

Cognitive Mirage: A Review of Hallucinations in Large Language Models*

Hongbin Ye^{1,*}, Tong Liu¹, Aijia Zhang¹, Wei Hua¹ and Weiqiang Jia¹

¹Zhejiang Lab, No. 1 Kechuang Avenue, Yuhang District, Hangzhou City, Zhejiang Province, China

Abstract

As large language models continue to develop in the field of AI, text generation systems are susceptible to a worrisome phenomenon known as *hallucination*. In this study, we summarize recent compelling insights into hallucinations in LLMs. We present a novel taxonomy of hallucinations from various text generation tasks, thus providing theoretical insights, detection methods and improvement approaches. Based on this, future research directions are proposed. Our contributions are threefold: (1) We provide a complete taxonomy for hallucinations appearing in text generation tasks; (2) We provide theoretical analyses of hallucinations in LLMs and provide existing detection and improvement methods; (3) We propose several research directions that can be developed in the future. Our literature library is available at <https://github.com/hongbinye/Cognitive-Mirage-Hallucinations-in-LLMs>.

Keywords

Taxonomy of Hallucination, Large Language Models, Hallucination Detection, Hallucination Correction

1. Introduction

In the ever-evolving realm of large language models (LLMs), a constellation of innovative creations has emerged, such as GPT-3 [1], InstructGPT [2], FLAN [3], PaLM [4], LLaMA [5] and other notable contributors [6, 7, 8, 9]. These models implicitly encode global knowledge within their parameters during the pre-training phase [10, 11], offering valuable insights as knowledge repositories for downstream tasks [12, 13, 14]. Nevertheless, the generalization of knowledge can result in *memory distortion*, an inherent limitation that may give rise to potential inaccuracies [15]. Moreover, their ability to represent knowledge is constrained by model scale and faces challenges in addressing long-tailed knowledge problems [16, 17]. While the privacy and timeliness of data in the real world [18, 19] unfortunately exacerbate this problem, leaving models difficult to maintain a comprehensive and up-to-date understanding of the facts. These challenges present a serious obstacle to the reliability of LLMs, which we refer to as *hallucination*. [20]. A prominent example of this drawback is that models typically generate statements that appear reasonable but are either cognitively irrelevant or factually incorrect. In light of this observation, hallucinations remain a critical challenge in medical [21, 22], financial [23] and other knowledge-intensive fields due to the exacting accuracy requirements. Particularly, the applications for legal case drafting showcase plausible interpretation as an aggregation of diverse subjective perspectives [24].

Definition of Hallucination. As depicted in Figure 1, *hallucination* refers to the generation of texts or responses that exhibit grammatical correctness, fluency, and authenticity, but deviate from the provided source inputs (*faithfulness*) or do not align with factual accuracy (*factualness*) [25]. In traditional NLP tasks [26], hallucinations are often synonymous with *faithfulness*: conflicting information leads to *Intrinsic Hallucination*, i.e., LMs conflict with the input information when generating a response; Conversely, generating ambiguous supplementary information may lead to *Extrinsic Hallucination*, i.e., LMs produce personal names, historical events, or technical documents that are challenging to

The First international OpenKG Workshop: Large Knowledge-Enhanced Models, August 03, 2024, Jeju Island, South Korea

*Corresponding author.

✉ yehongbin@zhejianglab.com (H. Ye); liutong@zhejianglab.com (T. Liu); zhangaijia@zhejianglab.com (A. Zhang); huawei@zhejianglab.com (W. Hua); jiaweiqiang@zhejianglab.com (W. Jia)

>ID 0000-0001-5727-5599 (H. Ye)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

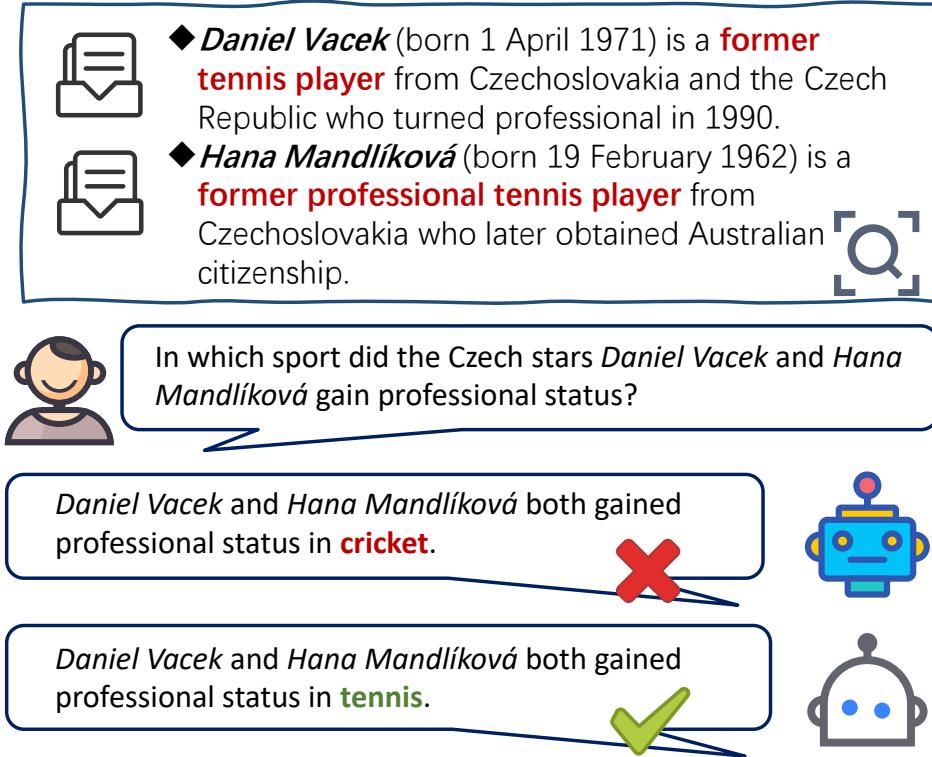


Figure 1: Illustration of Hallucination in LLMs. While the initial response appears fluent, it fails to align with the world knowledge retrieved from the external knowledge base.

verify. LLMs-oriented hallucinations instead prioritize *factualness*, focusing on whether the result can be evidenced or negated by reference to external facts in the real world. Uncritical trust in LLMs can give rise to a phenomenon **Cognitive Mirage**, contributing to misguided decision-making and a cascade of unintended consequences [27].

Present work To effectively control the risk of hallucinations, we summarize recent progress in hallucination theories and solutions in this paper. We propose to organize relevant work by a comprehensive survey (Figure 2):

- **Theoretical insight and mechanism analysis.** We provide in-depth theoretical and mechanism analysis from three typical perspectives: data collection, knowledge gap and optimization process, reviewing the recent and relevant theories related to hallucinations (§2).
- **Taxonomy of hallucination in LLMs.** We conduct a comprehensive review of hallucination in LLMs together with a task axis. We review the task-specific benchmarks with a comprehensive comparison and summary (§3).
- **Wide coverage on emerging hallucination detection and correction methods.** We propose a comprehensive investigation into the proactive detection (§4) and mitigation of hallucinations (§5) in the era of LLMs. This is critical to study the most popular techniques for inspiring future research directions (§6).

Related work As this topic is relatively nascent, only a few surveys exist. Closest to our work, [25] analyzes hallucinatory content in task-specific research progress, which focuses on early works in natural language generation field. Currently there are significant efforts to address hallucination in LLMs. [28] covers methods for effectively collecting high-quality instructions for LLM alignment, including the use of NLP benchmarks, human annotations, and leveraging strong LLMs. [29] discusses self-correcting methods where LLM itself is prompted or guided to correct the hallucinations from its

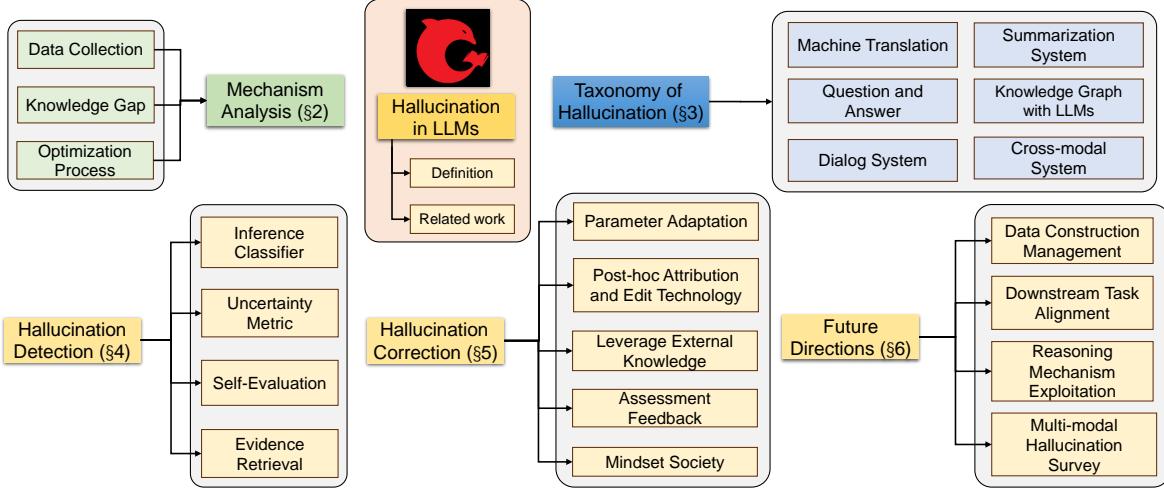


Figure 2: The overview structure of this review. We firstly analyze three crucial factors that contribute to hallucinations and refine the categorization of hallucinations across text generation tasks. Subsequently, we dutifully report current methods for detecting and mitigating hallucinations. Finally, we propose several potential research directions to address evolving problems of hallucinations.

own outputs. Despite some benchmarks [30, 31, 32] is constructed to evaluate whether LLMs are able to generate factual responses, these works scattered among various tasks have not been systematically reviewed and analyzed. Different from those surveys, in this paper, we conduct a literature review on hallucinations in LLMs, hoping to systematically understand the methodologies, compare different methods and inspire new ideas.

2. Mechanism Analysis

For the sake of clean exposition, this section provides theoretical insight into mechanism analysis for hallucinations in LLMs. As a regular LLM, the generative objective is modeled by a parameterized probabilistic model p_{gen} , and sampled to predict the next token in the sentence, thus generating the entire sentence:

$$p_{gen}(y_i) = \mathcal{F}_{\theta}(\mathcal{I}, \mathcal{D}, x, y_{i-1}) \quad (1)$$

where y_i represents probable tokens at each step that can be selected by beam search from vocabulary \mathcal{V} . Note that the instructions \mathcal{I} utilize a variety of predefined templates according to different tasks [33]. Multifarious and high-quality in-context demonstrations \mathcal{D} are aimed at providing analogy samples to reduce the cost of adapting models to new tasks [34]. Parameters θ implicitly memorize corpus knowledge through diverse architectural \mathcal{F} such as decoder-only, encoder-only, or encoder-decoder LLMs. As LLM-based systems can exhibit a variety of hallucinations, we summarise three primary mechanism types for theoretical analysis, and each mechanism is correlated with a distinct training factor.

Data Collection The parameters are implicitly stored within the model as a priori knowledge acquired during the pre-training process. Given the varying quality and range of knowledge within the pre-trained corpus, the information incorporated into the LLMs may be incomplete or outdated. In cases where pertinent memories are unavailable, the LLM’s performance may deteriorate, resorting to rudimentary corpus-based heuristics that rely on term frequencies to render judgements [35]. Another bias stems from the capacity for contextual learning [36] when a few demonstrations are introduced as input to the prefix context. Previous research [37, 38] has demonstrated that the acquisition of knowledge through model learning demonstrations depends on disparities in label categories and the order of demonstration samples. Likewise, multilingual LLMs encounter challenges related to hallucinations, particularly in

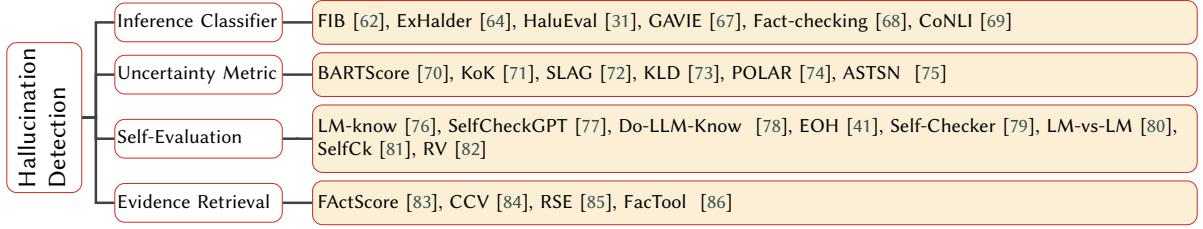


Figure 3: Taxonomy of Hallucination Detection.

handling language pairs with limited resources or non-English translations [39]. Furthermore, cutting-edge Large Vision-Language Models (LVLMs) exhibit instances of hallucinating common objects within visual instructional datasets and prone to objects that frequently co-occur in the same image [40, 41].

Knowledge Gap Knowledge gaps are typically attributed to differences in input format between the pre-training and fine-tuning stages [42]. Even when considering the automatic updating of textual knowledge bases, the output can deviate from the expected corrections [43]. For example, questions often do not align effectively with stored knowledge, and the available information remains unknown until the questions are presented. This knowledge gap poses thorny challenges in balancing memory with retrieved evidence, which is construed as a passive defense mechanism against the misuse of retrieval [44]. To delve into this issue, [45] and [46] propose that disregarding retrieved evidence introduces biased model knowledge, while mis-covering and over-thinking disrupt model behavior. Furthermore, in scenarios where a cache component is utilized to offer historical memory during training [47], the model also experiences inconsistency between the present hidden state and the hidden state stored in the cache.

Optimization Process The maximum likelihood estimation and teacher-forcing training have the potential to result in a phenomenon known as *stochastic parrotting* [48], wherein the model is prompted to imitate the training data without comprehension [49]. Specifically, exposure bias between the training and testing stages have been demonstrated to lead to hallucinations within LLMs, particularly when generating lengthy responses [50]. Besides, sampling techniques characterized by high uncertainty [51], such as top-p and top-k, exacerbate the issue of hallucination. Furthermore, [27] observes that LLMs tend to produce snowballing hallucinations to maintain coherence with earlier hallucinations, and even when directed with prompts as "Let's think step by step", they still generate ineffectual chains of reasoning [13].

3. Taxonomy of Hallucination

In this paper, we mainly consider representative hallucinations, which are widely observed in various downstream tasks, i.e. *Machine Translation*, *Question and Answer*, *Dialog System*, *Summarization System*, *Knowledge graph with LLMs*, and *Visual Question Answer*. As shown in Table 1, these hallucinations are identified complex taxonomy in numerous mainstream tasks associated with LLMs. In the following sections, we will introduce representative types of hallucinations to be resolved.

- **Machine Translation.** Since perturbations (e.g., spellings or capital errors) can induce hallucinations reliably, traditional machine translation models tend to validate instances memorised by the model when subjected to perturbations [87, 88]. It is worth noting that hallucinations generated by LLMs are mainly translation off-target, over-generation, or failed translation attempts [39]. While in low-resource language setting, most models exhibit subpar performance due to the lack of annotated data [54]. In contrast, they are vulnerable to increased amount of pre-trained languages in multilingual setting [89]. Subsequently, familial LLMs trained on different scales of monolingual data are proved to be viscous [39], as the source of *oscillatory hallucination* pathology.

Paper	Task	Architecture Resources		Hallucination Types		Research Method
Raunak et al. [52]	Machine Translation	Enc-Dec	IWSLT-2014	Under perturbation, Natural hallucination		Source perturbation
Guerreiro et al. [53]	Machine Translation	Enc-Dec	WMT2018	Oscillatory hallucination, Largely fluent hallucination		Consider a natural scenario
Dale et al. [54]	Machine Translation	Enc-Dec	FLORES-200, Jigsaw, Wikipedia	Full hallucination, Partial hallucination, Word-level hallucination		Introduce pathology detection
Pfeiffer et al. [55]	Multilingual Seq2seq	Enc-Dec	XQuAD, TyDi, XNLI, XL-Sum, MASSIVE	Source language hallucination		Evaluate source language hallucination
Lin et al. [30]	Question and Answer	Enc-Dec, Only-Dec	TruthfulQA	Imitative falsehoods		Cause imitative falsehoods
Zheng et al. [42]	Question and Answer	Only-Dec	HotpotQA, BoolQ	Comprehension, Factualness, Specificity, Inference Hallucination		Manual analysis of responses
Adlakha et al. [56]	Question and Answer	Enc-Dec, Only-Dec	NQ, HotpotQA, TopiOCQA	Semantic equivalence, Symbolic equivalence, Intrinsic ambiguity, Granularity discrepancies, Incomplete, Enumeration, Satisfactory Subset		Evaluate retrieval augmented QA
Umapathi et al. [22]	Question and Answer	Only-Dec	MEDMCQA, Headqa, US-MILE, Medqa, Pubmed	Reasoning hallucination, Memory-based hallucination		Medical benchmark <i>Med-HALT</i>
Dziri et al. [57]	Dialog System	Enc-Dec, Only-Dec	WoW, CMU-DOG, TopicalChat	Hallucination, Partial hallucination, Generic, Uncooperative		Infer exclusively from the knowledge-snippet
Das et al. [58]	Dialog System	Only-Dec	OpenDialKG	Extrinsic-Soft/Hard/ Grouped, Intrinsic-Soft/ Hard/Repetitive, History Corrupted		Analyze entity-level fact hallucination
Dziri et al. [59]	Dialog System	Enc-Dec, Only-Dec	WoW	Hallucination, Generic, Uncooperativeness		Hallucination-free benchmark <i>FaithDial</i>
Dziri et al. [60]	Dialog System	Enc-Dec, Only-Enc, Only-Dec	WoW, CMU-DOG, TopicalChat	Fully attributable, Not attributable, Generic		Knowledge-grounded interaction benchmark <i>Begin</i>
Sun et al. [61]	Dialog System	Enc-Dec, Only-Dec	WoW	Intrinsic hallucination, Extrinsic hallucination		Sample responses for conversation
Tam et al. [62]	Summarization System	Enc-Dec, Only-Dec	CNN/DM, XSum	Factually inconsistent summaries		Generate summaries from given models
Cao et al. [63]	Summarization System	Enc-Dec, Only-Dec	MENT	Non-hallucinated, Factual hallucination, Non-factual hallucination, Intrinsic hallucination		Label factual entities from summarizations
Shen et al. [64]	Summarization System	Enc-Dec, Only-Enc	NHNet	News headline hallucination		Majority vote of journalism degree holders
Qiu et al. [65]	Summarization System	Multiple ADapters	XL-Sum	Intrinsic hallucination, Extrinsic hallucination		In a cross-lingual transfer setting
Yu et al. [66]	Knowledge-based text generation	Enc-Dec, Only-Dec	Encyclopedic, ETC	Knowledge hallucination		Evaluate knowledge creating ability given known facts
Mihindukulasooriya et al. [32]	Knowledge graph generation	Only-Dec	TekGen, WebNLG	Subject hallucination, relation hallucination, object hallucination		Ontology driven KGC benchmark <i>Text2KGBench</i>
Li et al. [41]	Visual Question Answer	Enc-Dec	MSCOCO	Object hallucination		Caption hallucination assessment

Table 1
List of Representative Hallucination

- **Question and Answer.** Imperfect responses suffer from flawed external knowledge, knowledge recall cues and reasoning instruction [42]. For example, LLMs are mostly unable to avoid answering when provided with no relevant information, instead provide incomplete and plausible answers [56]. In

additon to external knowledge, memorized information without accurate, reliable and accessible source also contributes to different types of hallucinations [22]. Though scaling laws suggest that perplexity on the training distribution is positively correlated with parameter size, [30] further discovers that scaling up models should increase the rate of imitative falsehoods.

- **Dialog System.** Some studies view dialogue models as unobtrusive imitators, which simulates the distributional properties of data instead of generating faithful output. For example, *Uncooperativeness responses* [57] originating from discourse phenomena inclines to output an exact copy of the entire evidence. [58] reports more nuanced hallucinations in KG-grounded dialogue systems as analyzed through human feedback. Similarly, FaithDial [59], BEGIN [60], MixCL [61] all implement experiments on the wow dataset to conduct a meta-evaluation of the hallucination in knowledge grounded dialogue.

- **Summarization System.** Automatically generated abstracts based on LLMs may be fluent, but they still typically lack faithfulness to the source document. Compared to the human evaluation of traditional summarization models [26], the summarizations generated by LLMs can be categorized into two major types: *intrinsic hallucinations* that distort the information present in the document; *extrinsic hallucinations* that provide additional information that cannot be directly attributed to the document [65]. Note that extrinsic hallucination as a metrics of factually consistent continuation of inputs in LLMs is given more attention in summarisation systems [62, 64]. Furthermore, [63] subdivides extrinsic hallucinations into *factual* and *non-factual* hallucinations. The former provides additional world knowledge, which may benefit comprehensive understanding.

- **Knowledge Graph with LLMs.** Despite the promising progress in knowledge-based text geneartion, it encounters *intrinsic hallucinations* inherent to the process where the generated text not only covers the input information but also incorporates redundant details derived from its internal memorized knowledge [90]. To address this, [66] establish a distinction between correctly generated knowledge and *knowledge hallucinations* in terms of knowledge creation. Notably, the *Virtual Knowledge Extraction* [91] underscores the potential generalization capabilities of LLMs in the realms of constructing and inferring from knowledge graphs. [32] further empower LLMs to produce interpretable fact-checks through a neural symbolic approach. Based on their fidelity to the source, hallucinations are defined as *subject hallucination*, *relation hallucination*, and *object hallucination*.

- **Cross-modal System.** Augmented by the superior language capabilities of LLMs, performance of cross-modal tasks achieves promising progress [92, 40]. However, despite replacing the original language encoder with LLMs, Large Visual Language Models (LVLMs) [93] still generate object descriptions that not present in the target image, denoted as *object hallucinations* [41]. Especially, the various failure cases could be typically found in Visual Question Answering [41, 67], Image Captioning [94, 95, 96], Report Generation [68] etc.

4. Hallucination Detection

Conventional hallucination detection mainly depends on task-specific metrics, such as ROUGE and BLEU to evaluate the information overlap between source and target texts in summarization tasks [97], while knowledge F1 to estimate the knowledge-aware ability of response generation [98]. These metrics focus on measuring *faithfulness* of references and fail to provide an assessment of *factualness*. Despite some reference-free works are proposed, plugin-based methods [99] suffer from world knowledge limitation. QA-based matching metrics [100] lack knowledge completeness of source information. NLI-based methods [60] are unable to support finer-grained hallucination checking as they are sentence-level, besides entailment and hallucination problems are not equivalent. As the paradigm shift in hallucination detection arising from the rapid development of LLMs, we present a novel taxonomy in Fig 3 and introduce each category in following sections.

- **Inference Classifier.** The most straightforward strategy involves adopting classifiers to assess the likelihood of hallucinations. Concretely, given a question Q and an answer A , an inferential classifier C can be asked to determine whether the answer contains hallucinatory content \mathcal{H} via computing

$p(\mathcal{H}) = \mathcal{F}_{\mathcal{C}}(\mathcal{Q}, \mathcal{A})$. Therefore, [64] employs the state-of-the-art LLMs to do end-to-end text generation of detection results. Some other studies [31] finds that adding chains of thought indiscriminately may intervene in the final judgement, whereas retrieving the knowledge properly resulted in gains. Furthering this concept, the hinted classifier and explainer [64], used to generate intermediate process labels and high-quality natural language explanations, are demonstrated to enhance the final predicted class from a variety of perspectives. Subsequently, [62] suggests adopting a different classifier model to the generated model, contributing to easier judgement of factual consistency. For radiology report generation, binary classifiers [68] can be leveraged to measure the reliability by combining image and text embedding. Unlike previous work that employs complex human-crafted rules to evaluate object hallucinations, GAVIE [67] scores responses towards image content based on both accuracy and relevance criteria, which evaluates the LMMs output in an open-ended manner.

- **Uncertainty Metric.** It is important to examine the correlation between the hallucination metric and the quality of output from a variety of perspectives. One intuitive approach is to employ the probabilistic output of the model itself, as ASTSN [75] calculates the models' uncertainty about the identified concepts by utilising the logit output values. Similarly, BARTSCORE [70] employs a universal notion that models trained to convert generated text to reference output or source text will score higher when the generated text is superior. It is an unsupervised metric that supports the addition of appropriate prompts to improve the measure design, without human judgement to train. Furthermore, KoK [71] based on the work of [101] evaluates answer uncertainty from three categories, i.e., subjectivity, hedges and text uncertainty. However, SLAG [72] measures consistent factual beliefs in terms of paraphrase, logic, and entailment. In addition to this, KLD [73] combines information theory-based metrics (e.g., entropy and KL-divergence) to capture knowledge uncertainty. Beside expert-stipulated programmatic supervision, POLAR [74] introduces Pareto optimal learning assessed risk score for estimating the confidence level of a response.

- **Self-Evaluation.** To self-evaluate is challenging since the model might be overconfident about its generated samples being correct. The motivating idea of SelfCheckGPT [77] is to use the ability of the LLMs themselves to sample multiple responses and identify fictitious statements by measuring the consistency of information among responses. [76] further illustrates that both the increase in size and the demonstration of assessment can improve self-assessment. Beyond repetitive multiple direct queries, [78] uses open-ended indirect queries and compares their answers to each other for an agreed-upon score outcome. SelfCk [81] imposes appropriate constraints on the same LLM to generate pairs of sentences triggering self-contradictions, which prompt the detection. In contrast, Polling-based querying [41] reduce the complexity of judgement by randomly sampling query objects. Besides, Self-Checker [79] decomposes complex statements into multiple simple statements, fact-checking them one by one. However, [80] introduces two LLMs to drive the complex fact-checking reasoning process by crosscheck.

- **Evidence Retrieval.** Evidence retrieval accomplishes factual detection by retrieving supporting evidence related to hallucinations. To this end, Designing a claim-centric pipeline allows for a question-retrieve-summary chain to effectively collect original evidence [84, 85]. Consequently, FActScore [83] calculates the percentage of atomic facts supported by the given knowledge source. Towards adapting the tasks that users in interaction with generative models, FacTool [86] proposes to integrate a variety of tools into a task-agnostic and domain-agnostic detection framework, in order to assemble evidence about the authenticity of the generated content.

5. Hallucination Correction

In this section, we delve into the methods to correct hallucination in terms of different aspects. As shown in Figure 4, these hallucination correction paradigms have demonstrated strong dominance in many mainstream NLP tasks. Note that these methods are not entirely orthogonal but could complement each other as required by the tasks in practical applications. In the following sections, we will introduce each methods as shown in Figure 5.

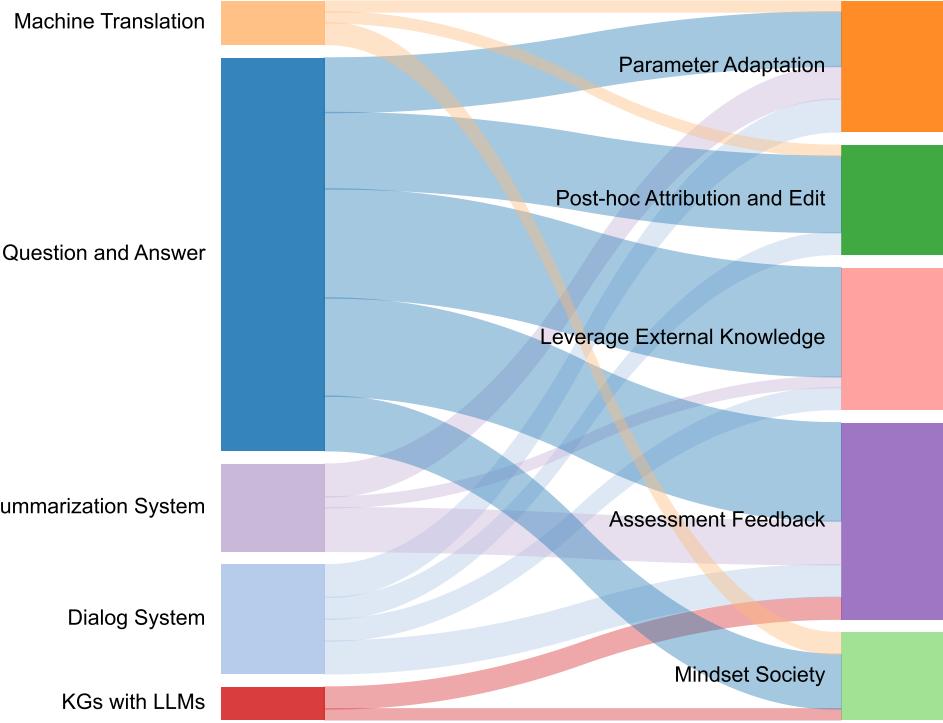


Figure 4: Sankey diagram of hallucination correction methods with different mainstream NLP tasks.

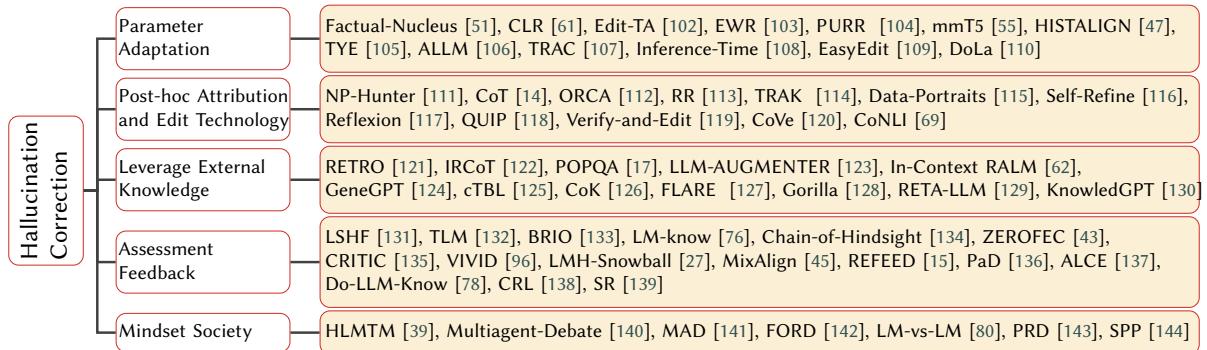


Figure 5: Taxonomy of Hallucination Correction.

• **Parameter Adaptation.** Parameters in LLMs store biases learned in pre-training, are often unaligned with user intent. A cutting-edge strategy is to guide effective knowledge through parameter conditioning, editing, and optimisation. For example, CLR [61] optimises to reduce the generation probability of negative samples at span level utilising contrastive learning parameters. While introducing contextual knowledge background that contradicts the model’s intrinsic prior knowledge, TYE [105] effectively reduces the weight of prior knowledge through context-aware decoding method. Besides, PURR [104] corrupts noise into the text, fine-tune compact editors, and denoise by merging relevant evidence. To introduce additional cache component, HISTALIGN [47] discovers that its hidden state is not aligned with the current hidden state, and proposes sequence information contrastive learning to improve the reliability of memory parameters. Consequently, Edit-TA [102] mitigates the biases learnt in pre-training from a task algorithm perspective. An intuition behind it is that parameter variations learnt through negative example tasks could be perceived through weight variances. However as this fails to take the importance of different negative examples into account, therefore EWR [103] proposes Fisher information matrices to measure the uncertainty of their estimation, which is applied for the dialogue systems to execute a parameter interpolation and remove hallucination. EasyEdit [109] summarises

methods for parameter editing, while minimising the influence to irrelevant parameter.

An efficient alternative is to identify task-specific parameters and exploit them. For example, ALLM [106] aligns the parameter module with task-specific knowledge, and then generates the relevant knowledge as additional context in background augmented prompts. Similarly, mmT5 [55] utilises language-specific modules during pre-training to separate language-specific information from language-independent information, demonstrating that adding language-specific modules can alleviate the curse of multilinguality. Instead, TRAC [107] combines conformal prediction and global testing to augment retrieval-based QA. The conservative strategy formulation ensures that a semantically equivalent answer to the truthful answer is included in the prediction set.

Another parameter adaptation idea focuses on flexible sampling consistent with user requirements. For instance, [51] observes that the randomness of sampling is more detrimental to factuality when generating the latter part of a sentence. The factual-nucleus sampling algorithm is introduced to keep the faithfulness of the generation while ensuring the quality and diversity. Besides, Inference-Time [108] firstly identifies a set of attentional heads with high linear probing accuracy, and then shifts activation in the inference process along the direction associated with factual knowledge.

• **Post-hoc Attribution and Edit Technology.** A source of hallucination is that LLMs may leverage the patterns observed in the pre-training data for inference in a novel form. Recently, ORCA [112] reveals problematic patterns in the behaviour of models by probing supporting data evidences from pre-training data. Similarly, TRAK [114] and Data-Portraits [115] analyse whether models plagiarise or reference existing resources by means of data attribution. QUIP [118] further demonstrates that providing text that has been observed in the pre-training phase can improve the ability of LLMs to generate more factual information. Furthermore, motivated by the gap between LLMs and human modes of thinking, one intuition is to align the two modes of reasoning. Thus CoT [14] elicits faithful reasoning via a kind of Chain-of-Thought (CoT) [13] prompts. Similarly, RR [113] retrieves relevant external knowledge based on decomposed reasoning steps obtained from a CoT prompt. Since LLMs do not produce the best output on the first attempt, Self-Refine [116] implements self-refinement algorithms through iterative feedback and improvement. Reflexion [117] also employs verbal reinforcement to generate reflective feedback by learning about prior failings. Verify-and-Edit [119] proposes a CoT-prompted verify-and-edit framework, which improves the fidelity of predictions by post-editing the inference chain based on externally retrieved knowledge. Cove [120] emphasises the importance of independent self-verification to prevent being influenced by other responses. Another source of hallucinations is to describe factual content with incorrect retrievals. To rectify this, NP-Hunter [111] follows a generate-then-refine strategy whereby a generated response is amended using the KG so that the dialogue system is able to correct potential hallucinations by querying the KG.

• **Leverage External Knowledge.** As an attempt to extend the language model for halucination mitigation, a suggestion is to retrieve relevant documents from large textual databases. RETRO [121] splits the input sequence into chunks and retrieves similar documents, while In-Context RALM [62] places the selected document before the input text to improve the prediction. Furthermore, IRCoT [122] interweaves CoT generation and document retrieval steps to guide LLMs. LLM-AUGMENTER [123] also bases the responses of LLMs on integrated external knowledge and automated feedback to improve the truthfulness score of the answers. Another work, CoK [126] iteratively analyses future content of upcoming sentences, and then applies them as a query to retrieve relevant documents for the purposes of re-generating sentences when they contain low confidence tokens. Similarly, RETA-LLM [129] creates a complete pipeline to assist users in building their own domain-based LLM retrieval systems. Note that in addition to document retrieval, diverse external knowledge queries coule be assembled into retrieval-augmented LLM systems. For example, FLARE [127] leverages structured knowledge bases to support complex queries and provide more straightforward factual statements. Further, KnowledGPT [130] adopts program of thoughts (PoT) prompting, which generates codes to interact with knowledge bases. While cTBL [125] proposes to enhance LLMs with tabular data in conversation settings. Besides, GeneGPT [124] demonstrates that expertise can be accessed more easily and accurately by detecting and executing API calls through contextual learning and augmented decoding algorithms. To support potentially millions of ever-changing APIs, Gorilla [128] explores self-instruct fine-tuning and retrieval

for efficient API exploitation.

- **Assessment Feedback.** As language models become more sophisticated, evaluation feedback can significantly improve the quality of generated text, as well as reduce the appearance of hallucinations. To realise this concept, LSHF [131], TLM [132] and Chain-of-Hindsight [134] predict human preferences through reinforcement learning and employs this as the reward function. In addition to enabling the model to learn directly from the feedback of factual metrics in a sample-efficient manner [138], it is also important to build in a self-evaluation function of the model to filter candidate generated texts. For example, BRIO [133] empowers summarization model assessment, estimating probability distributions of candidate outputs to rate the quality of candidate summaries. While LM-know [76] is devoted to investigating whether LLMs can evaluate the validity of their own claims by detecting the probability that they know the answer to a question. Consequently, Do-LLM-Know [78] queries exclusively with black-box LLMs, and the results of queries repeatedly generated multiple times are compared with each other to pass consistency checks. As missing citation quality evaluation affects the final performance, ALCE [137] employs a natural language reasoning model to measure citation quality and extends the integrated retrieval system. Similarly, CRITIC [135] proposes to interact with appropriate tools to assess certain aspects of the text, and then to modify the output based on the feedback obtained during the verification process. Note that automated error checking can also utilise LLMs to generate text that conforms to tool interfaces. PaD [136] distills the LLMs with a synthetic inference procedure, and the synthesis program obtained can be automatically compiled and executed by an explainer. Further, iterative refinement processes are validated to effectively identify appropriate details [96, 45, 15], and can stop early invalid reasoning chains, beneficially reducing the phenomenon of hallucination snowballing [27].

- **Mindset Society.** Human intelligence thrives on the concept of cognitive synergy, where collaboration between different cognitive processes produces better results than isolated individual cognitive processes. "Society of minds" [145] is believed to have the potential to significantly improve the performance of LLMs and pave the way for consistency in language production and comprehension. For the purpose of addressing hallucinations in large-scale multilingual models across different translation scenarios, HLMTM [39] proposes a hybrid setting in which other translation systems can be requested to act as a back-up system when the original system is hallucinating. Consequently, Multiagent-Debate [140] employs multiple LLMs in several rounds to propose and debate their individual responses and reasoning processes to reach a consensus final answer. As a result of this process, the models are encouraged to construct answers that are consistent with both internal criticisms and responses from other agents. Before a final answer is presented, the resultant community of models can hold and maintain multiple reasoning chains and possible answers simultaneously. Based on this idea, MAD [141] adds a judge-managed debate process, demonstrating that adaptive interruptions of debate and controlled "tit-for-tat" states help to complete factual debates. Furthermore, FORD [142] proposes roundtable debates that include more than two LLMs and emphasises that competent judges are essential to dominate the debate. LM-vs-LM [80] also proposes multi-round interactions between LM and another LM to check the factualness of original statements. Besides, PRD [143] proposes a peer rank and discussion-based evaluation framework to arrive at a well-recognised assessment result that all peers are in agreement with. In an effort to maintain strong reasoning, SPP [144] utilises LLMs to assign several fine-grained roles, which effectively stimulates knowledge acquisition and reduces hallucinations.

6. Future Directions

Though numerous technical solutions have been proposed in the survey for hallucinations in LLMs, there exist some potential directions:

- **Data Construction Management.** As previously discussed, the style, and knowledge of LLMs is basically learned during model pre-training. High quality data present promising opportunities for facilitating the reduction of hallucinations in LLMs [146]. Inspired by the basic rule of machine learning models "Garbage input, garbage output", [147] proposes that data quality and diversity outweigh

the importance of fine-tuning large-scale instructions [148, 3, 149] and RLHF [6, 2]. To perform efficiently in knowledge-intensive verticals, we argue that construction of entity-centred fine-tuned instructions [150, 151, 152] is a promising direction that it can enhance the factuality of generated entity information. Another feasible proposal is to incorporate a self-curation phase [153] in the instruction construction process to rate the quality of candidate pairs. During the iteration process, quality evaluation [154] based on manual or automated rule constraints could provide self-correction capacity.

- **Reasoning Mechanism Exploitation.** The emerging CoT technique [14] stimulates the emergent reasoning ability of LLMs by imitating intrinsic stream of thought. Recently, A primary improvement is ToT [155] introduces tree and into the thought process, and provides a novel backtrack function. However, the actual thinking process creates a complex network of ideas, as an example, people could explore a particular chain of reasoning, backtrack or start a new chain of reasoning. GoT [156] extends the dependencies between thoughts by constructing vertices with multiple incoming edges to aggregate arbitrary thoughts. Since previous methods have no storages for intermediate results, CR [156] uses cumulative and iterative manners to simulate human thought processes, and decompose the task into smaller components. In addition to self-heuristic methods, PAL [157] and PoT [158] introduce programming logic into the language space [159], expanding the ability to invoke external explainers. As a summary, research based on human cognition helps to provide brilliant insights into the analysis of hallucinations, such as Dual Process Theory [160], Three layer mental model [161], Computational Theory of Mind [162], and Connectionism [163].

- **Multi-modal Hallucination Survey.** It has become a community consensus to establish powerful Multimodal Large Language Models (MLLMs) [164, 165, 166] by taking advantage of excellent comprehension and reasoning capabilities of LLMs. [41] confirms the severity of hallucinations in MLLM by object detecting and polling-based querying. The results indicate that they are highly susceptible to object hallucination, and the generated description does not match the target image. Besides, [167] that MLLMs have limited multimodal reasoning ability as well as dependence on spurious cues. Though current study [168] provides a broad overview of MLLMs, the causation of hallucinations has not been comprehensively investigated. In the future, as more sophisticated multi-model applications emerge, in-depth analyses of the biased distribution resulting from misalignment among modes is a promising research direction, to provide faithful modal interactions.

7. Conclusion and Vision

In this paper, we provide an overview of hallucinations in LLMs with new taxonomy, theoretical insight, detection methods, correction methods and several future research directions. Note that it is crucial to utilize LLMs in a responsible and beneficial manner. Furthermore, with sophisticated and efficient detection methods proposed for various aspects, LLMs will provide human reliable and secure information in broad application scenarios.

References

- [1] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), NeurIPS 2020, 2020. URL: <https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html>.
- [2] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. F. Christiano, J. Leike, R. Lowe, Training language models to follow instructions with

- human feedback, in: NeurIPS, 2022. URL: http://papers.nips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html.
- [3] J. Wei, M. Bosma, V. Y. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, Q. V. Le, Finetuned language models are zero-shot learners, in: The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022, OpenReview.net, 2022. URL: <https://openreview.net/forum?id=gEZrGCozdqR>.
 - [4] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, P. Schuh, K. Shi, S. Tsvyashchenko, J. Maynez, A. Rao, P. Barnes, Y. Tay, N. Shazeer, V. Prabhakaran, E. Reif, N. Du, B. Hutchinson, R. Pope, J. Bradbury, J. Austin, M. Isard, G. Gur-Ari, P. Yin, T. Duke, A. Levskaya, S. Ghemawat, S. Dev, H. Michalewski, X. Garcia, V. Misra, K. Robinson, L. Fedus, D. Zhou, D. Ippolito, D. Luan, H. Lim, B. Zoph, A. Spiridonov, R. Sepassi, D. Dohan, S. Agrawal, M. Omernick, A. M. Dai, T. S. Pillai, M. Pellat, A. Lewkowycz, E. Moreira, R. Child, O. Polozov, K. Lee, Z. Zhou, X. Wang, B. Saeta, M. Diaz, O. Firat, M. Catasta, J. Wei, K. Meier-Hellstern, D. Eck, J. Dean, S. Petrov, N. Fiedel, Palm: Scaling language modeling with pathways, CoRR abs/2204.02311 (2022). URL: <https://doi.org/10.48550/arXiv.2204.02311>. doi:10.48550/arXiv.2204.02311. arXiv:2204.02311.
 - [5] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, Llama: Open and efficient foundation language models, CoRR abs/2302.13971 (2023). URL: <https://doi.org/10.48550/arXiv.2302.13971>. doi:10.48550/arXiv.2302.13971. arXiv:2302.13971.
 - [6] Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan, N. Joseph, S. Kadavath, J. Kernion, T. Conerly, S. E. Showk, N. Elhage, Z. Hatfield-Dodds, D. Hernandez, T. Hume, S. Johnston, S. Kravec, L. Lovitt, N. Nanda, C. Olsson, D. Amodei, T. B. Brown, J. Clark, S. McCandlish, C. Olah, B. Mann, J. Kaplan, Training a helpful and harmless assistant with reinforcement learning from human feedback, CoRR abs/2204.05862 (2022). URL: <https://doi.org/10.48550/arXiv.2204.05862>. doi:10.48550/arXiv.2204.05862. arXiv:2204.05862.
 - [7] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. T. Diab, X. Li, X. V. Lin, T. Mihaylov, M. Ott, S. Shleifer, K. Shuster, D. Simig, P. S. Koura, A. Sridhar, T. Wang, L. Zettlemoyer, OPT: open pre-trained transformer language models, CoRR abs/2205.01068 (2022). URL: <https://doi.org/10.48550/arXiv.2205.01068>. doi:10.48550/arXiv.2205.01068. arXiv:2205.01068.
 - [8] A. Zeng, X. Liu, Z. Du, Z. Wang, H. Lai, M. Ding, Z. Yang, Y. Xu, W. Zheng, X. Xia, W. L. Tam, Z. Ma, Y. Xue, J. Zhai, W. Chen, Z. Liu, P. Zhang, Y. Dong, J. Tang, GLM-130B: an open bilingual pre-trained model, in: The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023, OpenReview.net, 2023. URL: <https://openreview.net/pdf?id=Aw0rrrPUF>.
 - [9] C. Xu, Q. Sun, K. Zheng, X. Geng, P. Zhao, J. Feng, C. Tao, D. Jiang, Wizardlm: Empowering large language models to follow complex instructions, CoRR abs/2304.12244 (2023). URL: <https://doi.org/10.48550/arXiv.2304.12244>. doi:10.48550/arXiv.2304.12244. arXiv:2304.12244.
 - [10] X. Han, Z. Zhang, N. Ding, Y. Gu, X. Liu, Y. Huo, J. Qiu, Y. Yao, A. Zhang, L. Zhang, W. Han, M. Huang, Q. Jin, Y. Lan, Y. Liu, Z. Liu, Z. Lu, X. Qiu, R. Song, J. Tang, J. Wen, J. Yuan, W. X. Zhao, J. Zhu, Pre-trained models: Past, present and future, AI Open 2 (2021) 225–250. URL: <https://doi.org/10.1016/j.aiopen.2021.08.002>. doi:10.1016/j.aiopen.2021.08.002.
 - [11] J. Huang, K. C. Chang, Towards reasoning in large language models: A survey, in: A. Rogers, J. L. Boyd-Graber, N. Okazaki (Eds.), Findings of ACL 2023, ACL, 2023, pp. 1049–1065. URL: <https://doi.org/10.18653/v1/2023.findings-acl.67>. doi:10.18653/v1/2023.findings-acl.67.
 - [12] D. Pu, V. Demberg, Chatgpt vs human-authored text: Insights into controllable text summarization and sentence style transfer, in: V. Padmakumar, G. Vallejo, Y. Fu (Eds.), ACL 2023, ACL, 2023, pp. 1–18. URL: <https://doi.org/10.18653/v1/2023.acl-srw.1>. doi:10.18653/v1/2023.acl-srw.1.
 - [13] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, Y. Iwasawa, Large language models are zero-shot reasoners, in: NeurIPS, 2022. URL: http://papers.nips.cc/paper_files/paper/2022/hash/8bb0d291acd4acf06ef112099c16f326-Abstract-Conference.html.

- [14] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, D. Zhou, Chain-of-thought prompting elicits reasoning in large language models, in: NeurIPS, 2022. URL: http://papers.nips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.
- [15] W. Yu, Z. Zhang, Z. Liang, M. Jiang, A. Sabharwal, Improving language models via plug-and-play retrieval feedback, CoRR abs/2305.14002 (2023). URL: <https://doi.org/10.48550/arXiv.2305.14002>. doi:10.48550/arXiv.2305.14002. arXiv:2305.14002.
- [16] N. Kandpal, H. Deng, A. Roberts, E. Wallace, C. Raffel, Large language models struggle to learn long-tail knowledge, in: A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, J. Scarlett (Eds.), ICML 2023, volume 202 of *Proceedings of Machine Learning Research*, PMLR, 2023, pp. 15696–15707. URL: <https://proceedings.mlr.press/v202/kandpal23a.html>.
- [17] A. Mallen, A. Asai, V. Zhong, R. Das, D. Khashabi, H. Hajishirzi, When not to trust language models: Investigating effectiveness of parametric and non-parametric memories, in: A. Rogers, J. L. Boyd-Graber, N. Okazaki (Eds.), ACL 2023, ACL, 2023, pp. 9802–9822. URL: <https://doi.org/10.18653/v1/2023.acl-long.546>. doi:10.18653/v1/2023.acl-long.546.
- [18] A. Lazaridou, E. Gribovskaya, W. Stokowiec, N. Grigorev, Internet-augmented language models through few-shot prompting for open-domain question answering, CoRR abs/2203.05115 (2022). URL: <https://doi.org/10.48550/arXiv.2203.05115>. doi:10.48550/arXiv.2203.05115. arXiv:2203.05115.
- [19] W. Shi, S. Min, M. Yasunaga, M. Seo, R. James, M. Lewis, L. Zettlemoyer, W. Yih, REPLUG: retrieval-augmented black-box language models, CoRR abs/2301.12652 (2023). URL: <https://doi.org/10.48550/arXiv.2301.12652>. doi:10.48550/arXiv.2301.12652. arXiv:2301.12652.
- [20] W. Yu, C. Zhu, Z. Li, Z. Hu, Q. Wang, H. Ji, M. Jiang, A survey of knowledge-enhanced text generation, ACM Comput. Surv. 54 (2022) 227:1–227:38. URL: <https://doi.org/10.1145/3512467>. doi:10.1145/3512467.
- [21] D. Dash, R. Thapa, J. M. Banda, A. Swaminathan, M. Cheatham, M. Kashyap, N. Kotecha, J. H. Chen, S. Gombar, L. Downing, R. Pedreira, E. Goh, A. Arnaout, G. K. Morris, H. Magon, M. P. Lungren, E. Horvitz, N. H. Shah, Evaluation of GPT-3.5 and GPT-4 for supporting real-world information needs in healthcare delivery, CoRR abs/2304.13714 (2023). URL: <https://doi.org/10.48550/arXiv.2304.13714>. doi:10.48550/arXiv.2304.13714. arXiv:2304.13714.
- [22] L. K. Umapathi, A. Pal, M. Sankarasubbu, Med-halt: Medical domain hallucination test for large language models, CoRR abs/2307.15343 (2023). URL: <https://doi.org/10.48550/arXiv.2307.15343>. doi:10.48550/arXiv.2307.15343. arXiv:2307.15343.
- [23] S. S. Gill, M. Xu, P. Patros, H. Wu, R. Kaur, K. Kaur, S. Fuller, M. Singh, P. Arora, A. K. Parlakad, V. Stankovski, A. Abraham, S. K. Ghosh, H. Lutfiyya, S. S. Kanhere, R. Bahsoon, O. F. Rana, S. Dustdar, R. Sakellariou, S. Uhlig, R. Buyya, Transformative effects of chatgpt on modern education: Emerging era of AI chatbots, CoRR abs/2306.03823 (2023). URL: <https://doi.org/10.48550/arXiv.2306.03823>. doi:10.48550/arXiv.2306.03823. arXiv:2306.03823.
- [24] S. Curran, S. Lansley, O. Bethell, Hallucination is the last thing you need, CoRR abs/2306.11520 (2023). URL: <https://doi.org/10.48550/arXiv.2306.11520>. doi:10.48550/arXiv.2306.11520. arXiv:2306.11520.
- [25] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. Bang, A. Madotto, P. Fung, Survey of hallucination in natural language generation, ACM Comput. Surv. 55 (2023) 248:1–248:38. URL: <https://doi.org/10.1145/3571730>. doi:10.1145/3571730.
- [26] J. Maynez, S. Narayan, B. Bohnet, R. T. McDonald, On faithfulness and factuality in abstractive summarization, in: D. Jurafsky, J. Chai, N. Schluter, J. R. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5–10, 2020, ACL, 2020, pp. 1906–1919. URL: <https://doi.org/10.18653/v1/2020.acl-main.173>. doi:10.18653/v1/2020.acl-main.173.
- [27] M. Zhang, O. Press, W. Merrill, A. Liu, N. A. Smith, How language model hallucinations can snowball, CoRR abs/2305.13534 (2023). URL: <https://doi.org/10.48550/arXiv.2305.13534>. doi:10.48550/arXiv.2305.13534. arXiv:2305.13534.

- [28] Y. Wang, W. Zhong, L. Li, F. Mi, X. Zeng, W. Huang, L. Shang, X. Jiang, Q. Liu, Aligning large language models with human: A survey, CoRR abs/2307.12966 (2023). URL: <https://doi.org/10.48550/arXiv.2307.12966>. doi:10.48550/arXiv.2307.12966. arXiv:2307.12966.
- [29] L. Pan, M. Saxon, W. Xu, D. Nathani, X. Wang, W. Y. Wang, Automatically correcting large language models: Surveying the landscape of diverse self-correction strategies, CoRR abs/2308.03188 (2023). URL: <https://doi.org/10.48550/arXiv.2308.03188>. doi:10.48550/arXiv.2308.03188. arXiv:2308.03188.
- [30] S. Lin, J. Hilton, O. Evans, Truthfulqa: Measuring how models mimic human falsehoods, in: ACL 2022, ACL, 2022, pp. 3214–3252. URL: <https://doi.org/10.18653/v1/2022.acl-long.229>. doi:10.18653/v1/2022.acl-long.229.
- [31] J. Li, X. Cheng, W. X. Zhao, J. Nie, J. Wen, Halueval: A large-scale hallucination evaluation benchmark for large language models, CoRR abs/2305.11747 (2023). URL: <https://doi.org/10.48550/arXiv.2305.11747>. doi:10.48550/arXiv.2305.11747. arXiv:2305.11747.
- [32] N. Mihindukulasooriya, S. Tiwari, C. F. Enguiça, K. Lata, Text2kgbench: A benchmark for ontology-driven knowledge graph generation from text, CoRR abs/2308.02357 (2023). URL: <https://doi.org/10.48550/arXiv.2308.02357>. doi:10.48550/arXiv.2308.02357. arXiv:2308.02357.
- [33] F. Yin, J. Vig, P. Laban, S. Joty, C. Xiong, C. Wu, Did you read the instructions? rethinking the effectiveness of task definitions in instruction learning, in: A. Rogers, J. L. Boyd-Graber, N. Okazaki (Eds.), ACL 2023, ACL, 2023, pp. 3063–3079. URL: <https://doi.org/10.18653/v1/2023.acl-long.172>. doi:10.18653/v1/2023.acl-long.172.
- [34] M. Chen, J. Du, R. Pasunuru, T. Mihaylov, S. Iyer, V. Stoyanov, Z. Kozareva, Improving in-context few-shot learning via self-supervised training, in: M. Carpuat, M. de Marneffe, I. V. M. Ruiz (Eds.), NAACL 2022, ACL, 2022, pp. 3558–3573. URL: <https://doi.org/10.18653/v1/2022.nacl-main.260>. doi:10.18653/v1/2022.nacl-main.260.
- [35] N. McKenna, T. Li, L. Cheng, M. J. Hosseini, M. Johnson, M. Steedman, Sources of hallucination by large language models on inference tasks, CoRR abs/2305.14552 (2023). URL: <https://doi.org/10.48550/arXiv.2305.14552>. doi:10.48550/arXiv.2305.14552. arXiv:2305.14552.
- [36] S. Chan, A. Santoro, A. K. Lampinen, J. Wang, A. Singh, P. H. Richemond, J. L. McClelland, F. Hill, Data distributional properties drive emergent in-context learning in transformers, in: NeurIPS, 2022. URL: http://papers.nips.cc/paper_files/paper/2022/hash/77c6ccacf9962e2307fc64680fc5ace-Abstract-Conference.html.
- [37] S. Wang, K. Wei, H. Zhang, Y. Li, W. Wu, Let me check the examples: Enhancing demonstration learning via explicit imitation, in: A. Rogers, J. L. Boyd-Graber, N. Okazaki (Eds.), ACL 2023, ACL, 2023, pp. 1080–1088. URL: <https://doi.org/10.18653/v1/2023.acl-short.93>. doi:10.18653/v1/2023.acl-short.93.
- [38] Y. Lu, M. Bartolo, A. Moore, S. Riedel, P. Stenetorp, Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity, in: S. Muresan, P. Nakov, A. Villavicencio (Eds.), ACL 2022, ACL, 2022, pp. 8086–8098. URL: <https://doi.org/10.18653/v1/2022.acl-long.556>. doi:10.18653/v1/2022.acl-long.556.
- [39] N. M. Guerreiro, D. M. Alves, J. Waldendorf, B. Haddow, A. Birch, P. Colombo, A. F. T. Martins, Hallucinations in large multilingual translation models, CoRR abs/2303.16104 (2023). URL: <https://doi.org/10.48550/arXiv.2303.16104>. doi:10.48550/arXiv.2303.16104. arXiv:2303.16104.
- [40] H. Liu, C. Li, Q. Wu, Y. J. Lee, Visual instruction tuning, CoRR abs/2304.08485 (2023). URL: <https://doi.org/10.48550/arXiv.2304.08485>. doi:10.48550/arXiv.2304.08485. arXiv:2304.08485.
- [41] Y. Li, Y. Du, K. Zhou, J. Wang, W. X. Zhao, J. Wen, Evaluating object hallucination in large vision-language models, CoRR abs/2305.10355 (2023). URL: <https://doi.org/10.48550/arXiv.2305.10355>. doi:10.48550/arXiv.2305.10355. arXiv:2305.10355.
- [42] S. Zheng, J. Huang, K. C. Chang, Why does chatgpt fall short in answering questions faithfully?, CoRR abs/2304.10513 (2023). URL: <https://doi.org/10.48550/arXiv.2304.10513>. doi:10.48550/arXiv.2304.10513. arXiv:2304.10513.
- [43] K. Huang, H. P. Chan, H. Ji, Zero-shot faithful factual error correction, in: ACL 2023, ACL, 2023, pp. 5660–5676. URL: <https://doi.org/10.18653/v1/2023.acl-long.311>. doi:10.18653/v1/

2023.acl-long.311.

- [44] L. Gao, Z. Dai, P. Pasupat, A. Chen, A. T. Chaganty, Y. Fan, V. Y. Zhao, N. Lao, H. Lee, D. Juan, K. Guu, RARR: researching and revising what language models say, using language models, in: ACL 2023, ACL, 2023, pp. 16477–16508. URL: <https://doi.org/10.18653/v1/2023.acl-long.910>. doi:10.18653/v1/2023.acl-long.910.
- [45] S. Zhang, L. Pan, J. Zhao, W. Y. Wang, Mitigating language model hallucination with interactive question-knowledge alignment, CoRR abs/2305.13669 (2023). URL: <https://doi.org/10.48550/arXiv.2305.13669>. doi:10.48550/arXiv.2305.13669. arXiv:2305.13669.
- [46] D. Halawi, J. Denain, J. Steinhardt, Overthinking the truth: Understanding how language models process false demonstrations, CoRR abs/2307.09476 (2023). URL: <https://doi.org/10.48550/arXiv.2307.09476>. arXiv:2307.09476.
- [47] D. Wan, S. Zhang, M. Bansal, Histalign: Improving context dependency in language generation by aligning with history, CoRR abs/2305.04782 (2023). URL: <https://doi.org/10.48550/arXiv.2305.04782>. doi:10.48550/arXiv.2305.04782. arXiv:2305.04782.
- [48] S. Chiesurin, D. Dimakopoulos, M. A. S. Cabezudo, A. Eshghi, I. Papaioannou, V. Rieser, I. Konstas, The dangers of trusting stochastic parrots: Faithfulness and trust in open-domain conversational question answering, in: A. Rogers, J. L. Boyd-Graber, N. Okazaki (Eds.), Findings of ACL 2023, ACL, 2023, pp. 947–959. URL: <https://doi.org/10.18653/v1/2023.findings-acl.60>. doi:10.18653/v1/2023.findings-acl.60.
- [49] D. Kang, T. Hashimoto, Improved natural language generation via loss truncation, in: D. Jurafsky, J. Chai, N. Schluter, J. R. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5–10, 2020, ACL, 2020, pp. 718–731. URL: <https://doi.org/10.18653/v1/2020.acl-main.66>. doi:10.18653/v1/2020.acl-main.66.
- [50] C. Wang, R. Sennrich, On exposure bias, hallucination and domain shift in neural machine translation, in: D. Jurafsky, J. Chai, N. Schluter, J. R. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5–10, 2020, ACL, 2020, pp. 3544–3552. URL: <https://doi.org/10.18653/v1/2020.acl-main.326>. doi:10.18653/v1/2020.acl-main.326.
- [51] N. Lee, W. Ping, P. Xu, M. Patwary, P. Fung, M. Shoeybi, B. Catanzaro, Factuality enhanced language models for open-ended text generation, in: NeurIPS, 2022. URL: http://papers.nips.cc/paper_files/paper/2022/hash/df438caa36714f69277daa92d608dd63-Abstract-Conference.html.
- [52] V. Raunak, A. Menezes, M. Junczys-Dowmunt, The curious case of hallucinations in neural machine translation, in: NAACL 2021, ACL, 2021, pp. 1172–1183.
- [53] N. M. Guerreiro, E. Voita, A. F. T. Martins, Looking for a needle in a haystack: A comprehensive study of hallucinations in neural machine translation, in: EACL 2023, ACL, 2023, pp. 1059–1075. URL: <https://aclanthology.org/2023.eacl-main.75>.
- [54] D. Dale, E. Voita, J. Lam, P. Hansanti, C. Ropers, E. Kalbassi, C. Gao, L. Barrault, M. R. Costa-jussà, Halomi: A manually annotated benchmark for multilingual hallucination and omission detection in machine translation, CoRR abs/2305.11746 (2023). URL: <https://doi.org/10.48550/arXiv.2305.11746>. doi:10.48550/arXiv.2305.11746. arXiv:2305.11746.
- [55] J. Pfeiffer, F. Piccinno, M. Nicosia, X. Wang, M. Reid, S. Ruder, mmt5: Modular multilingual pre-training solves source language hallucinations, CoRR abs/2305.14224 (2023). URL: <https://doi.org/10.48550/arXiv.2305.14224>. doi:10.48550/arXiv.2305.14224. arXiv:2305.14224.
- [56] V. Adlakha, P. BehnamGhader, X. H. Lu, N. Meade, S. Reddy, Evaluating correctness and faithfulness of instruction-following models for question answering, CoRR abs/2307.16877 (2023). URL: <https://doi.org/10.48550/arXiv.2307.16877>. doi:10.48550/arXiv.2307.16877. arXiv:2307.16877.
- [57] N. Dziri, S. Milton, M. Yu, O. R. Zaïane, S. Reddy, On the origin of hallucinations in conversational models: Is it the datasets or the models?, in: NAACL 2022, ACL, 2022, pp. 5271–5285. URL: <https://doi.org/10.18653/v1/2022.nacl-main.387>. doi:10.18653/v1/2022.nacl-main.387.
- [58] S. Das, S. Saha, R. K. Srihari, Diving deep into modes of fact hallucinations in dialogue systems, in: Findings of EMNLP 2022, ACL, 2022, pp. 684–699. URL: <https://doi.org/10.18653/v1/2022>.

- findings-emnlp.48. doi:10.18653/v1/2022.findings-emnlp.48.
- [59] N. Dziri, E. Kamalloo, S. Milton, O. R. Zaiane, M. Yu, E. M. Ponti, S. Reddy, Faithdial: A faithful benchmark for information-seeking dialogue, *Trans. Assoc. Comput. Linguistics* 10 (2022) 1473–1490. URL: <https://transacl.org/ojs/index.php/tacl/article/view/4113>.
 - [60] N. Dziri, H. Rashkin, T. Linzen, D. Reitter, Evaluating attribution in dialogue systems: The BEGIN benchmark, *Trans. Assoc. Comput. Linguistics* 10 (2022) 1066–1083. URL: <https://transacl.org/ojs/index.php/tacl/article/view/3977>.
 - [61] W. Sun, Z. Shi, S. Gao, P. Ren, M. de Rijke, Z. Ren, Contrastive learning reduces hallucination in conversations, in: AAAI 2023, AAAI Press, 2023, pp. 13618–13626. URL: <https://ojs.aaai.org/index.php/AAAI/article/view/26596>.
 - [62] D. Tam, A. Mascarenhas, S. Zhang, S. Kwan, M. Bansal, C. Raffel, Evaluating the factual consistency of large language models through news summarization, in: Findings of ACL 2023, ACL, 2023, pp. 5220–5255. URL: <https://doi.org/10.18653/v1/2023.findings-acl.322>. doi:10.18653/v1/2023.findings-acl.322.
 - [63] M. Cao, Y. Dong, J. C. K. Cheung, Hallucinated but factual! inspecting the factuality of hallucinations in abstractive summarization, in: ACL 2022, ACL, 2022, pp. 3340–3354. URL: <https://doi.org/10.18653/v1/2022.acl-long.236>. doi:10.18653/v1/2022.acl-long.236.
 - [64] J. Shen, J. Liu, D. Finnie, N. Rahmati, M. Bendersky, M. Najork, "why is this misleading?": Detecting news headline hallucinations with explanations, in: WWW 2023, ACM, 2023, pp. 1662–1672. URL: <https://doi.org/10.1145/3543507.3583375>. doi:10.1145/3543507.3583375.
 - [65] Y. Qiu, Y. Ziser, A. Korhonen, E. M. Ponti, S. B. Cohen, Detecting and mitigating hallucinations in multilingual summarisation, *CoRR* abs/2305.13632 (2023). URL: <https://doi.org/10.48550/arXiv.2305.13632>. arXiv: 2305.13632. doi:10.48550/arXiv.2305.13632.
 - [66] J. Yu, X. Wang, S. Tu, S. Cao, D. Zhang-li, X. Lv, H. Peng, Z. Yao, X. Zhang, H. Li, C. Li, Z. Zhang, Y. Bai, Y. Liu, A. Xin, N. Lin, K. Yun, L. Gong, J. Chen, Z. Wu, Y. Qi, W. Li, Y. Guan, K. Zeng, J. Qi, H. Jin, J. Liu, Y. Gu, Y. Yao, N. Ding, L. Hou, Z. Liu, B. Xu, J. Tang, J. Li, Kola: Carefully benchmarking world knowledge of large language models, *CoRR* abs/2306.09296 (2023). URL: <https://doi.org/10.48550/arXiv.2306.09296>. doi:10.48550/arXiv.2306.09296. arXiv: 2306.09296.
 - [67] F. Liu, K. Lin, L. Li, J. Wang, Y. Yacoob, L. Wang, Aligning large multi-modal model with robust instruction tuning, *CoRR* abs/2306.14565 (2023). URL: <https://doi.org/10.48550/arXiv.2306.14565>. arXiv: 2306.14565. doi:10.48550/arXiv.2306.14565.
 - [68] R. Mahmood, G. Wang, M. K. Kalra, P. Yan, Fact-checking of ai-generated reports, *CoRR* abs/2307.14634 (2023). URL: <https://doi.org/10.48550/arXiv.2307.14634>. doi:10.48550/arXiv.2307.14634. arXiv: 2307.14634.
 - [69] D. Lei, Y. Li, M. Hu, M. Wang, V. Yun, E. Ching, E. Kamal, Chain of natural language inference for reducing large language model ungrounded hallucinations (2023). arXiv: 2310.03951.
 - [70] W. Yuan, G. Neubig, P. Liu, Bartscore: Evaluating generated text as text generation, in: M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, J. W. Vaughan (Eds.), NeurIPS 2021, 2021, pp. 27263–27277. URL: <https://proceedings.neurips.cc/paper/2021/hash/e4d2b6e6fdeca3e60e0f1a62fee3d9dd-Abstract.html>.
 - [71] A. Amayuelas, L. Pan, W. Chen, W. Y. Wang, Knowledge of knowledge: Exploring known-unknowns uncertainty with large language models, *CoRR* abs/2305.13712 (2023). URL: <https://doi.org/10.48550/arXiv.2305.13712>. doi:10.48550/arXiv.2305.13712. arXiv: 2305.13712.
 - [72] P. Hase, M. T. Diab, A. Celikyilmaz, X. Li, Z. Kozareva, V. Stoyanov, M. Bansal, S. Iyer, Methods for measuring, updating, and visualizing factual beliefs in language models, in: A. Vlachos, I. Augenstein (Eds.), EACL 2023, ACL, 2023, pp. 2706–2723. URL: <https://aclanthology.org/2023.eacl-main.199>.
 - [73] P. Pezeshkpour, Measuring and modifying factual knowledge in large language models, *CoRR* abs/2306.06264 (2023). URL: <https://doi.org/10.48550/arXiv.2306.06264>. doi:10.48550/arXiv.2306.06264. arXiv: 2306.06264.
 - [74] T. Zhao, M. Wei, J. S. Preston, H. Poon, Llm calibration and automatic hallucination detection via

- pareto optimal self-supervision, 2023. arXiv:2306.16564.
- [75] N. Varshney, W. Yao, H. Zhang, J. Chen, D. Yu, A stitch in time saves nine: Detecting and mitigating hallucinations of llms by validating low-confidence generation, CoRR abs/2307.03987 (2023). URL: <https://doi.org/10.48550/arXiv.2307.03987>. doi:10.48550/arXiv.2307.03987. arXiv:2307.03987.
- [76] S. Kadavath, T. Conerly, A. Askell, T. Henighan, D. Drain, E. Perez, N. Schiefer, Z. Hatfield-Dodds, N. DasSarma, E. Tran-Johnson, S. Johnston, S. E. Showk, A. Jones, N. Elhage, T. Hume, A. Chen, Y. Bai, S. Bowman, S. Fort, D. Ganguli, D. Hernandez, J. Jacobson, J. Kernion, S. Kravec, L. Lovitt, K. Ndousse, C. Olsson, S. Ringer, D. Amodei, T. Brown, J. Clark, N. Joseph, B. Mann, S. McCandlish, C. Olah, J. Kaplan, Language models (mostly) know what they know, CoRR abs/2207.05221 (2022). URL: <https://doi.org/10.48550/arXiv.2207.05221>. doi:10.48550/arXiv.2207.05221. arXiv:2207.05221.
- [77] P. Manakul, A. Liusie, M. J. F. Gales, Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models, CoRR abs/2303.08896 (2023). URL: <https://doi.org/10.48550/arXiv.2303.08896>. doi:10.48550/arXiv.2303.08896. arXiv:2303.08896.
- [78] A. Agrawal, L. Mackey, A. T. Kalai, Do language models know when they're hallucinating references?, CoRR abs/2305.18248 (2023). URL: <https://doi.org/10.48550/arXiv.2305.18248>. doi:10.48550/arXiv.2305.18248. arXiv:2305.18248.
- [79] M. Li, B. Peng, Z. Zhang, Self-checker: Plug-and-play modules for fact-checking with large language models, CoRR abs/2305.14623 (2023). URL: <https://doi.org/10.48550/arXiv.2305.14623>. doi:10.48550/arXiv.2305.14623. arXiv:2305.14623.
- [80] R. Cohen, M. Hamri, M. Geva, A. Globerson, LM vs LM: detecting factual errors via cross examination, CoRR abs/2305.13281 (2023). URL: <https://doi.org/10.48550/arXiv.2305.13281>. doi:10.48550/arXiv.2305.13281. arXiv:2305.13281.
- [81] N. Mündler, J. He, S. Jenko, M. T. Vechev, Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation, CoRR abs/2305.15852 (2023). URL: <https://doi.org/10.48550/arXiv.2305.15852>. doi:10.48550/arXiv.2305.15852. arXiv:2305.15852.
- [82] S. Yang, R. Sun, X. Wan, A new benchmark and reverse validation method for passage-level hallucination detection (2023). arXiv:2310.06498.
- [83] S. Min, K. Krishna, X. Lyu, M. Lewis, W. Yih, P. W. Koh, M. Iyyer, L. Zettlemoyer, H. Hajishirzi, Factscore: Fine-grained atomic evaluation of factual precision in long form text generation, CoRR abs/2305.14251 (2023). URL: <https://doi.org/10.48550/arXiv.2305.14251>. doi:10.48550/arXiv.2305.14251. arXiv:2305.14251.
- [84] J. Chen, G. Kim, A. Sriram, G. Durrett, E. Choi, Complex claim verification with evidence retrieved in the wild, CoRR abs/2305.11859 (2023). URL: <https://doi.org/10.48550/arXiv.2305.11859>. doi:10.48550/ARXIV.2305.11859. arXiv:2305.11859.
- [85] S. Huo, N. Arabzadeh, C. L. A. Clarke, Retrieving supporting evidence for llms generated answers, CoRR abs/2306.13781 (2023). URL: <https://doi.org/10.48550/arXiv.2306.13781>. doi:10.48550/ARXIV.2306.13781. arXiv:2306.13781.
- [86] I. Chern, S. Chern, S. Chen, W. Yuan, K. Feng, C. Zhou, J. He, G. Neubig, P. Liu, Factool: Factuality detection in generative AI - A tool augmented framework for multi-task and multi-domain scenarios, CoRR abs/2307.13528 (2023). URL: <https://doi.org/10.48550/arXiv.2307.13528>. doi:10.48550/arXiv.2307.13528. arXiv:2307.13528.
- [87] R. Bawden, F. Yvon, Investigating the translation performance of a large multilingual language model: the case of BLOOM, CoRR abs/2303.01911 (2023). URL: <https://doi.org/10.48550/arXiv.2303.01911>. doi:10.48550/arXiv.2303.01911. arXiv:2303.01911.
- [88] A. Hendy, M. Abdelrehim, A. Sharaf, V. Raunak, M. Gabr, H. Matsushita, Y. J. Kim, M. Afify, H. H. Awadalla, How good are GPT models at machine translation? A comprehensive evaluation, CoRR abs/2302.09210 (2023). URL: <https://doi.org/10.48550/arXiv.2302.09210>. doi:10.48550/arXiv.2302.09210. arXiv:2302.09210.
- [89] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, in:

- D. Jurafsky, J. Chai, N. Schluter, J. R. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, ACL, 2020, pp. 8440–8451. URL: <https://doi.org/10.18653/v1/2020.acl-main.747>. doi:10.18653/v1/2020.acl-main.747.
- [90] S. Yuan, M. Färber, Evaluating generative models for graph-to-text generation, CoRR abs/2307.14712 (2023). URL: <https://doi.org/10.48550/arXiv.2307.14712>. arXiv:2307.14712.
- [91] Y. Zhu, X. Wang, J. Chen, S. Qiao, Y. Ou, Y. Yao, S. Deng, H. Chen, N. Zhang, Llms for knowledge graph construction and reasoning: Recent capabilities and future opportunities, CoRR abs/2305.13168 (2023). URL: <https://doi.org/10.48550/arXiv.2305.13168>. doi:10.48550/arXiv.2305.13168. arXiv:2305.13168.
- [92] D. Zhu, J. Chen, X. Shen, X. Li, M. Elhoseiny, Minigpt-4: Enhancing vision-language understanding with advanced large language models, CoRR abs/2304.10592 (2023). URL: <https://doi.org/10.48550/arXiv.2304.10592>. arXiv:2304.10592.
- [93] P. Wang, A. Yang, R. Men, J. Lin, S. Bai, Z. Li, J. Ma, C. Zhou, J. Zhou, H. Yang, OFA: unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework, in: K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvári, G. Niu, S. Sabato (Eds.), ICML 2022, volume 162 of *Proceedings of Machine Learning Research*, PMLR, 2022, pp. 23318–23340. URL: <https://proceedings.mlr.press/v162/wang22al.html>.
- [94] A. F. Biten, L. Gómez, D. Karatzas, Let there be a clock on the beach: Reducing object hallucination in image captioning, in: IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022, Waikoloa, HI, USA, January 3-8, 2022, IEEE, 2022, pp. 2473–2482. URL: <https://doi.org/10.1109/WACV51458.2022.00253>. doi:10.1109/WACV51458.2022.00253.
- [95] S. Petryk, S. Whitehead, J. E. Gonzalez, T. Darrell, A. Rohrbach, M. Rohrbach, Simple token-level confidence improves caption correctness, CoRR abs/2305.07021 (2023). URL: <https://doi.org/10.48550/arXiv.2305.07021>. arXiv:2305.07021.
- [96] M. Ning, Y. Xie, D. Chen, Z. Song, L. Yuan, Y. Tian, Q. Ye, L. Yuan, Album storytelling with iterative story-aware captioning and large language models, CoRR abs/2305.12943 (2023). URL: <https://doi.org/10.48550/arXiv.2305.12943>. doi:10.48550/arXiv.2305.12943. arXiv:2305.12943.
- [97] A. Pagnoni, V. Balachandran, Y. Tsvetkov, Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics, in: NAACL 2021, ACL, 2021, pp. 4812–4829. URL: <https://doi.org/10.18653/v1/2021.nacl-main.383>. doi:10.18653/v1/2021.nacl-main.383.
- [98] Y. Li, B. Peng, Y. Shen, Y. Mao, L. Liden, Z. Yu, J. Gao, Knowledge-grounded dialogue generation with a unified knowledge representation, in: NAACL 2022, ACL, 2022, pp. 206–218. URL: <https://doi.org/10.18653/v1/2022.nacl-main.15>. doi:10.18653/v1/2022.nacl-main.15.
- [99] Y. Dong, J. Wieting, P. Verga, Faithful to the document or to the world? mitigating hallucinations via entity-linked knowledge in abstractive summarization, in: Findings of EMNLP 2022, ACL, 2022, pp. 1067–1082. URL: <https://doi.org/10.18653/v1/2022.findings-emnlp.76>. doi:10.18653/v1/2022.findings-emnlp.76.
- [100] E. Durmus, H. He, M. T. Diab, FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization, in: D. Jurafsky, J. Chai, N. Schluter, J. R. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, ACL, 2020, pp. 5055–5070. URL: <https://doi.org/10.18653/v1/2020.acl-main.454>. doi:10.18653/v1/2020.acl-main.454.
- [101] J. Pei, D. Jurgens, Measuring sentence-level and aspect-level (un)certainty in science communications, in: M. Moens, X. Huang, L. Specia, S. W. Yih (Eds.), EMNLP 2021, ACL, 2021, pp. 9959–10011. URL: <https://doi.org/10.18653/v1/2021.emnlp-main.784>. doi:10.18653/v1/2021.emnlp-main.784.
- [102] G. Ilharco, M. T. Ribeiro, M. Wortsman, L. Schmidt, H. Hajishirzi, A. Farhadi, Editing models with task arithmetic, in: The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023, OpenReview.net, 2023. URL: <https://openreview.net/pdf?id=6t0Kwf8-jrj>.

- [103] N. Daheim, N. Dziri, M. Sachan, I. Gurevych, E. M. Ponti, Elastic weight removal for faithful and abstractive dialogue generation, CoRR abs/2303.17574 (2023). URL: <https://doi.org/10.48550/arXiv.2303.17574>. arXiv: 2303.17574.
- [104] A. Chen, P. Pasupat, S. Singh, H. Lee, K. Guu, PURR: efficiently editing language model hallucinations by denoising language model corruptions, CoRR abs/2305.14908 (2023). URL: <https://doi.org/10.48550/arXiv.2305.14908>. doi:10.48550/arXiv.2305.14908. arXiv: 2305.14908.
- [105] W. Shi, X. Han, M. Lewis, Y. Tsvetkov, L. Zettlemoyer, S. W. Yih, Trusting your evidence: Hallucinate less with context-aware decoding, CoRR abs/2305.14739 (2023). URL: <https://doi.org/10.48550/arXiv.2305.14739>. doi:10.48550/arXiv.2305.14739. arXiv: 2305.14739.
- [106] Z. Luo, C. Xu, P. Zhao, X. Geng, C. Tao, J. Ma, Q. Lin, D. Jiang, Augmented large language models with parametric knowledge guiding, CoRR abs/2305.04757 (2023). URL: <https://doi.org/10.48550/arXiv.2305.04757>. arXiv: 2305.04757.
- [107] S. Li, S. Park, I. Lee, O. Bastani, TRAC: trustworthy retrieval augmented chatbot, CoRR abs/2307.04642 (2023). URL: <https://doi.org/10.48550/arXiv.2307.04642>. doi:10.48550/arXiv.2307.04642. arXiv: 2307.04642.
- [108] K. Li, O. Patel, F. B. Viégas, H. Pfister, M. Wattenberg, Inference-time intervention: Eliciting truthful answers from a language model, CoRR abs/2306.03341 (2023). URL: <https://doi.org/10.48550/arXiv.2306.03341>. doi:10.48550/arXiv.2306.03341. arXiv: 2306.03341.
- [109] P. Wang, N. Zhang, X. Xie, Y. Yao, B. Tian, M. Wang, Z. Xi, S. Cheng, K. Liu, G. Zheng, H. Chen, Easyedit: An easy-to-use knowledge editing framework for large language models, CoRR abs/2308.07269 (2023). URL: <https://doi.org/10.48550/arXiv.2308.07269>. doi:10.48550/arXiv.2308.07269. arXiv: 2308.07269.
- [110] Y.-S. Chuang, Y. Xie, H. Luo, Y. Kim, J. Glass, P. He, Dola: Decoding by contrasting layers improves factuality in large language models, arXiv preprint arXiv:2309.03883 (2023).
- [111] N. Dziri, A. Madotto, O. Zaïane, A. J. Bose, Neural path hunter: Reducing hallucination in dialogue systems via path grounding, in: M. Moens, X. Huang, L. Specia, S. W. Yih (Eds.), EMNLP 2021, ACL, 2021, pp. 2197–2214. URL: <https://doi.org/10.18653/v1/2021.emnlp-main.168>. doi:10.18653/v1/2021.emnlp-main.168.
- [112] X. Han, Y. Tsvetkov, ORCA: interpreting prompted language models via locating supporting data evidence in the ocean of pretraining data, CoRR abs/2205.12600 (2022). URL: <https://doi.org/10.48550/arXiv.2205.12600>. arXiv: 2205.12600.
- [113] H. He, H. Zhang, D. Roth, Rethinking with retrieval: Faithful large language model inference, CoRR abs/2301.00303 (2023). URL: <https://doi.org/10.48550/arXiv.2301.00303>. doi:10.48550/arXiv.2301.00303. arXiv: 2301.00303.
- [114] S. M. Park, K. Georgiev, A. Ilyas, G. Leclerc, A. Madry, TRAK: attributing model behavior at scale, in: A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, J. Scarlett (Eds.), ICML 2023, volume 202 of *Proceedings of Machine Learning Research*, PMLR, 2023, pp. 27074–27113. URL: <https://proceedings.mlr.press/v202/park23c.html>.
- [115] M. Marone, B. V. Durme, Data portraits: Recording foundation model training data, CoRR abs/2303.03919 (2023). URL: <https://doi.org/10.48550/arXiv.2303.03919>. doi:10.48550/arXiv.2303.03919. arXiv: 2303.03919.
- [116] A. Madaan, N. Tandon, P. Gupta, S. Hallinan, L. Gao, S. Wiegreffe, U. Alon, N. Dziri, S. Prabhumoye, Y. Yang, S. Welleck, B. P. Majumder, S. Gupta, A. Yazdanbakhsh, P. Clark, Self-refine: Iterative refinement with self-feedback, CoRR abs/2303.17651 (2023). URL: <https://doi.org/10.48550/arXiv.2303.17651>. doi:10.48550/arXiv.2303.17651. arXiv: 2303.17651.
- [117] N. Shinn, B. Labash, A. Gopinath, Reflexion: an autonomous agent with dynamic memory and self-reflection, CoRR abs/2303.11366 (2023). URL: <https://doi.org/10.48550/arXiv.2303.11366>. doi:10.48550/arXiv.2303.11366. arXiv: 2303.11366.
- [118] O. Weller, M. Marone, N. Weir, D. J. Lawrie, D. Khashabi, B. V. Durme, "according to ..." prompting language models improves quoting from pre-training data, CoRR abs/2305.13252 (2023). URL: <https://doi.org/10.48550/arXiv.2305.13252>. doi:10.48550/arXiv.2305.13252. arXiv: 2305.13252.

- [119] R. Zhao, X. Li, S. Joty, C. Qin, L. Bing, Verify-and-edit: A knowledge-enhanced chain-of-thought framework, in: A. Rogers, J. L. Boyd-Graber, N. Okazaki (Eds.), ACL 2023, ACL, 2023, pp. 5823–5840. URL: <https://doi.org/10.18653/v1/2023.acl-long.320>. doi:10.18653/v1/2023.acl-long.320.
- [120] S. Dhuliawala, M. Komeili, J. Xu, R. Raileanu, X. Li, A. Celikyilmaz, J. Weston, Chain-of-verification reduces hallucination in large language models, CoRR abs/2309.11495 (2023). URL: <https://doi.org/10.48550/arXiv.2309.11495>. doi:10.48550/arXiv.2309.11495. arXiv:2309.11495.
- [121] S. Borgeaud, A. Mensch, J. Hoffmann, T. Cai, E. Rutherford, K. Millican, G. van den Driessche, J. Lespiau, B. Damoc, A. Clark, D. de Las Casas, A. Guy, J. Menick, R. Ring, T. Hennigan, S. Huang, L. Maggiore, C. Jones, A. Cassirer, A. Brock, M. Paganini, G. Irving, O. Vinyals, S. Osindero, K. Simonyan, J. W. Rae, E. Elsen, L. Sifre, Improving language models by retrieving from trillions of tokens, in: K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvári, G. Niu, S. Sabato (Eds.), ICML 2022, volume 162 of *Proceedings of Machine Learning Research*, PMLR, 2022, pp. 2206–2240. URL: <https://proceedings.mlr.press/v162/borgeaud22a.html>.
- [122] H. Trivedi, N. Balasubramanian, T. Khot, A. Sabharwal, Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions, in: A. Rogers, J. L. Boyd-Graber, N. Okazaki (Eds.), ACL 2023, ACL, 2023, pp. 10014–10037. URL: <https://doi.org/10.18653/v1/2023.acl-long.557>. doi:10.18653/v1/2023.acl-long.557.
- [123] B. Peng, M. Galley, P. He, H. Cheng, Y. Xie, Y. Hu, Q. Huang, L. Liden, Z. Yu, W. Chen, J. Gao, Check your facts and try again: Improving large language models with external knowledge and automated feedback, CoRR abs/2302.12813 (2023). URL: <https://doi.org/10.48550/arXiv.2302.12813>. doi:10.48550/arXiv.2302.12813. arXiv:2302.12813.
- [124] Q. Jin, Y. Yang, Q. Chen, Z. Lu, Genegpt: Augmenting large language models with domain tools for improved access to biomedical information, CoRR abs/2304.09667 (2023). URL: <https://doi.org/10.48550/arXiv.2304.09667>. doi:10.48550/arXiv.2304.09667. arXiv:2304.09667.
- [125] Z. Ding, A. Srinivasan, S. MacNeil, J. Chan, Fluid transformers and creative analogies: Exploring large language models' capacity for augmenting cross-domain analogical creativity, in: Creativity and Cognition, C&C 2023, Virtual Event, USA, June 19-21, 2023, ACM, 2023, pp. 489–505. URL: <https://doi.org/10.1145/3591196.3593516>. doi:10.1145/3591196.3593516.
- [126] X. Li, R. Zhao, Y. K. Chia, B. Ding, L. Bing, S. R. Joty, S. Poria, Chain of knowledge: A framework for grounding large language models with structured knowledge bases, CoRR abs/2305.13269 (2023). URL: <https://doi.org/10.48550/arXiv.2305.13269>. doi:10.48550/arXiv.2305.13269. arXiv:2305.13269.
- [127] Z. Jiang, F. F. Xu, L. Gao, Z. Sun, Q. Liu, J. Dwivedi-Yu, Y. Yang, J. Callan, G. Neubig, Active retrieval augmented generation, CoRR abs/2305.06983 (2023). URL: <https://doi.org/10.48550/arXiv.2305.06983>. doi:10.48550/arXiv.2305.06983. arXiv:2305.06983.
- [128] S. G. Patil, T. Zhang, X. Wang, J. E. Gonzalez, Gorilla: Large language model connected with massive apis, CoRR abs/2305.15334 (2023). URL: <https://doi.org/10.48550/arXiv.2305.15334>. doi:10.48550/arXiv.2305.15334. arXiv:2305.15334.
- [129] J. Liu, J. Jin, Z. Wang, J. Cheng, Z. Dou, J. Wen, RETA-LLM: A retrieval-augmented large language model toolkit, CoRR abs/2306.05212 (2023). URL: <https://doi.org/10.48550/arXiv.2306.05212>. doi:10.48550/arXiv.2306.05212. arXiv:2306.05212.
- [130] X. Wang, Q. Yang, Y. Qiu, J. Liang, Q. He, Z. Gu, Y. Xiao, W. Wang, Knowledgpt: Enhancing large language models with retrieval and storage access on knowledge bases, CoRR abs/2308.11761 (2023). URL: <https://doi.org/10.48550/arXiv.2308.11761>. doi:10.48550/arXiv.2308.11761. arXiv:2308.11761.
- [131] N. Stiennon, L. Ouyang, J. Wu, D. M. Ziegler, R. Lowe, C. Voss, A. Radford, D. Amodei, P. F. Christiano, Learning to summarize with human feedback, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), NeurIPS 2020, 2020. URL: <https://proceedings.neurips.cc/paper/2020/hash/1f89885d556929e98d3ef9b86448f951-Abstract.html>.
- [132] J. Menick, M. Trebacz, V. Mikulik, J. Aslanides, H. F. Song, M. J. Chadwick, M. Glaese, S. Young, L. Campbell-Gillingham, G. Irving, N. McAleese, Teaching language models to support answers

- with verified quotes, CoRR abs/2203.11147 (2022). URL: <https://doi.org/10.48550/arXiv.2203.11147>. doi:10.48550/arXiv.2203.11147. arXiv:2203.11147.
- [133] Y. Liu, P. Liu, D. R. Radev, G. Neubig, BRIO: bringing order to abstractive summarization, in: S. Muresan, P. Nakov, A. Villavicencio (Eds.), ACL 2022, ACL, 2022, pp. 2890–2903. URL: <https://doi.org/10.18653/v1/2022.acl-long.207>. doi:10.18653/v1/2022.acl-long.207.
 - [134] H. Liu, C. Sferrazza, P. Abbeel, Chain of hindsight aligns language models with feedback, CoRR abs/2302.02676 (2023). URL: <https://doi.org/10.48550/arXiv.2302.02676>. doi:10.48550/arXiv.2302.02676. arXiv:2302.02676.
 - [135] Z. Gou, Z. Shao, Y. Gong, Y. Shen, Y. Yang, N. Duan, W. Chen, CRITIC: large language models can self-correct with tool-interactive critiquing, CoRR abs/2305.11738 (2023). URL: <https://doi.org/10.48550/arXiv.2305.11738>. doi:10.48550/arXiv.2305.11738. arXiv:2305.11738.
 - [136] X. Zhu, B. Qi, K. Zhang, X. Long, B. Zhou, Pad: Program-aided distillation specializes large models in reasoning, CoRR abs/2305.13888 (2023). URL: <https://doi.org/10.48550/arXiv.2305.13888>. doi:10.48550/arXiv.2305.13888. arXiv:2305.13888.
 - [137] T. Gao, H. Yen, J. Yu, D. Chen, Enabling large language models to generate text with citations, CoRR abs/2305.14627 (2023). URL: <https://doi.org/10.48550/arXiv.2305.14627>. doi:10.48550/arXiv.2305.14627. arXiv:2305.14627.
 - [138] T. Dixit, F. Wang, M. Chen, Improving factuality of abstractive summarization without sacrificing summary quality, in: A. Rogers, J. L. Boyd-Graber, N. Okazaki (Eds.), ACL 2023, ACL, 2023, pp. 902–913. URL: <https://doi.org/10.18653/v1/2023.acl-short.78>. doi:10.18653/v1/2023.acl-short.78.
 - [139] Z. Ji, T. Yu, Y. Xu, N. Lee, E. Ishii, P. Fung, Towards mitigating hallucination in large language models via self-reflection (2023). arXiv:2310.06271.
 - [140] Y. Du, S. Li, A. Torralba, J. B. Tenenbaum, I. Mordatch, Improving factuality and reasoning in language models through multiagent debate, CoRR abs/2305.14325 (2023). URL: <https://doi.org/10.48550/arXiv.2305.14325>. doi:10.48550/arXiv.2305.14325. arXiv:2305.14325.
 - [141] T. Liang, Z. He, W. Jiao, X. Wang, Y. Wang, R. Wang, Y. Yang, Z. Tu, S. Shi, Encouraging divergent thinking in large language models through multi-agent debate, CoRR abs/2305.19118 (2023). URL: <https://doi.org/10.48550/arXiv.2305.19118>. doi:10.48550/arXiv.2305.19118. arXiv:2305.19118.
 - [142] K. Xiong, X. Ding, Y. Cao, T. Liu, B. Qin, Examining the inter-consistency of large language models: An in-depth analysis via debate, CoRR abs/2305.11595 (2023). URL: <https://doi.org/10.48550/arXiv.2305.11595>. doi:10.48550/arXiv.2305.11595. arXiv:2305.11595.
 - [143] R. Li, T. Patel, X. Du, PRD: peer rank and discussion improve large language model based evaluations, CoRR abs/2307.02762 (2023). URL: <https://doi.org/10.48550/arXiv.2307.02762>. doi:10.48550/arXiv.2307.02762. arXiv:2307.02762.
 - [144] Z. Wang, S. Mao, W. Wu, T. Ge, F. Wei, H. Ji, Unleashing cognitive synergy in large language models: A task-solving agent through multi-persona self-collaboration, CoRR abs/2307.05300 (2023). URL: <https://doi.org/10.48550/arXiv.2307.05300>. doi:10.48550/arXiv.2307.05300. arXiv:2307.05300.
 - [145] M. Minsky, Society of mind, Simon and Schuster, 1988.
 - [146] Y. Kirstain, P. S. H. Lewis, S. Riedel, O. Levy, A few more examples may be worth billions of parameters, in: Y. Goldberg, Z. Kozareva, Y. Zhang (Eds.), Findings of EMNLP 2022, ACL, 2022, pp. 1017–1029. URL: <https://doi.org/10.18653/v1/2022.findings-emnlp.72>. doi:10.18653/v1/2022.findings-emnlp.72.
 - [147] C. Zhou, P. Liu, P. Xu, S. Iyer, J. Sun, Y. Mao, X. Ma, A. Efrat, P. Yu, L. Yu, S. Zhang, G. Ghosh, M. Lewis, L. Zettlemoyer, O. Levy, LIMA: less is more for alignment, CoRR abs/2305.11206 (2023). URL: <https://doi.org/10.48550/arXiv.2305.11206>. doi:10.48550/arXiv.2305.11206. arXiv:2305.11206.
 - [148] S. Mishra, D. Khashabi, C. Baral, H. Hajishirzi, Natural instructions: Benchmarking generalization to new tasks from natural language instructions, CoRR abs/2104.08773 (2021). URL: <https://arxiv.org/abs/2104.08773>. arXiv:2104.08773.

- [149] V. Sanh, A. Webson, C. Raffel, S. H. Bach, L. Sutawika, Z. Alyafeai, A. Chaffin, A. Stiegler, A. Raja, M. Dey, M. S. Bari, C. Xu, U. Thakker, S. S. Sharma, E. Szczechla, T. Kim, G. Chhablani, N. V. Nayak, D. Datta, J. Chang, M. T. Jiang, H. Wang, M. Manica, S. Shen, Z. X. Yong, H. Pandey, R. Bawden, T. Wang, T. Neeraj, J. Rozen, A. Sharma, A. Santilli, T. Févry, J. A. Fries, R. Teehan, T. L. Scao, S. Biderman, L. Gao, T. Wolf, A. M. Rush, Multitask prompted training enables zero-shot task generalization, in: The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022, OpenReview.net, 2022. URL: <https://openreview.net/forum?id=9Vrb9D0WI4>.
- [150] Z. Bao, W. Chen, S. Xiao, K. Ren, J. Wu, C. Zhong, J. Peng, X. Huang, Z. Wei, Disc-medllm: Bridging general large language models and real-world medical consultation, 2023. arXiv: 2308.14346.
- [151] H. Gui, J. Zhang, H. Ye, N. Zhang, Instructie: A chinese instruction-based information extraction dataset, CoRR abs/2305.11527 (2023). URL: <https://doi.org/10.48550/arXiv.2305.11527>. doi:10.48550/arXiv.2305.11527. arXiv: 2305.11527.
- [152] W. Y. Wei Zhu, X. Wang, Shennong-tcm: A traditional chinese medicine large language model, <https://github.com/michael-wzhu/ShenNong-TCM-LLM>, 2023.
- [153] X. Li, P. Yu, C. Zhou, T. Schick, L. Zettlemoyer, O. Levy, J. Weston, M. Lewis, Self-alignment with instruction backtranslation, CoRR abs/2308.06259 (2023). URL: <https://doi.org/10.48550/arXiv.2308.06259>. arXiv: 2308.06259.
- [154] L. Chen, S. Li, J. Yan, H. Wang, K. Gunaratna, V. Yadav, Z. Tang, V. Srinivasan, T. Zhou, H. Huang, H. Jin, Alpagasus: Training A better alpaca with fewer data, CoRR abs/2307.08701 (2023). URL: <https://doi.org/10.48550/arXiv.2307.08701>. doi:10.48550/arXiv.2307.08701. arXiv: 2307.08701.
- [155] S. Yao, D. Yu, J. Zhao, I. Shafran, T. L. Griffiths, Y. Cao, K. Narasimhan, Tree of thoughts: Deliberate problem solving with large language models, CoRR abs/2305.10601 (2023). URL: <https://doi.org/10.48550/arXiv.2305.10601>. arXiv: 2305.10601.
- [156] Y. Zhang, J. Yang, Y. Yuan, A. C. Yao, Cumulative reasoning with large language models, CoRR abs/2308.04371 (2023). URL: <https://doi.org/10.48550/arXiv.2308.04371>. doi:10.48550/arXiv.2308.04371. arXiv: 2308.04371.
- [157] L. Gao, A. Madaan, S. Zhou, U. Alon, P. Liu, Y. Yang, J. Callan, G. Neubig, PAL: program-aided language models, in: A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, J. Scarlett (Eds.), ICML 2023, volume 202 of *Proceedings of Machine Learning Research*, PMLR, 2023, pp. 10764–10799. URL: <https://proceedings.mlr.press/v202/gao23f.html>.
- [158] W. Chen, X. Ma, X. Wang, W. W. Cohen, Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks, CoRR abs/2211.12588 (2022). URL: <https://doi.org/10.48550/arXiv.2211.12588>. arXiv: 2211.12588.
- [159] Z. Bi, N. Zhang, Y. Jiang, S. Deng, G. Zheng, H. Chen, When do program-of-thoughts work for reasoning?, 2023. arXiv: 2308.15452.
- [160] K. Frankish, Dual-process and dual-system theories of reasoning, *Philosophy Compass* 5 (2010) 914–926.
- [161] K. Stanovich, *Rationality and the reflective mind*, Oxford University Press, USA, 2011.
- [162] G. Piccinini, The first computational theory of mind and brain: a close look at mcculloch and pitts’s “logical calculus of ideas immanent in nervous activity”, *Synthese* 141 (2004) 175–215.
- [163] E. L. Thorndike, Animal intelligence, *Nature* 58 (1898) 390–390.
- [164] J. Li, D. Li, S. Savarese, S. C. H. Hoi, BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models, in: A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, J. Scarlett (Eds.), ICML 2023, volume 202 of *Proceedings of Machine Learning Research*, PMLR, 2023, pp. 19730–19742. URL: <https://proceedings.mlr.press/v202/li23q.html>.
- [165] W. Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. Li, P. Fung, S. C. H. Hoi, Instructblip: Towards general-purpose vision-language models with instruction tuning, CoRR abs/2305.06500 (2023). URL: <https://doi.org/10.48550/arXiv.2305.06500>. doi:10.48550/arXiv.2305.06500. arXiv: 2305.06500.
- [166] Q. Ye, H. Xu, G. Xu, J. Ye, M. Yan, Y. Zhou, J. Wang, A. Hu, P. Shi, Y. Shi, C. Li, Y. Xu, H. Chen,

- J. Tian, Q. Qi, J. Zhang, F. Huang, mplug-owl: Modularization empowers large language models with multimodality, CoRR abs/2304.14178 (2023). URL: <https://doi.org/10.48550/arXiv.2304.14178>. arXiv:2304.14178.
- [167] W. Shao, Y. Hu, P. Gao, M. Lei, K. Zhang, F. Meng, P. Xu, S. Huang, H. Li, Y. Qiao, P. Luo, Tiny lilm-ehub: Early multimodal experiments with bard, CoRR abs/2308.03729 (2023). URL: <https://doi.org/10.48550/arXiv.2308.03729>. doi:10.48550/arXiv.2308.03729. arXiv:2308.03729.
- [168] S. Yin, C. Fu, S. Zhao, K. Li, X. Sun, T. Xu, E. Chen, A survey on multimodal large language models, CoRR abs/2306.13549 (2023). URL: <https://doi.org/10.48550/arXiv.2306.13549>. doi:10.48550/arXiv.2306.13549. arXiv:2306.13549.