	Acc	GPT-2, BioGPT, BioMedLM, GPT-3, ChatGPT, New Bing
Guha et al. (2023)	Law	Issue-spotting, Rule-recall, Rule-application, Rule-conclusion, Interpretation, Rhetorical-understanding	LegalBench	Acc, EM	GPT-4, GPT-3.5, Claude-1, Incite, OPT Falcon, LLaMA-2, FLAN-T5...
Fei et al. (2023)	Law	Legal QA, NER, Sentiment Analysis, Reading Comprehension	LawBench	F1, Acc, ROUGE-L, Normalized log-distance ...	GPT-4, ChatGPT, InternLM-Chat, StableBeluga2...

language model in comparison to baseline models such as BLOOM and BLOOMZ. Categories of these baseline models include pre-trained large models with decoder-only architecture like BLOOM, and instruction-following language models like BLOOMZ and ChatGPT. The evaluation metric of the EcomInstruct involves converting all tasks to generative paradigms and using text generation evaluation metrics like ROUGE-L. Classification tasks were evaluated with precision, recall, and F1 scores. The EcomInstruct dataset, comprising 12 tasks across four major categories, is divided into training and testing sections. The EcomGPT is trained on 85,746 instances of E-commerce data. The performance of the model is assessed based on its capacity to generalize unseen tasks or datasets, with emphasis on cross-language and cross-task paradigm settings.

Medicine: Wang et al. (2023d) propose a localized medical benchmark called CMB, or the Comprehensive Medical Benchmark in Chinese. CMB is rooted entirely in the native Chinese linguistic and cultural framework. While traditional Chinese medicine is a significant part of this evaluation, it does not make up the entire benchmark. Both prominent LLMs, such as ChatGPT and GPT-4, and localized Chinese LLMs, including those specializing in the health domain, are evaluated using CMB. However, the benchmark is not designed as a leaderboard competition, but rather as a tool for self-assessment and understanding the progression

of models in this field. CMB embodies a comprehensive, multi-layered medical benchmark in Chinese, comprised of hundreds of thousands of multiple-choice questions and complex case consultation questions. This wide range of queries covers all clinical medical specialties and various professional levels, seeking to evaluate a model's medical knowledge and clinical consultation capabilities comprehensively.

Li et al. (2023b) introduce Huatuo-26M dataset, the largest Chinese medical Question and Answer (QA) dataset to date, including over 26 million high-quality medical QA pairs. It covers a wide range of topics, such as diseases, symptoms, treatments, and drug information. The dataset is a valuable resource for anyone looking to improve AI applications in the medical field, such as chatbots and intelligent diagnostic systems. The Huatuo-26M dataset is gathered and integrated from various sources, including online medical encyclopedias, online medical knowledge bases, and online medical consultation records. Each QA pair in the dataset contains a problem description and a corresponding answer from a doctor or expert. Although a significant proportion of the Huatuo-26M dataset is constituted by the online medical consultation records, the text format data from these records is not publicly available, for unspecified reasons. This dataset is expected to be instrumental for multiple types of research and AI applications

in the medical field. These areas of application extend to Natural Language Processing tasks like developing QA systems, text classification, sentiment analysis, and Machine Learning model training tasks like disease prediction and personalized treatment recommendation. Consequently, the dataset is favorable for developing AI applications in the medical field, from intelligent diagnosis systems to medical consultation chatbots.

Jin et al. (2023) use nine tasks related to NCBI resources to evaluate the proposed GeneGPT model. The dataset used is GeneTuring benchmark Hou and Ji (2023), which contains 12 tasks with 50 question-answer pairs each. The tasks are divided into four modules - Nomenclature (Gene alias, Gene name conversion), Genomic location(Gene SNP association, Gene location, SNP location), Functional analysis(Gene disease association, Protein-coding genes), Sequence alignment (DNA to human genome, DNA to multiple species). Two settings of GeneGPT are assessed: a full setting where all prompt components are used, and a slim setting using only two components. The performance of GeneGPT is compared against various baselines including general-domain GPT-based LLMs like GPT-2, GPT-3, and ChatGPT. Additionally, GPT-2-sized biomedical domain-specific LLMs such as BioGPT and BioMedLM are evaluated. The new Bing, a retrieval-augmented LLM with access to relevant web pages, is also assessed. The result evaluation of the compared methods is based on the results reported in the original benchmark and are manually evaluated. However, the evaluation of the proposed GeneGPT method is determined through automatic evaluations.

Law: LegalBench (Guha et al., 2023) is a benchmark for legal reasoning introduced due to the increasing use of LLMs in the legal field. It consists of 162 tasks covering six types of legal reasoning and was collaboratively constructed through an interdisciplinary process with significant contributions from legal professionals. The tasks designed aim to either demonstrate practical legal reasoning capabilities or measure reasoning skills that are of interest to lawyers. To facilitate discussions between different fields relating to LLMs in law, LegalBench tasks correspond to popular legal frameworks for describing legal reasoning, thus creating a shared language between lawyers and LLM developers. The paper not only describes LegalBench, but also presents an evaluation of 20 different open-source and commercial LLMs and highlights the types of research opportunities that LegalBench can provide.

LawBench (Fei et al., 2023) is an evaluation framework dedicated to assessing the capabilities of LLMs in relation to legal tasks. The context-specificity and high-stakes nature of the law field make it crucial to have a clear grasp of LLMs' legal knowledge and their ability to execute legal tasks. LawBench dives deep into three cognitive evaluations of LLMs: the capability to memorize crucial legal details, understanding legal texts, and applying legal knowledge to resolve complicated legal problems. A total of 20 diverse tasks have been put together, covering five main task categories: single-label classification, multi-label classification, regression, extraction, and generation. Throughout the evaluation process, 51 LLMs were extensively tested under LawBench, compiling a spectrum of language models

including 20 multilingual LLMs, 22 Chinese-oriented LLMs, and 9 law-specific LLMs. The results concluded that GPT-4 ranked as the superior LLM in the law domain, significantly surpassing its competitors. Despite the noted improvements when LLMs were fine-tuned on law-specific texts, the study acknowledged that there is still a long road ahead in achieving highly reliable LLMs for legal tasks.

4 ANALYSIS OF FACTUALITY

In the previous Section 3, we provide quantitative statistics related to evaluating factuality. In this section, we delve deeper, exploring the underlying mechanisms that influence factuality in large language models.

4.1 Analysis of Factuality

This subsection delves into intriguing analyses concerning the factuality of LLMs, focusing on aspects that aren't directly tied to evaluation, or enhancement. Specifically, we explore the mechanisms through which LLMs process, interpret, and produce factual content. The subsequent sections offer an in-depth examination of different dimensions of factuality in LLMs, ranging from their knowledge storage, and awareness to their approach to managing conflicting data. [2]

4.1.1 Knowledge Storage

The language model serves as a repository of knowledge, storing a multitude of information about the world within its parameters Petroni et al. (2019b). However, the organization of this knowledge within LLMs remains largely mysterious. Meng et al. (2022) introduced a methodology called causal tracing to measure the indirect impact of hidden states or activations. This technique was employed to illustrate that factual knowledge is primarily stored in the early-layer feed-forward networks (FFNs) of such models. Similarly, Geva et al. (2021c) also suggests that a substantial portion of factual information is encoded within the FFN layers. They conceptualize the input of the FFN as a query, the first layer as keys, and the second layer as values. Consequently, the intermediate hidden dimension of the FFN can be interpreted as the number of memories within the layer, and the intermediate hidden state represents a vector comprising activation values for each memory. As a result, the final output of the FFN can be understood as the weighted sum of activated values. The authors further demonstrate that the value vectors often encapsulate human-interpretable concepts and knowledge Geva et al. (2023, 2022). In addition, Chen et al. (2023d) made an intriguing finding that the language model contains language-independent neurons that express multilingual knowledge and degenerate neurons that convey redundant information by applying the integrated gradients method Lundstrom et al. (2022). Nevertheless, it is important to note that the aforementioned studies primarily focus on the representation of individual facts, and the comprehensive understanding of how factual knowledge is precisely organized

2. While some research (Neeman et al., 2023; Wang et al., 2021) applies some similar analyses to Pretrained Language Models. This review excludes them as they are not typically considered Large Language Models.

and interconnected within these models remains an ongoing challenge.

4.1.2 Knowledge Completeness and Awareness

This subsubsection delves into the intriguing realm of LLMs' self-awareness, their capacity to discern their knowledge gaps, and the balance between their internally generated knowledge and externally retrieved information. We delve into the dichotomy between parametric knowledge and retrieved knowledge, exploring the promises and challenges these models bring to the forefront of knowledge-intensive tasks.

Knowledge Awareness: Several studies have investigated the knowledge awareness of Large Language Models, specifically assessing whether LLMs can accurately estimate the correctness of their own responses. Most of these studies treat LLMs as "black boxes," prompting the models to report their confidence levels or calculating the perplexity of the model's output as an indicator of response likelihood. Gou et al. (2023) explore the model's ability to validate and iteratively refine its outputs, akin to how humans interact with tools. The authors find that solely relying on self-correction without external feedback can lead to marginal improvements or even diminished performance. Ren et al. (2023) experiment with settings either augmented or not with external document retrieval to determine whether models recognize their own knowledge boundaries. Their findings indicate that LLMs possess an inaccurate perception of their factual knowledge boundaries and tend to be overly confident about their responses. LLMs often fail to fully harness the knowledge they possess; however, retrieval enhancement can somewhat compensate for this shortcoming. Yin et al. (2023) introduce a dataset named "SelfAware" to test if models recognize what they don't know, encompassing both answerable and unanswerable questions. The experiment suggests that models do possess some capacity to discern their own knowledge gaps, but they are still far from human levels. GPT-4 outperforms other models, instructions and In-Context-Learning Dong et al. (2023) can enhance a model's discriminatory ability. Kadavath et al. (2022) focus on LLM self-assessment based on Language Model calibration using multiple-choice questions. Their findings revealed that the "none of the above" option decreased accuracy, larger models showed better calibration, and RLHF hindered model calibration levels. However, simply adjusting the temperature parameter can rectify this issue. Azaria and Mitchell (2023) assess the truthfulness of statements generated by LLMs, by using the model's internal state and hidden layer activations. The authors, employing a feedforward neural network, can classify if the model is misleading by utilizing the hidden output states.

Parametric Knowledge vs Retrieved Knowledge: Yu et al. (2023) explore whether the internal knowledge of LLMs can replace the retrieved documents on knowledge-intensive tasks. They ask LLMs, such as InstructGPT, to directly generate contexts given a question rather than retrieving them from the database. They find the generated documents contain the golden answers more often than the top retrieved documents. Then they feed the generated docs and retrieved docs to the Fusion-in-Decoder model (Izacard and Grave, 2021b) for knowledge-intensive tasks such as Open-domain QA (Kwiatkowski et al., 2019) and find the generated docs are more effective than the retrieved docs, suggesting that the LLMs contain enough knowledge for knowledge-intensive tasks.

On the contrary, these observations have been contested in subsequent investigations. Kandpal et al. (2023) underscore the dependency of LLMs on the number of associated documents seen during pre-training. They argue that the success in answering fact-based questions is highly linked to the number of documents containing the topic of the question that were encountered in pre-training. The study further posited the necessity of scaling models extensively to achieve competitive performance for questions with minimum representation in the training data. Adding to these concerns, Sun et al. (2023b) critically evaluate the factual knowledge base of LLMs, using a specifically designed Head-to-Tail benchmark comprised of 18K question-answer pairs. The results show that the understanding of factual knowledge, particularly related to torso-to-tail entities, by currently available LLMs is suboptimal.

In summary, while LLMs show promise in handling knowledge-intensive tasks, their dependency on pre-training information and limitations in factual accuracy remain significant hurdles. It underscores the need for further advancements in the field and the importance of incorporating complementary methods, such as retrieval augmentation, to enhance the learning of long-tail knowledge in LLMs.

4.1.3 Contextual Influence and Knowledge Conflict

This sub-subsection examines the interplay between an LLM's inherent parametric knowledge and the provided contextual knowledge, exploring both the model's capacity to utilize context and its behavior when confronted with conflicting information.

Contextual Influence on Generation: Some works explore the model's capacity to utilize context, for example, Li et al. (2023a) observe that larger models tend to rely on their parametric knowledge, even when faced with counterfactual contexts. This suggests that as models increase in size, they might grow more confident in their internal knowledge, potentially sidelining external context. However, the introduction of irrelevant contexts can still influence their outputs. The balance between controllability (relying on relevant context) and robustness (resisting irrelevant context) emerges as a challenge in LLM training. The study indicates that reducing context noise improves controllability, but the effect on robustness remains to be seen. In contrast, Zhou et al. (2023b) propose prompt templates to guide LLMs towards more faithful text generation. Among these, opinion-based prompts prove most effective, indicating that when LLMs are queried about opinions, they adhere more closely to the context. Interestingly, the study finds that using counterfactual context enhances the model's faithfulness, while original context, sourced from platforms like Wikipedia, might induce a simplicity bias, leading LLMs to answer questions without heavily relying on the context. Chen et al. (2023b) conduct a comprehensive evaluation of LLMs' ability to effectively utilize retrieved information. The study reveals that while retrieved documents can boost

LLM performance, the presence of noise in these documents can hinder it. Yue et al. (2023) investigate the nature of LLM-generated content in relation to provided references. They categorize the generated content as attributable, contradictory, or extrapolatory to the reference. Both fine-tuned models and instruction-based LLMs struggle to accurately evaluate the alignment between generated content and references, underscoring the challenge of ensuring that LLMs produce content consistent with the provided context.

Handling Knowledge Conflicts: A series of studies are interested in LLMs' behavior when confronted with conflicting information. Longpre et al. (2021) introduce the concept of knowledge conflicts, where the provided context contradicts the model's learned information. Their findings suggest that such conflicts lead to increased prediction uncertainty, especially for in-domain examples. Observations across models, ranging from T5-60M to 11B, indicate that larger models tend to default to their parametric knowledge. Moreover, there's an inverse relationship between retrieval quality and the tendency to rely on internal knowledge: the more irrelevant the evidence, the more the model defaults to its parametric knowledge. Chen et al. (2022) conduct experiments on typical ODQA models, including FiD and RAG. Their results show that FiD models rarely resort to memorization (less than 3.6% for NQ) compared to RAG models. Instead, FiD primarily grounds its answers in the provided evidence. Interestingly, when confronted with conflicting retrieved passages, models tend to fall back on their parametric knowledge. Xie et al. (2023a) explore the behavior of recent LLMs, including ChatGPT and GPT-4. Contrary to findings on smaller LMs, they discover that LLMs can be highly receptive to external evidence, even if it contradicts their parametric memory, provided the external evidence is coherent and convincing. Additionally, LLMs exhibit a strong confirmation bias, especially when presented with evidence that aligns with their parametric memory. This bias becomes even more pronounced for widely accepted knowledge. In scenarios where no relevant evidence is provided, LLMs tend to express uncertainty. However, when presented with both relevant and irrelevant evidence, they demonstrate an ability to filter out irrelevant information.

In conclusion, while studies like Li et al. (2023a) and Zhou et al. (2023b) emphasize the challenges and potential solutions in making LLMs more context-aware, others like Yue et al. (2023) and Xie et al. (2023a) highlight the inherent biases and limitations of LLMs. The overarching theme is the need for a balanced approach, where LLMs effectively leverage both their internal knowledge and external context to produce accurate and coherent outputs.

4.2 Causes of Factual Errors

Understanding the root causes of these factual inaccuracies is crucial for refining these models and ensuring their reliable application in real-world scenarios. In this subsection, we delve into the multifaceted origins of these errors, categorizing them based on the stages of model operation: Model Level, Retrieval Level, Generation Level, and other miscellaneous causes. Table 1 shows examples of factuality errors caused by different factors.

4.2.1 Model-level Causes

This subsection delves into the intrinsic factors within large language models that contribute to factual errors, originating from their inherent knowledge and capabilities.

Domain Knowledge Deficit: The model may lack comprehensive expertise in specific domains, leading to inaccuracies. Every LLM has its limitations based on the data it was trained on. If an LLM hasn't been exposed to comprehensive data in a specific domain during its training, it's likely to produce inaccurate or generalized outputs when queried about that domain. For instance, while an LLM might be adept at answering general science questions, it might falter when asked about niche scientific subfields (Lu et al., 2022).

Outdated Information: The model's dependence on older datasets can make it unaware of recent developments or changes. LLMs are trained on datasets that, at some point, become outdated. This means that any events, discoveries, or changes post-dating the last training update won't be known to the model. For example, ChatGPT and GPT-4 are both trained on data up to 2021.09 might not be aware of events or advancements after then.

Immemorization: The model does not always retain knowledge from its training corpus. While it's a misconception that LLMs "memorize" data, they do form representations of knowledge based on their training. However, they might not always recall specific, less-emphasized details from their training parameters, especially if such details were rare or not reinforced through multiple examples. For example, ChatGPT has pretrained with Wikipedia, but it still fail in answering some questions from NaturalQuestions (Kwiatkowski et al., 2019) and TriviaQA (Joshi et al., 2017), which are constructed from Wikipedia (Wang et al., 2023a).

Forgetting: The model might not retain knowledge from its training phase or could forget prior knowledge as it undergoes further training. As models are further fine-tuned or trained on new data, there's a risk of "catastrophic forgetting" (Chen et al., 2020; Goodfellow et al., 2015; Wang et al., 2022; Zhai et al., 2023) where they might lose certain knowledge they were previously know. This is a well-known challenge in neural network training, where networks forget previously learned information when exposed to new data, which also happens in large language models (Luo et al., 2023b).

Reasoning Failure: While the model might possess relevant knowledge, it can sometimes fail to reason with it effectively to answer queries. Even if an LLM has the requisite knowledge to answer a question, it might fail to connect the dots or reason logically. For instance, ambiguity in the input (Liu et al., 2023a) can potentially lead to a failure in understanding by LLMs, consequently resulting in reasoning errors. In addition, Berglund et al. (2023) find the LLM traps in the reversal curse, for example, it knows that A is B's mother but fail to answer who is B's son. This is especially evident in complex multi-step reasoning tasks or when the model needs to infer based on a combination of facts (Kotha et al., 2023; Tan et al., 2023b).

4.2.2 Retrieval-level Causes

The retrieval process plays a pivotal role in determining the accuracy of LLMs' responses, especially in retrieval-

augmented settings. Several factors at this level can lead to factual errors:

Insufficient Information: If the retrieved data doesn't provide enough context or details, the LLM might struggle to generate a factual response. This can result in generic or even incorrect outputs due to the lack of comprehensive evidence.

Misinformation Not Recognized by LLMs: LLMs can sometimes accept and propagate misinformation present in the retrieved data. This is especially concerning when the model encounters knowledge conflicts, where the retrieved information contradicts its pre-trained knowledge, or multiple retrieved documents contradict each other Li et al. (2015). For instance, Longpre et al. (2021) observed that the more irrelevant the evidence, the more likely the model is to rely on its intrinsic knowledge. Recent studies, such as Pan et al. (2023c), have also shown that LLMs are susceptible to misinformation attacks within the retrieval process.

Distracting Information: LLMs can be misled by irrelevant or distracting information in the retrieved data. For example, if the evidence mentions a "Russian movie" and "Directors", the LLM might incorrectly infer that "The director is Russian". Luo et al. (2023a) highlighted this vulnerability, noting that LLMs can be significantly impacted by distracting retrieval results. They further proposed instruction tuning as a potential solution to enhance the model's ability to sift through and leverage retrieval results more effectively.

Additionally, when dealing with long retrieval inputs, the models tend to show their best performance when processing information given at the beginning or end of the input context, according to Liu et al. (2023c). In contrast, the models are likely to experience a significant decrease in performance when they are required to extract relevant data from the middle of these extensive contexts.

Misinterpretation of Related Information: Even when the retrieved information is closely related to the query, LLMs can sometimes misunderstand or misinterpret it. While this might be less frequent when the retrieval process is optimized, it remains a potential source of error. For instance, in the ReAct study (Yao et al., 2023a), the rate of errors dropped significantly when the retrieval process was improved.

4.2.3 Inference-level Causes

Snowballing: During the generation process, a minor error or deviation at the beginning can compound as the model continues to generate content. For instance, if an LLM misinterprets a prompt or starts with an inaccurate premise, the subsequent content can veer further from the truth (Varshney et al., 2023; Zhang et al., 2023b).

Erroneous Decoding: The decoding phase is crucial for translating the model's internal representations into human-readable content (Chuang et al., 2023; Massarelli et al., 2020). Mistakes during this phase, whether due to issues like beam search errors or suboptimal sampling strategies, can lead to outputs that misrepresent the model's actual "knowledge" or intention. This can manifest as inaccuracies, contradictions, or even nonsensical statements.

Exposure Bias: LLMs are a product of their training data. If they've been exposed more frequently to certain types of content or phrasing, they might have a bias toward generating similar content, even when it's not the most factual or relevant. This bias can be especially pronounced if the training data has imbalances or if certain factual scenarios are underrepresented (Felkner et al., 2023; Gallegos et al., 2023). The model's outputs, in such cases, reflect its training exposure rather than objective factuality. For example, research by Hossain et al. (2023) suggests that LLMs can correctly identify the gender of individuals who conform to the binary gender system, but they perform poorly when determining non-binary or neutral genders.

5 ENHANCEMENT

This section discusses methods to enhance factuality in LLMs across different phases, including LLM generation, retrieval-augmented generation, inference-phase enhancements, and domain-specific factuality improvements, outlined in Figure 2.

Table 8 provides a summary of enhancement methods and their respective improvements over baseline LLMs. It's essential to recognize that various research papers may employ distinct experimental settings, such as zero-shot, few-shot, or full settings. Consequently, when examining this table, it's important to note that performance metrics for different methods, even when evaluating the same metric on the same dataset, may not be directly comparable.

5.1 On Standalone LLM Generation

When focusing on standalone LLM generation, enhancement strategies can be broadly grouped into three main categories:

(1) Improving Factual Knowledge from Unsupervised Corpora (Sec 5.1.1): This involves refining the training data during pretraining, such as through deduplication and emphasizing informative words (Lee et al., 2022a). Techniques like TOPICPREFIX (Lee et al., 2022b) and sentence completion loss are also explored to enhance this approach.

(2) Enhancing Factual Knowledge from Supervised Data (Sec 5.1.2): Examples in this category include supervised finetuning strategies (Chung et al., 2022; Zhou et al., 2023a) focus on finetuning with labelled data or integrating structured knowledge from such as knowledge graphs (KGs) or making precise adjustments to model parameters (Li et al., 2023d).

(3) Optimally Eliciting Factual Knowledge from the Model (Sec 5.1.3, 5.1.4, 5.1.5): This category encompasses methods like Multi-agent collaboration (Du et al., 2023) and innovative prompts (Yu et al., 2023). Additionally, novel decoding methodologies, such as factual-nucleus sampling, are introduced to further improve factuality (Chuang et al., 2023; Lee et al., 2022b).

Some work (Liu et al., 2023b; Sun et al., 2023d) aim for improve the factuality of large multi-modal models, we choose to not emphasize them.

5.1.1 Pretraining-based

Pretraining plays a pivotal role in equipping the model with the factual knowledge derived from the corpus. By emphasizing strategies during this phase, the model's inherent
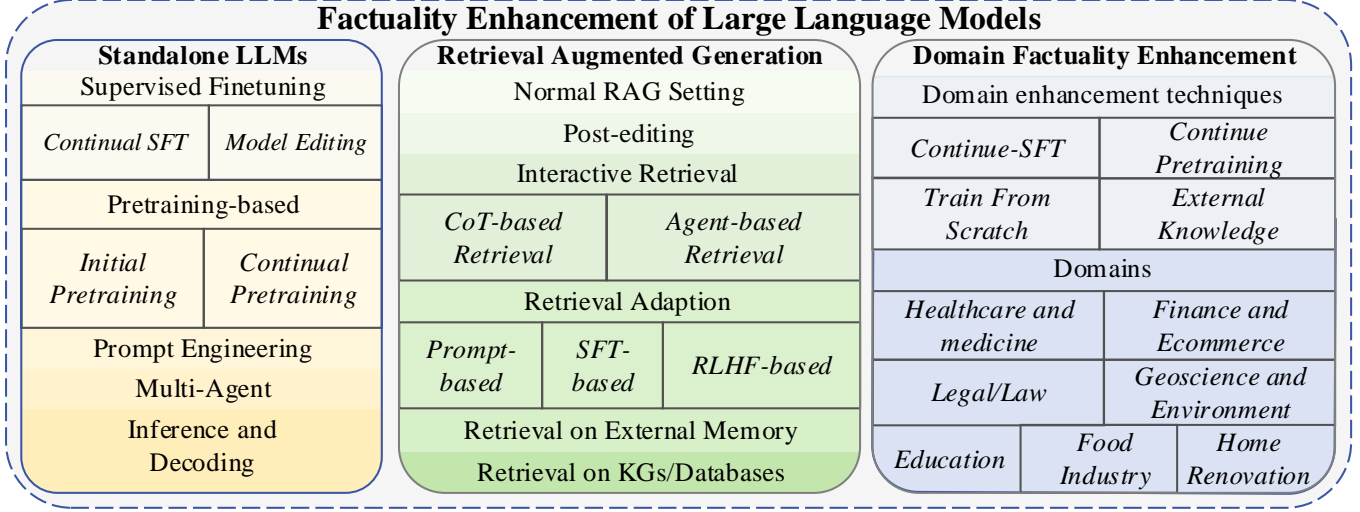
Factuality Enhancement of Large Language Models

Standalone LLMs	Retrieval Augmented Generation	Domain Factuality Enhancement

Standalone LLMs

Supervised Finetuning

Continual SFT	*Model Editing*

Pretraining-based

Initial Pretraining	*Continual Pretraining*

Prompt Engineering

Multi-Agent

Inference and Decoding

Retrieval Augmented Generation

Normal RAG Setting

Post-editing

Interactive Retrieval

CoT-based Retrieval	*Agent-based Retrieval*

Retrieval Adaption

Prompt-based	*SFT-based*	*RLHF-based*

Retrieval on External Memory

Retrieval on KGs/Databases

Domain Factuality Enhancement

Domain enhancement techniques

Continue-SFT	*Continue Pretraining*
Train From Scratch	*External Knowledge*

Domains

Healthcare and medicine	*Finance and Ecommerce*
Legal/Law	*Geoscience and Environment*

Education	*Food Industry*	*Home Renovation*

Fig. 2. An overview of methods to enhance factuality in large language models. This encompasses three primary areas: enhancement techniques for pure LLMs, strategies for retrieval-augmented LLMs, and methods employed by domain-specific LLMs to boost their factual accuracy within their respective domains.

factuality can be significantly enhanced. This approach is particularly crucial for addressing challenges like immemorization and forgetting.

Initial Pretraining: Methods employed during the foundational pretraining of the model.

Lee et al. (2022a) develop two distinct tools for deduplicating training data, addressing the issue of redundant texts and long repeated substrings present in the training sets of current LLMs. These tools effectively reduce the recall of memorized texts in model outputs. Remarkably, they achieve similar or even superior accuracy with fewer training steps.

Sadeq et al. (2023) introduce a modification to the Masked Language Model (MLM) training objective used in the pretraining of LLMs. They discover that high-frequency words do not consistently contribute to the model's ability to learn factual knowledge. To address this, they devise a strategy that encourages the language model to prioritize informative words during unsupervised training. This is achieved by masking tokens more frequently based on their informative relevance. To quantify this relevance, they utilize Pointwise Mutual Information (PMI) (Fano and Hawkins, 1961), positing that words with elevated PMI values, in relation to their adjacent words, are likely to be more informatively pertinent. Experimental results indicate that this innovative approach significantly bolsters the efficacy of pretrained language models across various tasks, including factual recall, question answering, sentiment analysis, and natural language inference, in a closed-book setting.

Continual Pretraining: Iterative pretraining processes that allow the model to progressively refine and update its knowledge base.

Lee et al. (2022b) introduce TOPICPREFIX as a preprocessing method and the sentence completion loss as training objective. Some factual sentences may be unclear when the LM training corpus is chunked, especially when these sentences contain pronouns (e.g., she, he, it). So they prepend TOPICPREFIX (e.g., Wikipedia document name) to sentences in the factual documents to transform each

sentence into an independent factual statement. They also introduce the sentence completion loss, with the aim of enabling the model to capture facts from entire sentences rather than just focusing on the associations between subwords. For implementation, they establish a pivot t for each sentence and require zero-masking for all token prediction losses before t. This pivot is only necessary during the training phase. Experiments show that such methods can further reduce the factual errors than standard factual-domain adaptive training.

5.1.2   Supervised Finetuning

Supervised fine-tuning leverages labeled datasets to refine the model's performance. This approach serves a dual purpose: it imparts specific task or knowledge base-oriented information to the model and addresses challenges like immemorization and forgetting. Several studies, such as Chung et al. (2022); Zhou et al. (2023a), emphasize the pivotal role of supervised fine-tuning in eliciting the inherent knowledge of the base model, subsequently enhancing its reasoning capabilities.

Continual SFT: A cyclic fine-tuning approach, where the model undergoes consistent refinement using sequential sets of labeled data.

Moiseev et al. (2022) investigate how to inject structured knowledge from a knowledge graph into LLMs. The approach involves directly training T5 using triplets containing relationship knowledge. Previous methods often describe the triplets using prompts and then train LLMs with masked language model task, but some triplets are not easy to describe. They compare three knowledge-enhancement fine-tuning methods, which are MLM training on the C4 corpus (Raffel et al., 2020b), masking the subject or object in KG triplets, and masking the subject or object in the KELM corpus (Agarwal et al., 2021). Experiments show that effectiveness of the latter two methods achieved better exact match scores in closed-book QA tasks (Jiang et al., 2019; Joshi et al., 2017; Kwiatkowski et al., 2019; Welbl et al., 2018). This demonstrates that training directly based on KG triplets is an effective way to inject knowledge into the model

TABLE 8
Performance of Select Factuality Enhancement Methods. The table displays performance metrics on various datasets for both baseline models and their enhanced counterparts, denoted with a →. Due to space constraints, only a subset of datasets, metrics, and models from each work is presented.

Reference	Dataset	Metrics	Baselines → Theirs	Dataset	Metrics	Baselines → Theirs
Li et al. (2022c)	NQ	EM	34.5 → 44.35 (T5 11B)	GSM8K	ACC	77.0 → 85.0 (ChatGPT)
Yu et al. (2023)	NQ	EM	20.9 → 28.0 (InstructGPT)	TriviaQA	EM	57.5 → 59.0 (InstructGPT)
				WebQA	EM	18.6 → 24.6 (InstructGPT)
Chuang et al. (2023)	FACTOR News	ACC	58.3 → 62.0 (LLaMa-7B)	FACTOR News	ACC	61.1 → 62.5 (LLaMa-13B)
	FACTOR News	ACC	63.8 → 65.4 (LLaMa-33B)	FACTOR News	ACC	63.6 → 66.2 (LLaMa-65B)
	FACTOR Wiki	ACC	58.6 → 62.2 (LLaMa-7B)	FACTOR Wiki	ACC	62.6 → 66.2 (LLaMa-13B)
	FACTOR Wiki	ACC	69.5 → 70.3 (LLaMa-33B)	FACTOR Wiki	ACC	72.2 → 72.4 (LLaMa-65B)
	TruthfulQA	%Truth * Info	32.4 → 44.6 (LLaMa-13B)	TruthfulQA	%Truth * Info	34.8 → 49.2 (LLaMa-65B)
Li et al. (2022b)	TruthfulQA	%Truth * Info	32.4 → 44.4 (LLaMa-13B)	TruthfulQA	%Truth * Info	31.7 → 36.7 (LLaMa-33B)
	TruthfulQA	%Truth * Info	34.8 → 43.4 (LLaMa-65B)			
Li et al. (2023d)	NQ	ACC	46.6 → 51.3 (LLaMA-7B)	TriviaQA	ACC	89.6 → 91.1 (LLaMA-7B)
	MMLU	ACC	35.7 → 40.1 (LLaMA-7B)	TruthfulQA	%Truth * Info	32.5 → 65.1 (Alpaca)
	TruthfulQA	%Truth * Info	26.9 → 43.5 (LLaMa-7B)	TruthfulQA	%Truth * Info	51.5 → 74.0 (Vicuna)
Cohen et al. (2023b)	LAMA	F1	50.7 → 80.8 (ChatGPT)	TriviaQA	F1	56.2 → 82.6 (ChatGPT)
	NQ	F1	60.6 → 79.1 (ChatGPT)	PopQA	F1	65.2 → 85.4 (ChatGPT)
	LAMA	F1	42.5 → 79.3 (GPT-3)	TriviaQA	F1	46.7 → 77.2 (GPT-3)
	NQ	F1	52.0 → 78.0 (GPT-3)	PopQA	F1	43.7 → 77.4 (GPT-3)
Weller et al. (2023)	TriviaQA	QUIP	31.6 → 33.6 (ChatGPT)	NQ	QUIP	32.8 → 34.3 (ChatGPT)
	HotpotQA	QUIP	28.3 → 29.2 (ChatGPT)	ELI5	QUIP	24.1 → 26.5 (ChatGPT)
	TriviaQA	EM	77.8 → 78.8 (ChatGPT)	NQ	EM	32.9 → 34.8 (ChatGPT)
	HotpotQA	F1	35.7 → 36.6 (ChatGPT)	ELI5	R-L	22.7 → 21.7 (ChatGPT)
Dhuliawala et al. (2023)	MultiSpanQA	F1	39.0 → 48.0 (LLaMA 65B)	-	FactScore	55.9 → 71.4 (LLaMA 65B)
				-	Avg. # facts	16.6 → 12.3 (LLaMA 65B)
Yao et al. (2023a)	HotpotQA	EM	28.7 → 35.1 (PaLM-540B)	FEVER	ACC	57.1 → 62.0 (PaLM-540B)
Jiang et al. (2023b)	2WikiMultihopQA	EM	28.2 → 51 (ChatGPT)	StrategyQA	EM	72.9 → 77.3 (ChatGPT)
	WikiAsp	UniEval	47.1 → 53.4 (ChatGPT)	ASQA	EM	33.8 → 41.3 (ChatGPT)
				ASQA-hint	EM	40.1 → 46.2 (ChatGPT)
Izacard et al. (2022)	MMLU	ACC	42.4 → 56.3 (T5-770M)	MMLU	ACC	50.4 → 59.9 (T5-3B)
				MMLU	ACC	54 → 65.8 (T5-11B)
Shi et al. (2023)	MMLU	ACC	68.3 → 73.2 (Codex)	NQ	EM	40.6 → 45.5(Codex)
				TriviaQA	EM	73.6 → 77.3 (Codex)

Sun et al. (2023c) introduce negative samples and contrastive learning to supervised finetuning process with MLE loss. The sources of these negative samples are either from the Knowledge Graph or are generated by large language models. While traditional contrastive learning can only function at the token or sentence level, the intrinsic value of span information cannot be overlooked, they employ Named Entity Recognition to extract this critical span data. The training process incorporates a blend of MLE loss, standard contrastive learning loss, and span-based contrastive learning loss, with the parameters finely tuned to optimize results. Experiments indicate that the method delivers performance comparable with SOTA KB-based methods but offers significant benefits in efficiency and scalability.

Yang et al. (2023b) presents a development framework for Knowledge Graph-enhanced Large Language Models (KGLLMs), drawing from established technologies. They detail various enhancement strategies, including a before-training enhancement that refines input quality by integrat-

ing factual data; a during-training enhancement that synergizes textual and structural knowledge, utilizing tools like graph neural networks and attention mechanisms; multi-task learning which focuses on knowledge-guided pre-training tasks to bolster the factual knowledge acquisition of LLMs; and a post-training enhancement that fine-tunes LLMs for domain-specific tasks using knowledge-rich data. Additionally, the significance of prompt learning in LLMs is highlighted, emphasizing the importance of choosing appropriate prompt templates. They also suggest knowledge graphs as a valuable resource for crafting these templates to harness domain-specific knowledge.

Model Editing: Instead of directly finetuning the model, model edit is a more precise approach to enhance the model's factuality. By editing specific areas that are related to the fact, the model can correctly express that fact without compromising other unrelated knowledge. Current editing methods (Yao et al., 2023b) can be categorized into weight-preserved and weight-modified paradigms. When

modifying the model's weight, KN (Dai et al., 2022) and ROME (Meng et al., 2022) first analyze representations to locate those underlying factual errors and then directly update the relevant weights. Meanwhile, KE (De Cao et al., 2021) and MEND (Mitchell et al., 2022a) employ a hypernetwork to learn the necessary weight changes. While effective, the robustness and generalization of directly updating weights remains an open question.

Li et al. (2023d) present a technique known as Inference-Time Intervention (ITI) designed to boost the factual accuracy of large language models. ITI adjusts model activations during inference to enhance the truthfulness of responses. This method, categorized under activation editing, is both adjustable and less intrusive, setting it apart from weight editing techniques. Drawing inspiration from prior research, they utilize steering vectors, which are proven effective for style transfer, to guide model activations. They detail a multi-step model selection process involving the calibration of intervention strength hyperparameters, pinpointing truth-telling related heads, and determining their truth-telling directions. The TruthfulQA dataset serves as the foundation for training and validation, with a rigorous 2-fold cross-validation employed to prevent data leakage. Experiments on the TruthfulQA benchmark demonstrate that ITI improves Alpaca's (Taori et al., 2023) truthfulness from 32.5% to 65.1%.

5.1.3 Multi-Agent

Engaging multiple models in a collaborative or competitive manner, enhancing factuality through their collective prowess, helps in the immemorization and reasoning failure problems.

Du et al. (2023) propose an approach to enhance the performance of language models by treating different LLMs as intelligent agents engaged in multi-agent debates. In this method, multiple instances of language models present and debate their respective answers and reasoning processes, ultimately reaching a consensus on the final answer after multiple rounds of debate. If the debate's answers fail to converge, prompts are modified to reduce the stubbornness of the two agents. This approach has been demonstrated to significantly improve mathematical and reasoning abilities across various tasks while enhancing the factual accuracy of generated content. Moreover, this method can be directly applied to existing black-box models, making it applicable to all research tasks using the same prompts.

Cohen et al. (2023b) develop a fact-checking mechanism. Drawing parallels to a scenario where a witness is interrogated for the veracity of their claims, they utilize a LLM to gather statements from a QA dataset, which are either factually correct or incorrect. During the statement generation, the model is provided with a golden answer, prompting it to produce both accurate and inaccurate statements, inherently labeling each statement. For every QA pair, another LLM, acting as an interrogator, generates a series of questions. A separate LLM, playing the role of the respondent, answers these queries. This iterative questioning and answering continues until the interrogator is satisfied, culminating in a conclusion. The authors conduct experiments on datasets like LAMA (Petroni et al., 2019a), PopQA (Mallen et al., 2023), NaturalQA (Kwiatkowski et al., 2019), and TriviaQA

(Joshi et al., 2017). The precision is the portion of incorrect claims, out of the claims rejected by the examiner and the recall is the portion of incorrect claims rejected by the examiner, out of all the incorrect claims. Measured by the F1 score, the LM vs LM method consistently outperforms the baseline by a significant margin, ranging from ten to over twenty points across all datasets.

5.1.4 Novel Prompt

Introducing innovative or tailored prompts to extract more factual and precise responses from the LLM, can better assist the model elicit the knowledge in its parameters and improve the reasoning ability.

Yu et al. (2023) introduce a novel approach called Generate-then-Read (GENREAD). Document retrievers are replaced by LLM generators. In this article, the LLM is prompted to generate multiple contextual documents on a given question. The authors clustered these document embeddings, and sampled documents from different clusters to ensure diversity of contextual documents. With these generated in-context demonstrations, LLMs achieved better results on knowledge-intensive tasks than retrieving from external corpus such as Wikipedia.

Weller et al. (2023) introduce a metric called QUIP-Score to measure the dependency on pre-trained data. They index Wikipedia to swiftly determine the dependency of an LLM's response. By using specific prompts, like "Based on evidence from Wikipedia:" they aim to evoke the LLM's recall of content from its training dataset. In addition to grounding prompts, they also introduce anti-grounding prompts to encourage the LLM to respond without referencing its training data, for instance, "Respond without using any information from Wikipedia." The motivation behind this approach is the belief that guiding the LLM to reference more of the knowledge it acquires during pre-training can reduce the generation of incorrect information. To quantify this grounding, they propose the QUIP-Score metric to gauge the similarity between the model's generated content and the most relevant content in Wikipedia. In their experiments conducted on datasets like TQ, NQ, HotpotQA, and ELI5, the results show that while adding the said prompt doesn't significantly improve traditional QA metrics, it notably boosts the scores on the QUIP metric.

Khot et al. (2023) present Decomposed Prompting. Complex tasks are broken down into multiple simpler tasks via prompting and then it can be addressed by task-specific LLMs. For instance, for tasks involving extremely long input sequences, this technique systematically decomposes the input into shorter sequences for individual processing. Notably, the authors observed that when combined with a retrieval module, this approach significantly enhances performance on open domain multi-hop QA tasks.

Dhuliawala et al. (2023) present Chain-of-Verification (CoVe) to reduce factual errors.The CoVe strategy involves the model initially crafting a response, subsequently formulating verification queries to assess its initial draft, independently responding to these queries to maintain unbiased answers, and ultimately producing a validated reply. The CoVe method encompasses four pivotal stages: 1) Drafting an initial reply based on a query using the LLM, 2) Generating verification questions from the query and initial answer

to pinpoint potential errors, 3) Responding to each verification question and comparing these answers with the initial response to detect discrepancies, and 4) If discrepancies are found, producing a revised answer that integrates the verification outcomes. The entire procedure is executed by prompting the same LLM differently to achieve the intended results. Experiments show that Chain-of-Verification can reduce errors in diverse tasks, including list-based questions from Wikidata (Vrandečić and Krötzsch, 2014), closed book MultiSpanQA (Li et al., 2022a) and longform text generation (Min et al., 2023).

5.1.5 Decoding

Decoding methodologies, such as beam search and nucleus sampling, play a crucial role in directing the model to produce outputs that are both factual and coherent. By refining the decoding process, challenges like snowballing errors or erroneous decoding, as detailed in Sec 4.2.3, can be effectively addressed.

Lee et al. (2022b) propose a new decoding sampling algorithm called factual-nucleus sampling that achieves a better trade-off between generation quality and factuality when compared to prevailing decoding algorithms. They postulate that the randomness of sampling is more harmful to factuality when applied to the generation of the latter portion of a sentence as opposed to its initial segment. So the factual-nucleus sampling algorithm, an adaptation of nucleus sampling, can dynamically adjust the 'nucleus' probability throughout the generation of each sentence, progressively reducing randomness with each successive generation step. The ω-bound parameter is provided to prevent the p-value from becoming too small and hurting diversity. Experimental results show that a factual-nucleus sampling algorithm can improve the factuality of generation while maintaining generation quality, e.g., diversity and repetition.

Chuang et al. (2023) propose "Decoding by Contrasting Layers" to mitigate these hallucinations. This approach leverages the differences in logits obtained from projecting the later layers versus earlier layers to the vocabulary space, taking advantage of the known localization of factual knowledge in LLMs. The results of this study demonstrate that DoLa consistently enhances the truthfulness of LLM-generated content across various tasks, such as multiple-choice and open-ended generation tasks, showcasing its potential to significantly improve the reliability of LLMs in generating accurate and truthful facts.

5.2 On Retrieval-Augmented Generation

Retrieval-Augmented Generation (RAG) has emerged as a widely adopted approach to address certain limitations inherent to standalone LLMs, such as outdated information and the inability to memorize (Chase, 2022; Liu, 2022). These challenges are elaborated upon in Sec 4.2.1. Yet, while RAG offers solutions to some issues, it introduces its own set of challenges, including the potential for insufficient information and the misinterpretation of related data, as detailed in Sec 4.2.2. This subsection delves into various strategies devised to mitigate these challenges. Within the realm of retrieval-augmented generation, enhancement techniques can be broadly categorized into several pivotal areas:

(1) The Normal Setting of Utilizing Retrieved Text for Generations (Sec 5.2.1):

(2) Interactive Retrieval and Generation (Sec 5.2.2): Examples here include the integration of Chain-of-Thoughts steps into query retrieval (He et al., 2022) and the use of an LLM-based agent framework that taps into external knowledge APIs (Yao et al., 2023a).

(3) Adapting LLMs to the RAG Setting (Sec 5.2.3): This involves methods like the one proposed by Peng et al. (2023), which combines a fixed LLM with a plug-and-play retrieval module. Another notable approach is REPLUG (Shi et al., 2023), a retrieval-augmented framework that treats the LLM as a black box and fine-tunes retrieval models using language modeling scores.

(4) Retrieving from Additional Knowledge Bases (Sec 5.2.5 and Sec 5.2.4): This category includes methods that retrieve from external parametric memories (Chen et al., 2023a) or knowledge graphs (Zhang et al., 2023d) to enhance the model's knowledge base.

5.2.1 Normal RAG Setting

Workflow of Normal RAG Setting: A normal RAG setting works by retrieving external data and passing it to a LLM during the generation phase. We follow the framework proposed by LlamaIndex (Liu, 2022) and LangChain (Chase, 2022) to decouple the process into the following modules and steps:

(1) Document loaders are used to load documents from various sources. These loaders can fetch different types of documents (HTML, PDF, code) from various locations (private s3 buckets, public websites).

(2) Document transformers are employed to extract relevant parts of the documents. This may involve splitting or chunking large documents into smaller chunks. Different algorithms for this task are employed, optimized for specific document types like code or markdown.

(3) Text embedding models are employed to capture the semantic meaning of the text.

(4) Vector stores are employed to efficiently store and search the embeddings.

(5) Retrievers, such as the Parent Document Retriever, Self Query Retriever, and Ensemble Retriever, are used to retrieve the data from the database.

Here, the Parent Document Retriever allows for the creation of multiple embeddings per parent document to retrieve smaller chunks while maintaining a larger context. The Self Query Retriever separates the semantic part of a query from other metadata filters, allowing for more accurate retrieval. The Ensemble Retriever enables the retrieval of documents from multiple sources or using different algorithms.

Borgeaud et al. (2022) suggest scaling the size of the text database for retrieval as a complementary path to scaling language models. There is a pre-collected text database with a total of over 5 trillion tokens. These chunks are stored in the form of key-value pairs, with each chunk as a unit, and similarity retrieval is performed on k-nearest neighbours from key-value database using the L_2 distance on Bert embeddings. The input sequence is splited into chunks, Retrieval Transformer (Retro) model retrieves text similar to the previous chunk to improve the predictions in the

current chunk. The model calculate cross-attention between the input text and the retrieved text chunks to generate better answers. With only 25× fewer parameters of GPT-3, its performance on Pile is quite comparable.

Lazaridou et al. (2022) present a method that capitalizes on the few-shot capabilities of large-scale language models to enhance their grounding in factual and current information. Drawing from semi-parametric language models, the approach conditions LMs using few-shot prompts based on data sourced from Google Search. For any given query, the method retrieves pertinent documents from the web, extracts the top 20 URLs, and processes them to obtain clear text. These documents are segmented into paragraphs, and the most relevant ones are chosen using TF-IDF based on their similarity to the query. The LMs are then conditioned using few-shot prompts that incorporate the retrieved paragraphs. This k-shot prompting technique is augmented with an evidence paragraph, creating a prompt structure that encompasses evidence, query, and response. The method also involves generating multiple answers from the model and reranking them using different probabilistic factorizations. The experimental results indicate that by conditioning on retrieved evidence, the 7B Gopher LM (Rae et al., 2022a) surpassed the performance of the 280B Gopher LM, with relative improvements reaching up to 30% on the NQ (Kwiatkowski et al., 2019) dataset.

5.2.2   Interactive Retrieval

While retrieval systems are designed to source relevant information, they may occasionally fail to retrieve accurate or comprehensive data. Additionally, LLMs might struggle to recognize, or even be misled by, the retrieved content, as detailed in Sec 4.2.2. Implementing an interactive retrieval mechanism can address these challenges, aiding in sourcing more appropriate information and guiding the LLM towards improved content generation. In this subsubsection, we explore methods that employ the Chain-of-Thoughts and Agents mechanisms to achieve effective interactive retrieval.

CoT-based Retrieval: In recent studies, there is a growing interest in integrating Chain-of-Thoughts (Wei et al., 2022) steps into query retrieval. He et al. (2022) introduce a method that generates multiple reasoning paths and their corresponding predictions for each query. This process involves retrieving relevant knowledge from external sources like Wikidata (Vrandečić and Krötzsch, 2014), WordNet (Miller, 1992), and ConceptNet (Speer et al., 2017). The faithfulness of each reasoning path determines based on entailment scores, contradiction scores, and MPNet similarities (Song et al., 2020) with the retrieved knowledge. The prediction with the highest faithfulness score is chosen as the final result. This method demonstrates superior performance in tasks such as commonsense reasoning (Geva et al., 2021b), temporal reasoning (Jia et al., 2018), and tabular reasoning (Gupta et al., 2020b), outperforming the baseline CoT reasoning and self-consistency methods (Wang et al., 2023e). On a related note, Trivedi et al. (2023) propose IRCoT, an innovative retrieval technique that interweaves the CoT process. In this approach, each generated sentence during the CoT combines with the question to form a retrieval query. The subsequent reasoning step then produces by the

Language Model using both the retrieval results and the prior reasoning. This interleaving method finds to enhance the performance of both retrieval and CoT in Open-domain QA. Experiments proves that this is beneficial for models of different sizes, including GPT-3(175B) (Brown et al., 2020) and Flan-T5 (Chung et al., 2022) families.

FLARE (Jiang et al., 2023b) is a dynamic solution to address the limitations of previous RAG works which either retrieve only once at the onset of generation or do so based on fixed intervals. Single retrievals are insufficient for long-form generation due to the evolving information needs during the process, and fixed intervals for previous token queries can be inappropriate, FLARE determines "when and what to retrieve" dynamically. The decision of "when" is based on whether the current sentence contains a token with a generation probability below a set threshold. If it doesn't, the sentence is accepted and moves to the next generation step; otherwise, retrieval augmented generation occurs. For the "what", the current sentence is used as a query. To address the challenge of low-probability tokens affecting retrieval accuracy, two solutions are proposed: masking low-probability tokens, and using LLM for generation of these tokens as queries. Testing on tasks like Multihop QA (Ho et al., 2020), Commonsense Reasoning (Geva et al., 2021a), Long-form QA (Stelmakh et al., 2022), and Open-domain Summarization (Hayashi et al., 2021) using GPT3.5, results show FLARE outperforms baselines, with both query generation methods showing comparable performance.

Agent-based Retrieval: Using an LLM-based agent framework that leverages external knowledge APIs as tools or requesting such APIs as actions.

Yao et al. (2023a) present a new framework named ReAct that integrates Chain-of-Thoughts reasoning with action. Through in-context learning, the LLM's CoT output is transformed into descriptions of reasoning processes and action behaviors. Subsequently, these action descriptions are standardized and executed, with the results being incorporated into the next prompt. In terms of results, for tasks like Fact checking and QA, while CoT has 14% of its correct answers containing incorrect reasoning steps or facts, ReAct only has 6%. In the incorrect answers, 56% of CoT's errors were due to errors in reasoning steps or facts, whereas ReAct has no factual errors. However, ReAct have 23% of its errors resulting from search errors.

Shinn et al. (2023) propose a prompt engineering framework named Reflextion to enable LLMs to reflect on and correct previous errors. They use linguistic feedback to strengthen an agent's actions instead of adjusting model weights. Specifically, the Reflextion Agent, an LLM, first interacts with its environment by generating an "Action" via ReAct (Yao et al., 2023a) or Chain-og-Thoughts (Wei et al., 2022) in few-shot scenarios, which results in an "Observation". This "Observation", whether a reward, error message, or natural language feedback, provides insights on the Agent's current "Action". When the Agent receives a failure signal, it triggers a self-reflection mechanism, utilizing the LLM, to summarize the reasons for the failure into its Memory module, creating a "long-term memory". On subsequent generations, the Agent can review all past reflection memories to prevent mistakes. Experimental findings indicate that Reflextion achieves a 10-20% performance

increase over the baseline methods ReACT and CoT on datasets like AlfWorld (Shridhar et al., 2021), HotPotQA (Yang et al., 2018), and HumanEval (Chen et al., 2021a).

Varshney et al. (2023) introduce a comprehensive framework aimed at reducing factual inaccuracies. They use models to recognize entities and generate questions, we recognize these models as tools for the LLM-based agent. During the generation process, pivotal concepts encompassing names, geographical locales, and temporal references are ascertained within the contextual sentence employing entity extraction, keyword distillation, or directives to the LLM. The logit output values corresponding to these discerned concepts act as surrogates for confidence estimates. Should these values fall beneath a predetermined threshold, the mechanism then fetches a pertinent document to corroborate the generated information. The query methodology employed for such a retrieval hinges on posing a binary (Yes/No) query to the LLM. In scenarios where the validation is unsuccessful, the framework directs the model to rectify the erroneous output, either by omission or by substitution, drawing upon the knowledge from the consulted document. Empirical evaluations, specifically in the domain of article generation, underscore the efficacy of the delineated approach. Notably, factual error rates exhibited by GPT-3 witnessed a substantial decline, from 47.5% to a mere 14.5%, when subjected to this methodology. The diagnostic facet of their approach manifests an 80% recall, and the rectification mechanism adeptly rectifies 57.6% of the factually incorrect outputs that were accurately pinpointed.

5.2.3 Retrieval Adaptation

Recent research (Ren et al., 2023; Wang et al., 2023a) has highlighted that merely using the retrieved information in LLMs doesn't always enhance their ability to answer factual questions. This underscores the importance of enabling LLMs to better adapt to the retrieved data to produce more accurate content. In this section, we delve into various strategies that facilitate this adaptation. Specifically, we explore three methodological approaches: prompt-based methods, SFT-based methods, and RLHF-based methods.

Prompt-based: Leveraging prompts to navigate the retrieval process, ensuring the extraction of pertinent and factual data.

Peng et al. (2023) introduce the LLM-Augmenter system, a system using a fixed LLM combined with a plug-and-play retrieval module to help the LLM perform better in tasks that are particularly sensitive to factual errors. This system enhances its performance by enabling the LLM to use a series of modules (e.g. allowing the LLM to interact with external knowledge) to assist the LLM in generating results grounded in evidence. And they use automated feedback generated by utility functions (e.g., the factuality score of a LLM-generated response) to modify LLM's candidate response options. The author evaluates the system's performance on information-seeking dialog (Galley et al., 2019b) and Wiki QA (Chen et al., 2021b) and experiments show that the system can significantly reduce ChatGPT's errors without sacrificing the fluency and informativeness of the generated content.

SFT-based: Optimizing the LLM or retrieval system through training to enhance the alignment between generation tasks and the retrieved content.

Izacard et al. (2022) introduce a comprehensive architecture named ATLAS which is composed of the Contriever (Izacard et al., 2021) retriever of the dual-encoder architecture and the T5 (Raffel et al., 2020a) language model with Fusion-in-Decoder (Izacard and Grave, 2021b). The training objectives for the retriever consist of four components: Attention Distillation (Izacard and Grave, 2021a), where the retriever is trained on the average attention scores from the language model for each article; End-to-end training of Multi-Document Reader and Retriever (EMDR2) (Singh et al., 2021), which involves using the query and the top-K retrieved articles from the current retriever as input and loss computation against the standard answers to train the retriever; Perplexity Distillation (PDist), where the retriever is trained to predict how much the perplexity of standard answers would improve for each document; Leave-one-out Perplexity Distillation (LOOP), which trains the retriever to predict how much worse the prediction of the language model gets when removing a document from the top-K results. The LM's training objectives consist of three parts: prefix language modeling, masked language modeling, and title to section generation. Besides, they optimize and accelerate retriever training using techniques such as full index update, Re-ranking, and Query-side fine-tuning. Atlas achieves notable accuracy on Natural Questions using only 64 examples, outperforming a 540B model with 50x fewer parameters.

Shi et al. (2023) introduce REPLUG, a retrieval-augmented framework that considers the LLM as a black box, freezes its parameters and tunes retrieval models with supervision signals using language modeling scores. In this framework, the input context and the document are encoded through the dual encoder architecture, cosine similarity is then calculated to retrieve related documents. The likelihood for each retrieved document and the language model scores are computed. Then they can update the retrieval model parameters by minimizing the KL divergence between retrieved document likelihood and the language model"s score distribution. Ablation experiments demonstrate that this method significantly improves the performance of the original language models and the improvements are not coming from ensembling random documents.

Luo et al. (2023a) focus on utilizing instruction-tuning to denoise the retrieval results. They gather retrieval outcomes from various search APIs and domains, leading to the creation of a new search-grounded dataset. This dataset encompasses instructions, grounding information, and responses. Notably, it includes both pertinent results and those that are unrelated or disputed. The model needs to learn to ground on useful search results. After fine-tuning the LLaMa-7B model on this dataset, the resulting model, named SAIL-7B, exhibits superior performance in transparency-sensitive tasks such as open-ended QA and fact-checking.

RLHF-based: Menick et al. (2022a) use reinforcement learning from human preferences (RLHP) to train a 280 billion parameter model named GopherCite that generate answers along with high quality supporting evidence. They firstly collect data from existing models and have it rated

by humans. The data is used for fine-tuning and reward model training. A supervised fine-tuning model is trained to produce accurate quotes with proper syntax. A reward model is created to rank model outputs based on overall quality. Finally, a reinforcement learning policy is optimized to align model behavior with human preferences, improving quoting performance. The model may decline to answer when the reward model score is too low. According to human evaluation, the model achieves better supported and plausible rating on the subset of Natural Questions dataset (Kwiatkowski et al., 2019) than the previous SOTA (FiD-DPR) (Izacard and Grave, 2021b).

5.2.4 Retrieval on External Memory

Currently, most LLMs enhance their factuality by retrieving knowledge in the form of text snippets from external storage and incorporating them into the context. Some researchers are exploring the storage of knowledge in non-textual forms and integrating this knowledge into models through specialized methods.

Li et al. (2022c) store knowledge in the form of key-value pairs in memory. The key is obtained by encoding knowledge using a Doc Retrieval Embedder, while the value is encoded using a Transformer encoder. Similar to traditional retrieval-based LLMs, the model encodes the input using a Query retrieval embedder and retrieves knowledge from memory. The retrieved value is then integrated into the model's multi-head attention layer through cross-attention to enhance factuality.

G-MAP (Wan et al., 2022) does not explicitly store knowledge in storage but uses a general domain PLM as external memory. To mitigate catastrophic forgetting during adaptive pretraining, G-MAP introduces a frozen-parameter general domain PLM (PLM-G) during the fine-tuning of the domain-specific PLM (PLM-D). During fine-tuning, the input is provided to both PLM-G and PLM-D. The hidden states from each layer of PLM-G are stored in a cache, and a Memory-Augmented Strategy is used to extract hidden states from certain layers, which are then concatenated and integrated into PLM-D's Memory-Augmented Layer. The study also compared four Memory-Augmented Strategies, with Chunk-based Gated Memory Transfer performing the best.

Fine-tuning methods and some model editing methods store new knowledge in new model parameters through continual pretraining. The difference is that fine-tuning methods store many pieces of knowledge in a matrix parameter, while model editing establishes a new neuron for each piece of knowledge.

Houlsby et al. (2019) propose to add Adapter modules to fine-tune pre-trained deep learning models on a new task with minimal changes to the original model by inserting small, trainable modules between existing layers. During fine-tuning, the main body of the pre-trained model is frozen, and the Adapter module learns knowledge specific to downstream tasks. The Adapter method reduces the computational requirements for model fine-tuning while enhancing the model's factuality for specific domains. However, the addition of the Adapter module also increases the overall parameter count of the model, somewhat reducing the model's inference performance.

To enhance the model's understanding of entities and thereby improve factuality, KALA (Kang et al., 2022), EaE (Févry et al., 2020), and Mention Memory (de Jong et al., 2022) store encoded entities in external memory. During generation, the retrieved entity embeddings are integrated into the model layers.

Kang et al. (2022) introduced KALA to reduce overhead and catastrophic forgetting during Adaptive Training. KALA not only establishes a memory for entities and their encodings but also uses a KG to store relationships between these entities. For a given mention in the input, the corresponding entity is first determined. Based on the KG, the encoding of this entity and its neighboring entities is retrieved from memory. Through GNN weighted aggregation, the encoding for this mention's corresponding entity is obtained. Finally, the Knowledge-conditioned Feature Modulation (KFM) is introduced in the model layer, integrating the encoding result into the representation of all tokens involved in the mention.

Févry et al. (2020) introduce mention detection, entity linking, and MLM during model training. The model queries the top 100 entity embeddings from the entity storage that is closest to the current mention and integrates them using attention.

de Jong et al. (2022)'s TOME model is an improvement over the EaE model. Instead of storing entity embeddings in memory, TOME stores the embeddings of entity mentions. For marked entity mentions in the input, TOME retrieves all related entity mention embeddings from memory and integrates them into the model through a memory attention layer.

Similar to the three methods mentioned above, knowledge plugin (Zhang et al., 2023g) also introduce entity-related knowledge. However, instead of integrating the knowledge directly into the model layers, they utilize a pretrained mapping network. This network maps the entity embeddings to the token embedding space of the Pretrained Language Model (PLM). Ultimately, the mapped entity embeddings are injected at the input embedding level, facilitating the knowledge insertion process.

To further tackle the factual correction task (Thorne and Vlachos, 2021), Gao et al. (2023a) explore the integration of LLMs, such as GPT-3, with search engines to improve their precision and memory. The goal is to use search engines to search for evidence and correct sentences generated by LLMs. The proposed method RARR is that, for each input sentence, a set of questions is generated, and web pages are searched to verify the consistency of information with the input sentence. The paper evaluates the modifications based on attribution and preservation criteria, with both manual and automatic verification methods. The primary evaluation metric is F1, considering both attribution and preservation aspects to assess the effectiveness of the approach in enhancing LLM-generated sentences.

Chen et al. (2023a) is a follow-up work to (Gao et al., 2023a) and (Thorne and Vlachos, 2021). Similar to EFEC, this paper fine-tunes a T5 model to serve as an editor, but it introduces negative samples during fine-tuning. The model PURR is trained to take user questions, perform Google searches to retrieve the top 5 web page summaries (used as positive samples), and generate noise by replacing some

content in these positive samples using the language model. A sequence-to-sequence model is then trained to correct the noisy sentences back to their correct versions. This approach differs from EFEC, which used a mask-and-fill approach. PURR represents an improvement over EFEC, focusing on directly training a language model to edit incorrect sentences into correct ones using Google search for generating positive samples, ultimately leading to increased F1 scores.

5.2.5 Retrieval on Structured Knowledge Source

We discuss studies that retrieves on structured repositories, such as knowledge graphs and databse, to source factual data during generation in this subsubsection.

Zhang et al. (2023d) utilize knowledge graphs (KG) for retrieval to tackle factual errors. They observe that there can be inconsistencies between a user's request and the content in the KG. For instance, when a user mentions a full name, the KG might only have its abbreviation, leading to imperfect retrieval results. To rectify this, they propose a method to rephrase the user's request. Their approach involves generating an SQL query based on the user's input and the database metadata using a LLM. And then they query the database and ask the LLM to identify which entity in the sentence corresponds to an entity in the database, thereby creating a mapping. Using the entity names from the database, the LLM is prompted to reformulate the question. If a database query results in multiple rows for a selected column and item, a new question is generated using a greedy approach, prompting the user for more specific details until a conclusive answer is reached. Experiments show that this method exhibits notable enhancements in comparison to contemporary state-of-the-art techniques in mitigating language model inaccuracies.

StructGPT (Jiang et al., 2023a) is a general prompt framework to support LLMs reasoning on structured data (e.g., KG, Table, and Database). In general, the core of this framework is that they construct the specialized interfaces to collect relevant evidence from structured data (i.e., reading), and let LLMs concentrate on the reasoning task based on the collected information (i.e., reasoning). Specially, they propose an invoking-linearization-generation procedure to support LLMs in reasoning on the structured data with the help of the interfaces. By iterating this procedure with provided interfaces, our approach can gradually approach the target answers to a given query. Experiments conducted on three types of structured data, including KGQA, TableQA, and Text-to-SQL, show that StructGPT greatly improves the performance of LLMs, under the few-shot or zero-shot settings.

Baek et al. (2023) propose to inject the factual knowledge from knowledge graphs into (large) language models (up to GPT-3.5), by retrieving the relevant facts from knowledge graphs based on their textual similarities with the input question and then injecting them as the prompt of language models. This approach improves the performance of language models on knowledge graph question answering tasks (Sen et al., 2022) by up to 48% on average, compared to baselines without knowledge graphs.

5.3 Domain Factuality Enhanced LLMs

Domain Knowledge Deficit is not only an important reason for limiting the application of LLM in specific fields, but also a subject of great concern to both academia and industry. In this subsection, we discuss how those Domain-Specific LLMs enhance their domain factuality.

Table 9 lists the domain-factuality enhanced LLMs. Here, we include several domains, including healthcare/medicine (H), finance (F), law/legal (L), geoscience/environment (G), education (E), food testing (FT), and home renovation (HR).

Based on the actual scenarios of Domain-Specific LLMs and our previous categorization of enhancement methods, we have summarized several commonly used enhancement techniques for Domain-Specific LLMs:

(1) Continual Pretraining: A method that involves continuously updating and fine-tuning a pre-trained language model using domain-specific data. This process ensures that the model stays up-to-date and relevant within a specific domain or field. It starts with an initial pre-trained model, often a general-purpose language model, and then fine-tunes it using domain-specific text or data. As new information becomes available, the model can be further fine-tuned to adapt to the evolving knowledge landscape. Continual pretraining is a powerful approach for maintaining the accuracy and relevance of AI models in rapidly changing domains, such as technology or medicine (Yang et al., 2023c; Zhang et al., 2023a).

(2) Continual SFT: Another strategy for enhancing the factuality of AI models. In this approach, the model is fine-tuned using labeled or annotated data specific to the domain of interest. This fine-tuning process allows the model to learn and adapt to the nuances and specifics of the domain, improving its ability to provide accurate and contextually relevant information. It can be particularly useful in applications where access to domain-specific labeled data is available over time, such as in the case of legal databases, medical records, or financial reports (Bao et al., 2023; Li et al., 2023f).

(3) Train From Scratch: It involves starting the learning process with minimal prior knowledge or pretraining. This approach can be likened to teaching a machine learning model with a blank slate. While it may not have the advantage of leveraging pre-existing knowledge, training from scratch can be advantageous when dealing with completely new domains or tasks where there is limited relevant data available. It allows the model to build its understanding from the ground up, although it may require substantial computational resources and time (Ross et al., 2022; Venigalla et al., 2022).

(4) External knowledge: involves augmenting a language model's internal knowledge with information from external sources. This method allows the model to access databases, websites, or other structured data repositories to verify facts or gather additional information when responding to user queries. By integrating external knowledge, the model can enhance its fact-checking capabilities and provide more accurate and contextually relevant answers, especially when dealing with dynamic or rapidly changing information. Below, we introduce these methods (Fan et al., 2023; Wang et al., 2023f).

TABLE 9
LLMs Enhanced for Domain-Specific Factuality. In the 'domain' column, we utilize the following abbreviations: healthcare/medicine (H), finance (F), law/legal (L), geoscience/environment (G), education (E), food testing (FT), and home renovation (HR).

Reference	Domain	Model	Eval Task	Eval Dataset	Continual Pretrained?	Continual SFT?	Train From Scratch?	External Knowledge
Zhang et al. (2023a)	H	Baichuan-7B, Ziya-LLaMA-13B	QA	cMedQA2, WebMedQA, Huatuo-26M	✓			
Yang et al. (2023c)	H	Ziya-LLaMA-13B	QA	CMtMedQA, huatuo-26M	✓	✓		
Wang et al. (2023f)	H	GPT-3.5-Turbo, LLaMA-2-13B	QA	MedQAUSMLE, MedQAMCMLE, MedMCQA				✓
Ross et al. (2022)	H	MOLFORMER	Molecule properties prediction				✓	
Bao et al. (2023)	H	Baichuan-13B	QA	CMB-Clin, CMD, CMID		✓		
Guan et al. (2023)	H	ChatGPT	IU-RR, MIMIC-CXR					✓
Liu et al. (2023f)	H	GPT-4	Medical Text De-Identification					✓
Li et al. (2023f)	H	LLaMA	QA			✓		
Venigalla et al. (2022)	H	GPT (2.7b)	QA				✓	
Xiong et al. (2023b)	H	ChatGLM-6B	QA			✓		
Tan et al. (2023a)	H	Baichuan-7B	QA	C-Eval, MMLU		✓		
Luo et al. (2022)	H	GPT-2	QA, DC, RE				✓	
Jin et al. (2023)	H	Codex	QA	GeneTuring				✓
Zakka et al. (2023)	H	text-davinci-003	QA	ClinicalQA				✓
Liu et al. (2023e)	H	GPT-2medium	Molecular Property Prediction, Molecule-text translation			✓	✓	
Nguyen (2023)	L	GPT3				✓		
Savelka et al. (2023)	L	GPT-4						✓
Huang et al. (2023a)	L	LLaMA	CN Legal Tasks		✓	✓		
Cui et al. (2023a)	L	Ziya-LLaMA-13B	QA	national judicial examination question	✓			✓
Li et al. (2023e)	F	BLOOMZ	4 major tasks 12 subtasks	EcomInstruct		✓		
Wu et al. (2023)	F	BLOOM	Financial NLP (SA, BC, NER, NER+NED, QA)	Financial Datasets			✓	
Deng et al. (2023)	G	LLaMA-7B		GeoBench	✓			
Bai et al. (2023)	G	ChatGLM-6B			✓			
Fan et al. (2023)	E	phoenix-inst-chat-7b	Chinese Grammatical Error Correction	ChatGPT-generated, Human-annotated		✓		
Qi et al. (2023)	FT	Chinese-LLaMA2-13B	QA		✓			✓
Wen et al. (2023)	HR	Baichuan-13B		C-Eval, CMMLU, EvalHome		✓		

For each Domain-specific LLM, we list its respective enhancement methods, which are presented in Table 9.

5.3.1 Healthcare domain-enhanced LLMs

These LLMs have emerged as powerful tools in the medical field, offering a diverse range of capabilities. These models, such as CohortGPT(Guan et al., 2023), ChatDoctor(Li et al., 2023f), DeID-GPTLiu et al., 2023f, BioMedLMVenigalla et al., 2022, DoctorGLMXiong et al., 2023a, MedChatZHTan et al., 2023a, BioGPT(Luo et al., 2022), GeneGPT(Jin et al., 2023), Almanac(Zakka et al., 2023), and MolXPT(Liu et al., 2023e), harness the potential of LLMs to revolutionize healthcare. They are equipped with features like classifying unstructured medical text into disease labels, improving performance with knowledge graphs and sample selection strategies, fine-tuning on large datasets of patient-doctor dialogues, enabling automatic medical text de-identification, excelling in medical question-answering tasks, handling traditional Chinese medical question-answering, and outperforming in various biomedical NLP tasks. Some models interact with web APIs for genomics questions, while others specialize in clinical guidelines and treatment recommendations. These LLMs not only demonstrate state-of-the-art performance on healthcare domain but also emphasize

the importance of domain-specific training and evaluation, showcasing their potential in transforming healthcare and clinical decision-making.

Zhang et al. (2023a) present HuatuoGPT, a medical language model that uses data from ChatGPT and doctors, resulting in state-of-the-art performance in medical consultations. It is based on Baichuan-7B and Ziya-LLaMA-13B-Pretrain-v1, continually pre-trained on both distilled data (from ChatGPT) and real-world data (from Doctors).

Likewise, Yang et al. (2023c) introduce Zhongjing, the first Chinese medical language model based on LLaMA, which utilizes a comprehensive training pipeline and a multi-turn medical dialogue dataset. Specifically, it is enhanced with a multi-turn medical dialogue dataset called CMtMedQA, consisting of 70,000 authentic doctor-patient dialogues, enabling complex dialogue and proactive inquiry. The backbone model used is Ziya-LLaMA-13B-v1, and the evaluation dataset is CMtMedQA and huatuo-26M (Li et al., 2023b).

Wang et al. (2023f) unveil a system, LLM-AMT, that improves large-scale language models like GPT-3.5-Turbo and LLaMA-2-13B with medical textbooks, notably enhancing open-domain medical question-answering tasks. At the same time, the external knowledge source is a Hybrid Textbook Retriever comprising 51 textbooks from the MedQA dataset and Wikipedia.

Bao et al. (2023) present DISC-MedLLM, a solution that uses LLMs to provide accurate medical responses in conversational healthcare services, utilizing strategies like medical knowledge graphs, real-world dialogue reconstruction, and human-guided preference rephrasing to create high-quality SFT datasets, applied on Baichuan-13B-Base. The paper uses various datasets for fine-tuning, including Re-constructed AI Doctor-Patient Dialogue, MedDialog, cMedQA, Knowledge Graph QA pairs (CMeKG), Behavioral Preference Dataset (Manual selection), MedMCQA, MOSS6, and Alpaca-GPT.

Similarly, Guan et al. (2023) introduce CohortGPT, a model that uses LLMs for participant recruitment in clinical research by classifying complex medical text into disease labels. CohortGPT enhances ChatGPT performance with the use of a knowledge graph as auxiliary information and a CoT sample selection strategy. The tasks involve IU-RR (Preparing a collection of radiology examinations for distribution and retrieval) and MIMIC-CXR (Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports). Li et al. (2023f) introduce ChatDoctor, a refined LLaMA-based model. It is fine-tuned using a dataset of 100,000 patient-doctor dialogues, and equipped with a self-directed information retrieval mechanism.

Liu et al. (2023f) introduce DeID-GPT, a framework leveraging GPT-4 for automatic medical text de-identification. Additionally, the paper mentions the use of HIPAA identifiers as an extra knowledge source to enhance the de-identification process.

Venigalla et al. (2022) present BioMedLM, a domain-specific LLM trained on PubMed data, for medical QA tasks. Xiong et al. (2023b) introduce DoctorGLM, a Chinese-focused language model fine-tuned for healthcare-specific tasks. Both Tan et al. (2023a) and Luo et al. (2022) introduce dialogue and generative Transformer language mod-

els for traditional Chinese medical question-answering and biomedical NLP tasks, respectively.

Jin et al. (2023) present GeneGPT, a method for teaching LLMs to answer genomics questions using NCBI Web APIs. Zakka et al. (2023) introduce Almanac, a LLM with retrieval capabilities for medical guideline recommendations. Lastly, Liu et al. (2023e) introduce MolXPT, a unified language model adept in molecular property prediction and molecular generation.

5.3.2 Legal domain enhanced LLMs

These LLMs, such as LawGPT (Nguyen, 2023), and ChatLaw (Cui et al., 2023b), have been fine-tuned to provide comprehensive legal assistance, from answering intricate legal queries and generating legal documents to offering expert legal advice. Leveraging extensive corpora of legal text, these models ensure context-aware and accurate responses. Moreover, their continual development involves injecting domain knowledge, designing supervised fine-tuning tasks, and incorporating retrieval modules to address issues like hallucination and ensure high-quality legal assistance. These innovations not only pave the way for more accessible and reliable legal services but also open up new avenues for research and exploration within the legal domain.

Nguyen (2023) introduce LawGPT 1.0, a fine-tuned GPT-3 language model for the legal domain to provide conversational legal assistance, including answering legal questions, generating legal documents, and offering legal advices. The paper mentions the use of a large corpus of legal text for fine-tuning the model to adapt it to the legal domain.

Savelka et al. (2023) evaluate the performance of GPT-4 in generating explanations of legal terms in legislation, comparing a baseline approach to an augmented approach that uses a legal information retrieval module to provide context from case law, revealing improvements in quality and addressing issues of factual accuracy and hallucination.

Huang et al. (2023a) address the challenge of enhancing LLMs like LLaMA for domain-specific tasks, particularly in the legal domain, by injecting domain knowledge during continual training, designing appropriate supervised fine-tune tasks and incorporating a retrieval module to improve factuality during text generation. They release their data and model for further research in Chinese legal tasks.

Cui et al. (2023a) introduce ChatLaw, an open-source legal LLM, designed for the Chinese legal domain. The paper introduces a method to improve model factuality during data screening, and a self-attention method for error handling. The paper uses various datasets for fine-tuning ChatLaw, including a collection of original legal data, data constructed based on legal regulations and judicial interpretations, and crawled real legal consultation data. The primary model used in this paper is Ziya-LLaMA-13B, which serves as the backbone for ChatLaw, tailored for the Chinese legal domain and optimized to handle legal questions and tasks. Additionally, the paper uses of a vector database retrieval method, keyword retrieval, and a self-attention method to enhance the model's performance in the legal domain.

5.3.3 Finance Domain-enhanced LLMs

These LLMs combine sophisticated language models designed specifically for commercial and financial tasks to deliver robust processing capabilities. They focus on creating tailor-made solutions optimized for both financial text analysis and e-commerce settings, trained on datasets containing myriad business-related tasks and copious financial tokens. They are designed to perform a plethora of functions, ranging from understanding and generating instructions for various E-commerce assignments to identifying sentiment, recognizing named entities, and answering questions in financial contexts. The models are further fine-tuned for zero-shot generalization on diverse tasks and benchmarks.

Li et al. (2023e) introduce EcomGPT, a language model tailored for E-commerce scenarios, trained on the newly created EcomInstruct dataset, which consists of 2.5 million instruction data spanning various E-commerce tasks and data types. The dataset covers product information, user reviews, and more. It defines atomic tasks and Chain-of-Task tasks to enable comprehensive training for E-commerce scenarios. The backbone model used is BLOOMZ, which is fine-tuned with the EcomInstruct dataset. The evaluation dataset includes 12 tasks, encompassing classification, generation, extraction, and other E-commerce-related tasks.

Wu et al. (2023) introduce BloombergGPT, a specialized 50 billion-parameter language model for the financial domain, trained on a massive 363 billion token dataset, which combines Bloomberg's extensive financial data sources with general-purpose datasets. The dataset used in this paper is an extensive 363 billion token dataset, which includes a significant portion of financial data from Bloomberg's sources (51.27% of the training data). BloombergGPT is based on a decoder-only causal language model architecture known as BLOOM. The evaluation includes various financial NLP tasks such as sentiment analysis, named entity recognition, binary classification, and question answering.

5.3.4 Other Domain-Enhanced LLMs

Geoscience and Environment domain-enhanced LLMs: are expertly designed, leveraging vast corpora to provide precise and robust results pertaining to geoscience and renewable energy. K2, a trailblazer in geoscience LLM, was trained on a massive geoscience text corpus and further refined using the GeoSignal dataset. Meanwhile, the HouYi model, another pioneering LLM focusing on renewable energy, harnessed the Renewable Energy Academic Paper dataset, containing over a million academic literature sources. These LLMs are fine-tuned to deliver adept performance in their respective fields, showing substantial capabilities in aligning their responses with user queries and renewable energy academic literature.

Deng et al. (2023) introduce K2, the first LLM designed specifically for geoscience, which is a LLaMA-7B continuously trained on a 5.5 billion token geoscience text corpus and fine-tuned using the GeoSignal dataset. The paper also presents resources like GeoSignal, a geoscience instruction tuning dataset, and GeoBench, the first geoscience benchmark for evaluating LLMs in the context of geoscience.

Bai et al. (2023) present the development of the HouYi model, the first LLM specifically designed for renewable energy, utilizing the newly created Renewable Energy Academic Paper (REAP) dataset, which contains over 1.1 million academic literature sources related to renewable energy, and the HouYi model is fine-tuned based on general LLMs such as ChatGLM-6B.

Education domain-enhanced LLMs: are used for assisting education scenarios. An example is GrammarGPT (Fan et al., 2023), which provides an innovative approach to language learning, particularly focusing on error correction in Chinese grammar. It is an open-source LLM designed for native Chinese grammatical error correction, which leverages a hybrid dataset of ChatGPT-generated and human-annotated data, along with heuristic methods to guide the model in generating ungrammatical sentences. The backbone model used is phoenix-inst-chat-7b.

Food domain-enhanced LLMs: are language models specifically designed to meet the distinct requirements of food testing protocols. For example, Qi et al. (2023) introduce FoodGPT, a LLM for food testing that incorporates structured knowledge and scanned documents using an incremental pre-training approach, with a focus on addressing machine hallucination by constructing a knowledge graph as an external knowledge base, utilizing the Chinese-LLaMA2-13B as the backbone model and collecting food-related data for training.

Home renovation domain-enhanced LLMs: are domain-specific language models tailored for home renovation tasks. For example, Wen et al. (2023) introduce ChatHome, which uses a dual-pronged methodology involving domain-adaptive pretraining and instruction-tuning on an extensive dataset comprising professional articles, standard documents, and web content relevant to home renovation. The backbone model is Baichuan-13B, and the evaluation datasets include C-Eval, CMMLU, and the newly created "EvalHome" domain dataset, while the fine-tuning data sources encompass National Standards, Domain Books, Domain Websites, and WuDaoCorpora.

6 CONCLUSION

Throughout this survey, we have systematically explored the intricate landscape of factuality issues within large language models (LLMs). We began by defining the concept of factuality (Sec 2.2) and proceeded to discuss its broader implications (Sec 2.3). Our journey took us through the multifaceted realm of factuality evaluation, encompassing benchmarks (Sec 3.2), metrics (Sec 3.1), specific evaluation studies (Sec 3.3), and domain-specific evaluations (Sec 3.4). We then delved deeper, probing the intrinsic mechanisms that underpin factuality in LLMs (Sec 4). Our exploration culminated in the discussion of enhancement techniques, both for pure LLMs (Sec 5.1) and retrieval-augmented LLMs (Sec 5.2), with a special focus on domain-specific LLM enhancements (Sec 5.3).

Despite the advancements detailed in this survey, several challenges loom large. The evaluation of factuality remains an intricate puzzle, complicated by the inherent variability and nuances of natural languages. The core processes governing how LLMs store, update, and produce facts are yet not fully revealed. And while certain techniques, like continual training and retrieval, show promise, they are

not without limitations. Looking ahead, the quest for fully factual LLMs presents both challenges and opportunities. Future research might delve deeper into understanding the neural architectures of LLMs, develop more robust evaluation metrics, and innovate on enhancement techniques. As LLMs become increasingly integrated into our digital ecosystem, ensuring their factual reliability will remain paramount, with implications that impact across the AI community and beyond.

REFERENCES

Introducing claude.

Oshin Agarwal, Heming Ge, Siamak Shakeri, and Rami Al-Rfou. 2021. Knowledge graph based synthetic corpus generation for knowledge-enhanced language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3554–3565, Online. Association for Computational Linguistics.

Badr AlKhamissi, Millicent Li, Asli Celikyilmaz, Mona Diab, and Marjan Ghazvininejad. 2022. A review on language models as knowledge bases.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Heslow, Julien Launay, Quentin Malartic, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. Falcon-40B: an open large language model with state-of-the-art performance.

Rami Aly, Zhijiang Guo, Michael Sejr Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. The fact extraction and VERification over unstructured and structured information (FEVEROUS) shared task. In *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*, pages 1–13, Dominican Republic. Association for Computational Linguistics.

Amos Azaria and Tom Mitchell. 2023. The internal state of an llm knows when its lying. *arXiv preprint arXiv:2304.13734*.

Jinheon Baek, Alham Fikri Aji, and Amir Saffari. 2023. Knowledge-augmented language model prompting for zero-shot knowledge graph question answering.

Mingliang Bai, Zhihao Zhou, Ruidong Wang, Yusheng Yang, Zizhen Qin, Yunxiao Chen, Chunjin Mu, Jinfu Liu, and Daren Yu. 2023. Houyi: An open-source large language model specially designed for renewable energy and carbon neutrality field. *arXiv preprint arXiv:2308.01414*.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*.

Zhijie Bao, Wei Chen, Shengze Xiao, Kuang Ren, Jiaao Wu, Cheng Zhong, Jiajie Peng, Xuanjing Huang, and Zhongyu Wei. 2023. Disc-medllm: Bridging general large language models and real-world medical consultation. *arXiv preprint arXiv:2308.14346*.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on Freebase from question-answer pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1533–1544, Seattle, Washington, USA. Association for Computational Linguistics.

Lukas Berglund, Meg Tong, Max Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. 2023. The reversal curse: Llms trained on "a is b" fail to learn "b is a".

Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. 2021. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow. If you use this software, please cite it using these metadata.

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.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4.

Meng Cao, Yue Dong, and Jackie Cheung. 2022. Hallucinated but factual! inspecting the factuality of hallucinations in abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3340–3354, Dublin, Ireland. Association for Computational Linguistics.

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2021. Evaluation of text generation: A survey. In *arXiv preprint arXiv:2006.14799*.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Kaijie Zhu, Hao Chen, Linyi Yang, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2023. A survey on evaluation of large language models. *arXiv preprint arXiv:2307.03109*.

Harrison Chase. 2022. Langchain.

Anthony Chen, Panupong Pasupat, Sameer Singh, Hongrae Lee, and Kelvin Guu. 2023a. Purr: Efficiently editing language model hallucinations by denoising language model corruptions. *arXiv preprint arXiv:2305.14908*.

Hung-Ting Chen, Michael Zhang, and Eunsol Choi. 2022. Rich knowledge sources bring complex knowledge conflicts: Recalibrating models to reflect conflicting evidence. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2292–2307.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2023b. Benchmarking large language models in retrieval-augmented generation.

Liang Chen, Yang Deng, Yatao Bian, Zeyu Qin, Bingzhe Wu, Tat-Seng Chua, and Kam-Fai Wong. 2023c. Beyond factuality: A comprehensive evaluation of large language models as knowledge generators. *arXiv preprint arXiv:2310.07289*.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021a. Evaluating large language models trained on code.

Sanyuan Chen, Yutai Hou, Yiming Cui, Wanxiang Che, Ting Liu, and Xiangzhan Yu. 2020. Recall and learn: Fine-tuning deep pretrained language models with less forgetting. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7870–7881.

Wenhu Chen, Ming-Wei Chang, Eva Schlinger, William Wang, and William W. Cohen. 2021b. Open question answering over tables and text.

Yuheng Chen, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. 2023d. Journey to the center of the knowledge neurons: Discoveries of language-independent knowledge neurons and degenerate knowledge neurons.

Zhihong Chen, Feng Jiang, Junying Chen, Tiannan Wang, Fei Yu, Guiming Chen, Hongbo Zhang, Juhao Liang, Chen Zhang, Zhiyi Zhang, Jianquan Li, Xiang Wan, Benyou Wang, and Haizhou Li. 2023e. Phoenix: Democratizing chatgpt across languages. *arXiv preprint arXiv:2304.10453*.

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.

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

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. 2023. Dola: Decoding by contrasting layers improves factuality in large language models. *arXiv preprint arXiv:2309.03883*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Roi Cohen, Eden Biran, Ori Yoran, Amir Globerson, and Mor Geva. 2023a. Evaluating the ripple effects of knowledge editing in language models.

Roi Cohen, May Hamri, Mor Geva, and Amir Globerson. 2023b. Lm vs lm: Detecting factual errors via cross examination.

Together Computer. 2023. Redpajama: An open source recipe to reproduce llama training dataset.

Jiaxi Cui, Zongjian Li, Yang Yan, Bohua Chen, and Li Yuan. 2023a. Chatlaw: Open-source legal large language model with integrated external knowledge bases. *arXiv preprint arXiv:2306.16092*.

Jiaxi Cui, Zongjian Li, Yang Yan, Bohua Chen, and Li Yuan. 2023b. Chatlaw: Open-source legal large language model with integrated external knowledge bases.

Yiming Cui, Ziqing Yang, and Xin Yao. 2023c. Efficient and effective text encoding for chinese llama and alpaca. *arXiv preprint arXiv:2304.08177*.

Curation. 2020. Curation corpus base.

Shawn Curran, Sam Lansley, and Oliver Bethell. 2023. Hallucination is the last thing you need.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8493–8502, Dublin, Ireland. Association for Computational Linguistics.

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2021. Editing factual knowledge in language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6491–6506, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Michiel de Jong, Yury Zemlyanskiy, Nicholas FitzGerald, Fei Sha, and William W. Cohen. 2022. Mention memory: in-

corporating textual knowledge into transformers through entity mention attention. In *International Conference on Learning Representations*.

Cheng Deng, Tianhang Zhang, Zhongmou He, Qiyuan Chen, Yuanyuan Shi, Le Zhou, Luoyi Fu, Weinan Zhang, Xinbing Wang, Chenghu Zhou, et al. 2023. Learning a foundation language model for geoscience knowledge understanding and utilization. *arXiv preprint arXiv:2306.05064*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2023. Chain-of-verification reduces hallucination in large language models. *arXiv preprint arXiv:2309.11495*.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents.

Qingxiu Dong, Damai Dai, Yifan Song, Jingjing Xu, Zhifang Sui, and Lei Li. 2022. Calibrating factual knowledge in pretrained language models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5937–5947, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate.

Julian Eisenschlos, Bhuwan Dhingra, Jannis Bulian, Benjamin Börschinger, and Jordan Boyd-Graber. 2021. Fool me twice: Entailment from Wikipedia gamification. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 352–365, Online. Association for Computational Linguistics.

Hady Elsahar, Pavlos Vougiouklis, Arslen Remaci, Christophe Gravier, Jonathon Hare, Frederique Laforest, and Elena Simperl. 2018. T-REx: A large scale alignment of natural language with knowledge base triples. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567, Florence, Italy. Association for Computational Linguistics.

Yaxin Fan, Feng Jiang, Peifeng Li, and Haizhou Li. 2023. Grammargpt: Exploring open-source llms for native chinese grammatical error correction with supervised fine-tuning. *arXiv preprint arXiv:2307.13923*.

Robert M Fano and David Hawkins. 1961. Transmission of information: A statistical theory of communications. *American Journal of Physics*, 29(11):793–794.

Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Songyang Zhang, Kai Chen, Zongwen Shen, and Jidong Ge. 2023. Lawbench: Benchmarking legal knowledge of large language models. *arXiv preprint arXiv:2309.16289*.

Virginia Felkner, Ho-Chun Herbert Chang, Eugene Jang, and Jonathan May. 2023. Winoqueer: A community-in-the-loop benchmark for anti-lgbtq+ bias in large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9126–9140.

James Ferguson, Matt Gardner, Hannaneh Hajishirzi, Tushar Khot, and Pradeep Dasigi. 2020. IIRC: A dataset of incomplete information reading comprehension questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1137–1147, Online. Association for Computational Linguistics.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166*.

Thibault Févry, Livio Baldini Soares, Nicholas FitzGerald, Eunsol Choi, and Tom Kwiatkowski. 2020. Entities as experts: Sparse memory access with entity supervision.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2023. Bias and fairness in large language models: A survey. *arXiv preprint arXiv:2309.00770*.

Michel Galley, Chris Brockett, Xiang Gao, Jianfeng Gao, and Bill Dolan. 2019a. Grounded response generation task at dstc7. In *AAAI Dialog System Technology Challenges Workshop*.

Michel Galley, Chris Brockett, Xiang Gao, Jianfeng Gao, and Bill Dolan. 2019b. Grounded response generation task at dstc7. In *AAAI Dialog System Technology Challenges Workshop*.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The pile: An 800gb dataset of diverse text for language modeling.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, et al. 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)*, pages 16477–16508.

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

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, Online. Association for Computational Linguistics.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir

Globerson. 2023. Dissecting recall of factual associations in auto-regressive language models. *CoRR*, abs/2304.14767.

Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. 2022. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 30–45, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021a. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021b. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021c. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ian J. Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. 2015. An empirical investigation of catastrophic forgetting in gradient-based neural networks.

Google. 2023. Bard. *bard.google.com*.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. 2023. Critic: Large language models can self-correct with tool-interactive critiquing.

Zihan Guan, Zihao Wu, Zhengliang Liu, Dufan Wu, Hui Ren, Quanzheng Li, Xiang Li, and Ninghao Liu. 2023. Cohortgpt: An enhanced gpt for participant recruitment in clinical study. *arXiv preprint arXiv:2307.11346*.

Neel Guha, Julian Nyarko, Daniel E Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel N Rockmore, et al. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *arXiv preprint arXiv:2308.11462*.

Vivek Gupta, Maitrey Mehta, Pegah Nokhiz, and Vivek Srikumar. 2020a. INFOTABS: Inference on tables as semi-structured data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2309–2324, Online. Association for Computational Linguistics.

Vivek Gupta, Maitrey Mehta, Pegah Nokhiz, and Vivek Srikumar. 2020b. Infotabs: Inference on tables as semi-structured data. *arXiv preprint arXiv:2005.06117*.

Hiroaki Hayashi, Prashant Budania, Peng Wang, Chris Ackerson, Raj Neervannan, and Graham Neubig. 2021. Wiki-Asp: A dataset for multi-domain aspect-based summarization. *Transactions of the Association for Computational Linguistics*, 9:211–225.

Hangfeng He, Hongming Zhang, and Dan Roth. 2022. Rethinking with retrieval: Faithful large language model inference.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Takanobu Hirosawa, Yukinori Harada, Masashi Yokose, Tetsu Sakamoto, Ren Kawamura, and Taro Shimizu. 2023. Diagnostic accuracy of differential-diagnosis lists generated by generative pretrained transformer 3 chatbot for clinical vignettes with common chief complaints: A pilot study. *International journal of environmental research and public health*, 20(4):3378.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6609–6625, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Tamanna Hossain, Sunipa Dev, and Sameer Singh. 2023. MISGENDERED: Limits of large language models in understanding pronouns. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5352–5367, Toronto, Canada. Association for Computational Linguistics.

Wenpin Hou and Zhicheng Ji. 2023. Geneturing tests gpt models in genomics. *bioRxiv*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp.

Xuming Hu, Junzhe Chen, Xiaochuan Li, Yufei Guo, Lijie Wen, Philip S Yu, and Zhijiang Guo. 2023. Do large language models know about facts? *arXiv preprint arXiv:2310.05177*.

Jie Huang and Kevin Chen-Chuan Chang. 2022. Towards reasoning in large language models: A survey. *arXiv preprint arXiv:2212.10403*.

Jie Huang and Kevin Chen-Chuan Chang. 2023. Towards reasoning in large language models: A survey.

Quzhe Huang, Mingxu Tao, Zhenwei An, Chen Zhang, Cong Jiang, Zhibin Chen, Zirui Wu, and Yansong Feng. 2023a. Lawyer llama technical report. *arXiv preprint arXiv:2305.15062*.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. 2023b. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *arXiv preprint arXiv:2305.08322*.

Zeyu Huang, Yikang Shen, Xiaofeng Zhang, Jie Zhou, Wenge Rong, and Zhang Xiong. 2023c. Transformer-patcher: One mistake worth one neuron. In *The Eleventh International Conference on Learning Representations*.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning. *arXiv preprint arXiv:2112.09118*.

Gautier Izacard and Edouard Grave. 2021a. Distilling Knowledge from Reader to Retriever for Question Answering. In *ICLR 2021 - 9th International Conference on*

Learning Representations, Vienna, Austria.

Gautier Izacard and Edouard Grave. 2021b. 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*, pages 874–880, Online. Association for Computational Linguistics.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2022. Atlas: Few-shot learning with retrieval augmented language models.

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

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

Zhen Jia, Abdalghani Abujabal, Rishiraj Saha Roy, Jannik Strötgen, and Gerhard Weikum. 2018. Tempquestions: A benchmark for temporal question answering. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, page 1057–1062, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Jinhao Jiang, Kun Zhou, Zican Dong, Keming Ye, Wayne Xin Zhao, and Ji-Rong Wen. 2023a. Structgpt: A general framework for large language model to reason on structured data.

Kelvin Jiang, Dekun Wu, and Hui Jiang. 2019. FreebaseQA: A new factoid QA data set matching trivia-style question-answer pairs with Freebase. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 318–323, Minneapolis, Minnesota. Association for Computational Linguistics.

Yichen Jiang, Shikha Bordia, Zheng Zhong, Charles Dognin, Maneesh Singh, and Mohit Bansal. 2020. HoVer: A dataset for many-hop fact extraction and claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3441–3460, Online. Association for Computational Linguistics.

Zhengbao Jiang, Frank F. Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023b. Active retrieval augmented generation.

Qiao Jin, Yifan Yang, Qingyu Chen, and Zhiyong Lu. 2023. Genegpt: Augmenting large language models with domain tools for improved access to biomedical information. *ArXiv*.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, Vancouver, Canada. Association for Computational Linguistics.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. 2022. Language models (mostly) know what they know.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *International Conference on Machine Learning*, pages 15696–15707. PMLR.

Minki Kang, Jinheon Baek, and Sung Ju Hwang. 2022. KALA: knowledge-augmented language model adaptation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5144–5167, Seattle, United States. Association for Computational Linguistics.

Jungo Kasai, Keisuke Sakaguchi, Yoichi Takahashi, Ronan Le Bras, Akari Asai, Xinyan Yu, Dragomir Radev, Noah A. Smith, Yejin Choi, and Kentaro Inui. 2022. Real-Time QA: What's the answer right now?

Tushar Khot, Kyle Richardson, Daniel Khashabi, and Ashish Sabharwal. 2022. Hey ai, can you solve complex tasks by talking to agents?

Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. 2023. Decomposed prompting: A modular approach for solving complex tasks.

Suhas Kotha, Jacob Mitchell Springer, and Aditi Raghunathan. 2023. Understanding catastrophic forgetting in language models via implicit inference.

Tiffany H Kung, Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, Maria Madriaga, Rimel Aggabao, Giezel Diaz-Candido, James Maningo, et al. 2023. Performance of chatgpt on usmle: Potential for ai-assisted medical education using large language models. *PLoS digital health*, 2(2):e0000198.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association of Computational Linguistics*.

Angeliki Lazaridou, Elena Gribovskaya, Wojciech Stokowiec, and Nikolai Grigorev. 2022. Internet-augmented language models through few-shot prompting for open-domain question answering.

Rémi Lebret, David Grangier, and Michael Auli. 2016. Neural text generation from structured data with application to the biography domain. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1203–1213, Austin, Texas. Association for Computational Linguistics.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas

Carlini. 2022a. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8424–8445, Dublin, Ireland. Association for Computational Linguistics.

Nayeon Lee, Wei Ping, Peng Xu, Mostofa Patwary, Pascale N Fung, Mohammad Shoeybi, and Bryan Catanzaro. 2022b. Factuality enhanced language models for open-ended text generation. *Advances in Neural Information Processing Systems*, 35:34586–34599.

Douglas B. Lenat. 1995. Cyc: A large-scale investment in knowledge infrastructure. *Commun. ACM*, 38(11):33–38.

Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. 2017. Zero-shot relation extraction via reading comprehension. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 333–342, Vancouver, Canada. Association for Computational Linguistics.

Daliang Li, Ankit Singh Rawat, Manzil Zaheer, Xin Wang, Michal Lukasik, Andreas Veit, Felix Yu, and Sanjiv Kumar. 2023a. Large language models with controllable working memory. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1774–1793, Toronto, Canada. Association for Computational Linguistics.

Haonan Li, Martin Tomko, Maria Vasardani, and Timothy Baldwin. 2022a. MultiSpanQA: A dataset for multi-span question answering. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1250–1260, Seattle, United States. Association for Computational Linguistics.

Jianquan Li, Xidong Wang, Xiangbo Wu, Zhiyi Zhang, Xiaolong Xu, Jie Fu, Prayag Tiwari, Xiang Wan, and Benyou Wang. 2023b. Huatuo-26m, a large-scale chinese medical qa dataset. *arXiv preprint arXiv:2305.01526*.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023c. Halueval: A large-scale hallucination evaluation benchmark for large language models.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023d. Inference-time intervention: Eliciting truthful answers from a language model.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2022b. Contrastive decoding: Open-ended text generation as optimization. *arXiv preprint arXiv:2210.15097*.

Yaliang Li, Jing Gao, Chuishi Meng, Qi Li, Lu Su, Bo Zhao, Wei Fan, and Jiawei Han. 2015. A survey on truth discovery. *SIGKDD Explor.*, 17(2):1–16.

Yangning Li, Shirong Ma, Xiaobin Wang, Shen Huang, Chengyue Jiang, Hai-Tao Zheng, Pengjun Xie, Fei Huang, and Yong Jiang. 2023e. Ecomgpt: Instruction-tuning large language model with chain-of-task tasks for e-commerce. *arXiv preprint arXiv:2308.06966*.

Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023f. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Cureus*, 15(6).

Zonglin Li, Ruiqi Guo, and Sanjiv Kumar. 2022c. Decoupled context processing for context augmented language modeling. In *Advances in Neural Information Processing Systems*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022a. Teaching models to express their uncertainty in words. *arXiv preprint arXiv:2205.14334*.

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

Yen-Ting Lin and Yun-Nung Chen. 2023. Llm-eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models. *arXiv preprint arXiv:2305.13711*.

Chen Ling, Xujiang Zhao, Jiaying Lu, Chengyuan Deng, Can Zheng, Junxiang Wang, Tanmoy Chowdhury, Yun Li, Hejie Cui, Xuchao Zhang, Tianjiao Zhao, Amit Panalkar, Wei Cheng, Haoyu Wang, Yanchi Liu, Zhengzhang Chen, Haifeng Chen, Chris White, Quanquan Gu, Jian Pei, and Liang Zhao. 2023a. Domain specialization as the key to make large language models disruptive: A comprehensive survey.

Chen Ling, Xujiang Zhao, Jiaying Lu, Chengyuan Deng, Can Zheng, Junxiang Wang, Tanmoy Chowdhury, Yun Li, Hejie Cui, Tianjiao Zhao, et al. 2023b. Beyond one-model-fits-all: A survey of domain specialization for large language models. *arXiv preprint arXiv:2305.18703*.

Alisa Liu, Zhaofeng Wu, Julian Michael, Alane Suhr, Peter West, Alexander Koller, Swabha Swayamdipta, Noah A Smith, and Yejin Choi. 2023a. We're afraid language models aren't modeling ambiguity. *arXiv preprint arXiv:2304.14399*.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2023b. Mitigating hallucination in large multi-modal models via robust instruction tuning.

Jerry Liu. 2022. LlamaIndex.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023c. Lost in the middle: How language models use long contexts. *arXiv preprint arXiv:2307.03172*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Yixin Liu, Alex Fabbri, Pengfei Liu, Yilun Zhao, Linyong Nan, Ruilin Han, Simeng Han, Shafiq Joty, Chien-Sheng Wu, Caiming Xiong, and Dragomir Radev. 2023d. Revisiting the gold standard: Grounding summarization evaluation with robust human evaluation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4140–4170, Toronto, Canada. Association for Computational Linguistics.

Zequn Liu, Wei Zhang, Yingce Xia, Lijun Wu, Shufang Xie, Tao Qin, Ming Zhang, and Tie-Yan Liu. 2023e. Molxpt: Wrapping molecules with text for generative pre-training. *arXiv preprint arXiv:2305.10688*.

Zhengliang Liu, Xiaowei Yu, Lu Zhang, Zihao Wu, Chao Cao, Haixing Dai, Lin Zhao, Wei Liu, Dinggang Shen,

Quanzheng Li, et al. 2023f. Deid-gpt: Zero-shot medical text de-identification by gpt-4. *arXiv preprint arXiv:2303.11032*.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7052–7063.

Ryan Lowe, Michael Noseworthy, Iulian Vlad Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 2017. Towards an automatic Turing test: Learning to evaluate dialogue responses. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*.

Daniel D Lundstrom, Tianjian Huang, and Meisam Razaviyayn. 2022. A rigorous study of integrated gradients method and extensions to internal neuron attributions. In *International Conference on Machine Learning*, pages 14485–14508. PMLR.

Hongyin Luo, Yung-Sung Chuang, Yuan Gong, Tianhua Zhang, Yoon Kim, Xixin Wu, Danny Fox, Helen Meng, and James Glass. 2023a. Sail: Search-augmented instruction learning. *arXiv preprint arXiv:2305.15225*.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6):bbac409.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. 2023b. An empirical study of catastrophic forgetting in large language models during continual fine-tuning.

Dakota Mahan, Ryan Carlow, Louis Castricato, Nathan Cooper, and Christian Laforte. Stable beluga models.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822, Toronto, Canada. Association for Computational Linguistics.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models.

Luca Massarelli, Fabio Petroni, Aleksandra Piktus, Myle Ott, Tim Rocktäschel, Vassilis Plachouras, Fabrizio Silvestri, and Sebastian Riedel. 2020. How decoding strategies affect the verifiability of generated text. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 223–235.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

J. McCarthy, M. L. Minsky, N. Rochester, and C. E. Shannon. 1955. A PROPOSAL FOR THE DARTMOUTH SUMMER RESEARCH PROJECT ON ARTIFICIAL INTELLIGENCE. http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html.

Alan Melikdjanian. 2018. Captain disillusion's escape from the USSR. *https://www.youtube.com/watch?v=MaDz0FCxzR8*.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. *Advances in Neural Information Processing Systems*, 36.

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

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

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*.

Microsoft. 2023. Bing chat. *https://www.bing.com/new*.

George A. Miller. 1992. WordNet: A lexical database for English. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*.

George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. 1990. Introduction to WordNet: An On-line Lexical Database*. *International Journal of Lexicography*, 3(4):235–244.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wentau 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. *arXiv preprint arXiv:2305.14251*.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. AmbigQA: Answering ambiguous open-domain questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5783–5797, Online. Association for Computational Linguistics.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D Manning. 2022a. Fast model editing at scale. In *International Conference on Learning Representations*.

Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D. Manning, and Chelsea Finn. 2022b. Memory-based model editing at scale. In *International Conference on Machine Learning*.

Fedor Moiseev, Zhe Dong, Enrique Alfonseca, and Martin Jaggi. 2022. SKILL: Structured knowledge infusion for large language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1581–1588, Seattle, United States. Association for Computational Linguistics.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf,

et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

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

Ella Neeman, Roee Aharoni, Or Honovich, Leshem Choshen, Idan Szpektor, and Omri Abend. 2023. DisentQA: Disentangling parametric and contextual knowledge with counterfactual question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10056–10070, Toronto, Canada. Association for Computational Linguistics.

Allen Newell and Herbert A. Simon. 1976. Computer science as empirical inquiry: Symbols and search. *Commun. ACM*, 19(3):113–126.

Ha-Thanh Nguyen. 2023. A brief report on lawgpt 1.0: A virtual legal assistant based on gpt-3. *arXiv preprint arXiv:2302.05729*.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems.

OpenAI. 2022a. Gpt-3.5 - openai. *https://platform.openai.com/docs/models/gpt-3-5*.

OpenAI. 2022b. Introducing chatgpt. *https://openai.com/blog/chatgpt*.

OpenAI. 2023. Gpt-4 technical report. *arXiv*.

Liangming Pan, Xiaobao Wu, Xinyuan Lu, Anh Tuan Luu, William Yang Wang, Min-Yen Kan, and Preslav Nakov. 2023a. Fact-checking complex claims with program-guided reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6981–7004, Toronto, Canada. Association for Computational Linguistics.

Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. 2023b. Unifying large language models and knowledge graphs: A roadmap.

Yikang Pan, Liangming Pan, Wenhu Chen, Preslav Nakov, Min-Yen Kan, and William Yang Wang. 2023c. On the risk of misinformation pollution with large language models. *arXiv preprint arXiv:2305.13661*.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Ngoc Quan Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. 2016. The LAMBADA dataset: Word prediction requiring a broad discourse context. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1525–1534, Berlin, Germany. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*.

Daniel Park. 2023. Open-llm-leaderboard-report.

Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, and Jianfeng Gao. 2023. Check your facts and try again: Improving large language models with external knowledge and automated feedback.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. 2021. KILT: a benchmark for knowledge intensive language tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2523–2544, Online. Association for Computational Linguistics.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019a. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019b. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. 2019c. Language models as knowledge bases?

Pouya Pezeshkpour. 2023. Measuring and modifying factual knowledge in large language models.

Will Oremus Pranshu Verma. 2023. Chatgpt invented a sexual harassment scandal and named a real law prof as the accused. *The Washington Post*.

Xiao Pu, Mingqi Gao, and Xiaojun Wan. 2023. Summarization is (almost) dead.

Zhixiao Qi, Yijiong Yu, Meiqi Tu, Junyi Tan, and Yongfeng Huang. 2023. Foodgpt: A large language model in food testing domain with incremental pre-training and knowledge graph prompt. *arXiv preprint arXiv:2308.10173*.

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.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine

Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew Johnson, Blake Hechtman, Laura Weidinger, Iason Gabriel, William Isaac, Ed Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. 2022a. Scaling language models: Methods, analysis & insights from training gopher.

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew Johnson, Blake Hechtman, Laura Weidinger, Iason Gabriel, William Isaac, Ed Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. 2022b. Scaling language models: Methods, analysis & insights from training gopher.

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

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

Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. 2023. Measuring attribution in natural language generation models. *Computational Linguistics*, pages 1–66.

Vipula Rawte, Amit Sheth, and Amitava Das. 2023. A survey of hallucination in large foundation models.

Ruiyang Ren, Yuhao Wang, Yingqi Qu, Wayne Xin Zhao, Jing Liu, Hao Tian, Hua Wu, Ji-Rong Wen, and Haifeng Wang. 2023. Investigating the factual knowledge boundary of large language models with retrieval augmentation.

Jerret Ross, Brian Belgodere, Vijil Chenthamarakshan, Inkit Padhi, Youssef Mroueh, and Payel Das. 2022. Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intelligence*, 4(12):1256–1264.

Nafis Sadeq, Byungkyu Kang, Prarit Lamba, and Julian McAuley. 2023. Unsupervised improvement of factual knowledge in language models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2960–2969, Dubrovnik, Croatia. Association for Computational Linguistics.

Leo Sands. 2023. Chatgpt falsely told voters their mayor was jailed for bribery. he may sue. *The Washington Post*.

Jaromir Savelka, Kevin D Ashley, Morgan A Gray, Hannes Westermann, and Huihui Xu. 2023. Explaining legal concepts with augmented large language models (gpt-4). *arXiv preprint arXiv:2306.09525*.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Priyanka Sen, Alham Fikri Aji, and Amir Saffari. 2022. Mintaka: A complex, natural, and multilingual dataset for end-to-end question answering. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1604–1619, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Raj Sanjay Shah, Kunal Chawla, Dheeraj Eidnani, Agam Shah, Wendi Du, Sudheer Chava, Natraj Raman, Charese Smiley, Jiaao Chen, and Diyi Yang. 2022. When flue meets flang: Benchmarks and large pre-trained language model for financial domain. *arXiv preprint arXiv:2211.00083*.

Yiqiu Shen, Laura Heacock, Jonathan Elias, Keith D Hentel, Beatriu Reig, George Shih, and Linda Moy. 2023. Chatgpt and other large language models are double-edged swords.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen tau Yih. 2023. Replug: Retrieval-augmented black-box language models.

Noah Shinn, Federico Cassano, Beck Labash, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning.

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2019. Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. 2021. Alfworld: Aligning text and embodied environments for interactive learning.

Devendra Singh, Siva Reddy, Will Hamilton, Chris Dyer,

and Dani Yogatama. 2021. End-to-end training of multi-document reader and retriever for open-domain question answering. *Advances in Neural Information Processing Systems*, 34:25968–25981.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature*, pages 1–9.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *Advances in Neural Information Processing Systems*, 33:16857–16867.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Ivan Stelmakh, Yi Luan, Bhuwan Dhingra, and Ming-Wei Chang. 2022. ASQA: Factoid questions meet long-form answers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8273–8288, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hao Sun, Xiao Liu, Yeyun Gong, Anlei Dong, Jingwen Lu, Yan Zhang, Daxin Jiang, Linjun Yang, Rangan Majumder, and Nan Duan. 2023a. Beamsearchqa: Large language models are strong zero-shot qa solver.

Kai Sun, Yifan Ethan Xu, Hanwen Zha, Yue Liu, and Xin Luna Dong. 2023b. Head-to-tail: How knowledgeable are large language models (llm)? aka will llms replace knowledge graphs? *arXiv preprint arXiv:2308.10168*.

Weiwei Sun, Zhengliang Shi, Shen Gao, Pengjie Ren, Maarten de Rijke, and Zhaochun Ren. 2023c. Contrastive learning reduces hallucination in conversations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13618–13626.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. 2023d. Aligning large multimodal models with factually augmented rlhf.

Derek Tam, Anisha Mascarenhas, Shiyue Zhang, Sarah Kwan, Mohit Bansal, and Colin Raffel. 2023. Evaluating the factual consistency of large language models through news summarization. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5220–5255, Toronto, Canada. Association for Computational Linguistics.

Yang Tan, Mingchen Li, Zijie Huang, Huiqun Yu, and Guisheng Fan. 2023a. Medchatzh: a better medical adviser learns from better instructions. *arXiv preprint arXiv:2309.01114*.

Yiming Tan, Dehai Min, Yu Li, Wenbo Li, Nan Hu, Yongrui Chen, and Guilin Qi. 2023b. Can chatgpt replace traditional kbqa models? an in-depth analysis of the question answering performance of the gpt llm family.

Xiangru Tang, Arman Cohan, and Mark Gerstein. 2023. Aligning factual consistency for clinical studies summarization through reinforcement learning. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 48–58, Toronto, Canada. Association for Computational Linguistics.

Xiangru Tang, Arjun Nair, Borui Wang, Bingyao Wang, Jai Desai, Aaron Wade, Haoran Li, Asli Celikyilmaz, Yashar Mehdad, and Dragomir Radev. 2022. CONFIT: Toward faithful dialogue summarization with linguistically-informed contrastive fine-tuning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5657–5668, Seattle, United States. Association for Computational Linguistics.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

InternLM Team. 2023a. Internlm: A multilingual language model with progressively enhanced capabilities.

MosaicML NLP Team. 2023b. Introducing mpt-30b: Raising the bar for open-source foundation models. Accessed: 2023-06-22.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, pages 1–11.

James Thorne and Andreas Vlachos. 2021. Evidence-based factual error correction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3298–3309.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification.

Together. 2023. Releasing 3b and 7b redpajama-incite family of models including base, instruction-tuned & chat models. https://www.together.xyz/blog/redpajama-models-v1.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan

Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. MuSiQue: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10014–10037, Toronto, Canada. Association for Computational Linguistics.

Neeraj Varshney, Wenlin Yao, Hongming Zhang, Jianshu Chen, and Dong Yu. 2023. A stitch in time saves nine: Detecting and mitigating hallucinations of llms by validating low-confidence generation.

A Venigalla, J Frankle, and M Carbin. 2022. Biomedlm: a domain-specific large language model for biomedical text. *MosaicML. Accessed: Dec*, 23(3):2.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, and Thang Luong. 2023. Freshllms: Refreshing large language models with search engine augmentation.

Zhongwei Wan, Yichun Yin, Wei Zhang, Jiaxin Shi, Lifeng Shang, Guangyong Chen, Xin Jiang, and Qun Liu. 2022. G-map: General memory-augmented pre-trained language model for domain tasks.

Cunxiang Wang, Sirui Cheng, Zhikun Xu, Bowen Ding, Yidong Wang, and Yue Zhang. 2023a. Evaluating open question answering evaluation.

Cunxiang Wang, Pai Liu, and Yue Zhang. 2021. Can generative pre-trained language models serve as knowledge bases for closed-book QA? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3241–3251, Online. Association for Computational Linguistics.

Cunxiang Wang, Fuli Luo, Yanyang Li, Runxin Xu, Fei Huang, and Yue Zhang. 2023b. Knowledgeable salient span mask for enhancing language models as knowledge base. In *Natural Language Processing and Chinese Computing*, pages 444–456, Cham. Springer Nature Switzerland.

Haochun Wang, Chi Liu, Nuwa Xi, Zewen Qiang, Sendong Zhao, Bing Qin, and Ting Liu. 2023c. Huatuo: Tuning llama model with chinese medical knowledge.

Xidong Wang, Guiming Hardy Chen, Dingjie Song, Zhiyi Zhang, Zhihong Chen, Qingying Xiao, Feng Jiang, Jianquan Li, Xiang Wan, Benyou Wang, et al. 2023d. Cmb: A comprehensive medical benchmark in chinese. *arXiv preprint arXiv:2308.08833*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023e. Self-consistency improves chain of thought reasoning in language models.

Yihan Wang, Si Si, Daliang Li, Michal Lukasik, Felix Yu, Cho-Jui Hsieh, Inderjit S Dhillon, and Sanjiv Kumar. 2022. Preserving in-context learning ability in large language model fine-tuning. *arXiv preprint arXiv:2211.00635*.

Yubo Wang, Xueguang Ma, and Wenhu Chen. 2023f. Augmenting black-box llms with medical textbooks for clinical question answering. *arXiv preprint arXiv:2309.02233*.

Yufei Wang, Wanjun Zhong, Liangyou Li, Fei Mi, Xingshan Zeng, Wenyong Huang, Lifeng Shang, Xin Jiang, and Qun Liu. 2023g. Aligning large language models with human: A survey. *arXiv preprint arXiv:2307.12966*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. Constructing Datasets for Multi-hop Reading Comprehension Across Documents. *Transactions of the Association for Computational Linguistics*, 6:287–302.

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.

Cheng Wen, Xianghui Sun, Shuaijiang Zhao, Xiaoquan Fang, Liangyu Chen, and Wei Zou. 2023. Chathome: Development and evaluation of a domain-specific language model for home renovation. *arXiv preprint arXiv:2307.15290*.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance.

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

Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. 2023b. Pixiu: A large language model, instruction data and evaluation benchmark for finance. *arXiv preprint arXiv:2306.05443*.

Honglin Xiong, Sheng Wang, Yitao Zhu, Zihao Zhao, Yuxiao Liu, Linlin Huang, Qian Wang, and Dinggang Shen. 2023a. Doctorglm: Fine-tuning your chinese doctor is not a herculean task.

Honglin Xiong, Sheng Wang, Yitao Zhu, Zihao Zhao, Yuxiao Liu, Qian Wang, and Dinggang Shen. 2023b. Doctorglm: Fine-tuning your chinese doctor is not a herculean task. *arXiv preprint arXiv:2304.01097*.

Ming Xu. 2023. Medicalgpt: Training medical gpt model. https://github.com/shibing624/MedicalGPT.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, Fan Yang, et al. 2023a. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Linyao Yang, Hongyang Chen, Zhao Li, Xiao Ding, and Xindong Wu. 2023b. Chatgpt is not enough: Enhancing large language models with knowledge graphs for fact-

aware language modeling.

Songhua Yang, Hanjia Zhao, Senbin Zhu, Guangyu Zhou, Hongfei Xu, Yuxiang Jia, and Hongying Zan. 2023c. Zhongjing: Enhancing the chinese medical capabilities of large language model through expert feedback and real-world multi-turn dialogue. *arXiv preprint arXiv:2308.03549*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023a. React: Synergizing reasoning and acting in language models.

Yunzhi Yao, Peng Wang, Bozhong Tian, Siyuan Cheng, Zhoubo Li, Shumin Deng, Huajun Chen, and Ningyu Zhang. 2023b. Editing large language models: Problems, methods, and opportunities.

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

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. 2023. Do large language models know what they don't know? In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8653–8665, Toronto, Canada. Association for Computational Linguistics.

Wenhao Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, Soumya Sanyal, Chenguang Zhu, Michael Zeng, and Meng Jiang. 2023. Generate rather than retrieve: Large language models are strong context generators. In *International Conference for Learning Representation (ICLR)*.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34:27263–27277.

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

Cyril Zakka, Akash Chaurasia, Rohan Shad, Alex R. Dalal, Jennifer L. Kim, Michael Moor, Kevin Alexander, Euan Ashley, Jack Boyd, Kathleen Boyd, Karen Hirsch, Curt Langlotz, Joanna Nelson, and William Hiesinger. 2023. Almanac: Retrieval-augmented language models for clinical medicine.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*.

Yuexiang Zhai, Shengbang Tong, Xiao Li, Mu Cai, Qing Qu, Yong Jae Lee, and Yi Ma. 2023. Investigating the catastrophic forgetting in multimodal large language models. *arXiv preprint arXiv:2309.10313*.

Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Jianquan Li, Guiming Chen, Xiangbo Wu, Zhiyi Zhang, Qingying Xiao, et al. 2023a. Huatuogpt, towards taming language model to be a doctor. *arXiv preprint arXiv:2305.15075*.

Jiaxing Zhang, Ruyi Gan, Junjie Wang, Yuxiang Zhang, Lin Zhang, Ping Yang, Xinyu Gao, Ziwei Wu, Xiaoqun Dong, Junqing He, Jianheng Zhuo, Qi Yang, Yongfeng Huang,

Xiayu Li, Yanghan Wu, Junyu Lu, Xinyu Zhu, Weifeng Chen, Ting Han, Kunhao Pan, Rui Wang, Hao Wang, Xiaojun Wu, Zhongshen Zeng, and Chongpei Chen. 2022a. Fengshenbang 1.0: Being the foundation of chinese cognitive intelligence. *CoRR*, abs/2209.02970.

Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A Smith. 2023b. How language model hallucinations can snowball. *arXiv preprint arXiv:2305.13534*.

Shuang Zhang, Rui Fan, Yuti Liu, Shuang Chen, Qiao Liu, and Wanwen Zeng. 2023c. Applications of transformer-based language models in bioinformatics: a survey. *Bioinformatics Advances*, 3(1):vbad001.

Shuo Zhang, Liangming Pan, Junzhou Zhao, and William Yang Wang. 2023d. Mitigating language model hallucination with interactive question-knowledge alignment.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022b. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Tianhua Zhang, Hongyin Luo, Yung-Sung Chuang, Wei Fang, Luc Gaitskell, Thomas Hartvigsen, Xixin Wu, Danny Fox, Helen Meng, and James Glass. 2023e. Interpretable unified language checking.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

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. 2023f. Siren's song in the ai ocean: A survey on hallucination in large language models.

Zhengyan Zhang, Zhiyuan Zeng, Yankai Lin, Huadong Wang, Deming Ye, Chaojun Xiao, Xu Han, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. 2023g. Plug-and-play knowledge injection for pre-trained language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10641–10658, Toronto, Canada. Association for Computational Linguistics.

Chao Zhao, Spandana Gella, Seokhwan Kim, Di Jin, Devamanyu Hazarika, Alexandros Papangelis, Behnam Hedayatnia, Mahdi Namazifar, Yang Liu, and Dilek Hakkani-Tur. 2023a. "what do others think?": Task-oriented conversational modeling with subjective knowledge.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023b. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and

Nan Duan. 2023a. Agieval: A human-centric benchmark for evaluating foundation models.

Zexuan Zhong, Zhengxuan Wu, Christopher D. Manning, Christopher Potts, and Danqi Chen. 2023b. Mquake: Assessing knowledge editing in language models via multi-hop questions.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023a. Lima: Less is more for alignment.

Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2023b. Context-faithful prompting for large language models. *arXiv preprint arXiv:2303.11315*.