Through the Lens of Core Competency: Survey on Evaluation of Large Language Models

Ziyu Zhuang, Qiguang Chen, Longxuan Ma, Mingda Li, Yi Han, Yushan Qian, Haopeng Bai, Weinan Zhang*, Ting Liu
Research Center for Social Computing and Information Retrieval, Harbin Institute of Technology
{zyzhuang, qgchen, lxma, mdli, yihan, ysqian, hpbai, wnzhang, tliu}@ir.hit.edu.cn

Abstract

From pre-trained language model (PLM) to large language model (LLM), the field of natural language processing (NLP) has witnessed steep performance gains and wide practical uses. The evaluation of a research field guides its direction of improvement. However, LLMs are extremely hard to thoroughly evaluate for two reasons. First of all, traditional NLP tasks become inadequate due to the excellent performance of LLM. Secondly, existing evaluation tasks are difficult to keep up with the wide range of applications in real-world scenarios. To tackle these problems, existing works proposed various benchmarks to better evaluate LLMs. To clarify the numerous evaluation tasks in both academia and industry, we investigate multiple papers concerning LLM evaluations. We summarize 4 core competencies of LLM, including reasoning, knowledge, reliability, and safety. For every competency, we introduce its definition, corresponding benchmarks, and metrics. Under this competency architecture, similar tasks are combined to reflect corresponding ability, while new tasks can also be easily added into the system. Finally, we give our suggestions on the future direction of LLM's evaluation.

1 Introduction

Large language models(LLMs) have achieved great progresses in many areas. One representative, ChatGPT[0], which applies the ability of LLMs in the form of dialogue, has received much attention due to its incredible versatility such as creative writing, coding, planning, etc. The evaluation of such a model thus becomes necessary to benchmark and build up its ability while preventing potential harmfulness.

Existing works on the evaluation of LLMs can be divided into three paradigms. The first line of work is evaluating LLMs with traditional NLP tasks like dialogue, summarization, etc. Since LLMs are actually pre-trained language models(PLMs) with huge model parameter size and data size (Kaplan et al., 2020), benchmarks like GLUE (Wang et al., 2019b), SuperGLUE (Wang et al., 2019a) can be adopted to evaluate its language understanding ability. The problem is that LLMs work really well on less restrictive tasks like translation, summarization, and natural language understanding tasks. Sometimes LLMs generated outputs' third-party scores are even higher than human generations (Liang et al., 2022), showing the need for higher-quality tasks. Secondly, advanced ability evaluations are proposed to completely test language models. The parameter size difference between LLMs and PLMs brings an amazing phenomenon, emergence (Wei et al., 2022a; Srivastava et al., 2022), which means that scaled models exhibit abilities that are not possessed in small-scaled language models. For instance, in tasks like reasoning, and tool manipulation, the correlation curve between the number of model parameters and the task effect is non-linear. And the effect will rise sharply when the model parameter exceeds a certain parameter scale. They're called "advanced" because they're more closely related to human abilities and harder for models to complete (Zhong et al., 2023). Thirdly, test language models' intrinsic abilities independent of the specific tasks. It can be tested in parallel with almost every task above. Robustness is a classic ability

*Corresponding author

[0]https://openai.com/blog/chatgpt/

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 88-109, Harbin, China, August 3 - 5, 2023.
Volume 2: Frotier Forum
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

in this paradigm. Due to the black-box nature of neural networks (Szegedy et al., 2014), robustness problems exist for every modality of input data(vision, audio, test, etc.).

Current evaluation benchmarks (Liang et al., 2022; Srivastava et al., 2022; Gao et al., 2021; Zhong et al., 2023; Li et al., 2023a) are mostly a mixture of the former three paradigms. They emphasize a complete system of evaluation tasks, in which all tasks are of equal importance. But the significance of marginal increases in model effects on tasks with excellent performance is debatable. Thus numerous evaluation tasks and benchmarks are proposed to follow and challenge the ever-evolving LLMs, while, oddly, seldom being reviewed in a systematic way. How to link numerous tasks and benchmarks, better present the evaluation results, and thus facilitate the research of LLMs is an urgent problem.

An ideal large language model needs to be capable, reliable, and safe (Ouyang et al., 2022). One surely needs extensive tests on multiple datasets to meet these miscellaneous standards. Moreover, to avoid the prevalent training set leakage, test sets also should be updated regularly (Huang et al., 2023). This is similar to the competency (Hoffmann, 1999) tests adopted in corporate recruitment. In competency tests, different task sets are combined to test the corresponding competency. And task sets also need renewal to prevent possible fraud.

In this survey, **we draw on the concept of the core competency to integrate multiple evaluation research for LLMs.** We investigated **540+** tasks widely used in various papers, aggregating tasks corresponding to a certain competency. During this process, 4 core competencies are summarized, including knowledge, reasoning, reliability, and safety. We will introduce the definition, taxonomy, and metrics for these competencies. Through this competency test, superabundant evaluation tasks and benchmarks are combed and clarified for their aiming utility. Furthermore, the evaluation results presented with this procedure will be direct, concise, and focused. Updated new tasks can also be added comprehensively. To support the community in taking this competency test further, We also create an extensible project, which will show the many-to-many relationship between competencies and tasks precisely. Due to the length of the paper, we can only present part of the surveyed results in this paper. A more comprehensive study will be released in a later version.

2 Core Competencies

In this section, we introduce the definition and taxonomy of the core competencies we summarized.

2.1 Knowledge

Knowledge is generally defined as the cognition of humans when practicing in the subjective and objective world, which is verified and can be reused over time[1]. The large language models (LLMs) nowadays obtain human knowledge from a large scale of training corpus, so that it can use the knowledge to solve various downstream tasks. In this section, we focus on the fundamental knowledge competency of LLMs that facilitates communication and other downstream tasks (such as reasoning). Specifically, we divide the fundamental knowledge into **linguistic knowledge** and **world knowledge** (Day et al., 1998) and introduce the definitions of them and the benchmarks that can evaluate them.

2.1.1 Linguistic Knowledge Competency

Linguistic knowledge includes grammatical, semantic, and pragmatic knowledge (Fromkin et al., 2018). The grammar of a natural language is its set of structural constraints on speakers' or writers' composition of clauses, phrases, and words. The term can also refer to the study of such constraints, a field that includes domains such as phonology, morphology, and syntax, often complemented by phonetics, semantics, and pragmatics. Semantic (Austin, 1975) studies the meaning of words, phrases, and sentences, focusing on general meanings rather than on what an individual speaker may want them to mean. Pragmatics (Austin, 1975) studies language use and how listeners bridge the gap between sentence meaning and the speaker's meaning. It is concerned with the relationship between semantic meaning, the context of use, and the speaker's meaning.

[1]https://plato.stanford.edu/entries/epistemology/

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 88-109, Harbin, China, August 3 - 5, 2023.
Volume 2: Frotier Forum

Dataset	Knowledge Category	LLM evaluated	Task Format	Lang
BLiMP	grammatical	MT-NLG;BLOOM	Classification	En
linguistic_mappings	grammar/syntax	Gopher;Chinchilla;FLAN-T5;GLM;etc.	Generation	En
minute_mysteries_qa	semantic	Gopher;Chinchilla;FLAN-T5;GLM;etc.	Generation/QA	En
metaphor_boolean	pragmatic/semantic	Gopher;Chinchilla;FLAN-T5;GLM;etc.	Classification	En
LexGLUE	domain	BLOOM	Multiple choice	En
WikiFact	world	BLOOM	Generation	En
TruthfulQA	world	GPT-3/InstructGPT/GPT-4	Generation	En
HellaSwag	commonsense	GPT-3/InstructGPT/GPT-4	Generation	En

Table 1: Datasets that are used to evaluate the knowledge Competency of LLMs.

The Linguistic Knowledge competency is embodied in almost all NLP tasks, researchers usually design specific scenarios to test the linguistic competency of LLMs. Some examples are shown in the upper group of Table 1. BLiMP (Warstadt et al., 2020) evaluates what language models (LMs) know about major grammatical phenomena. Linguistic_mappings [2] task aims to explore the depth of linguistic knowledge in enormous language models trained on word prediction. It aims to discover whether such knowledge is structured so as to support the use of grammatical abstractions, both morphological (past tense formation and pluralization) and syntactic (question formation, negation, and pronominalization). The minute_mysteries_qa [3] is a reading comprehension task focusing on short crime and mystery stories where the goal is to identify the perpetrator and to explain the reasoning behind the deduction and the clues that support it. The metaphor_boolean [4] task presents a model with a metaphoric sentence and asks it to identify whether a second sentence is the correct interpretation of the first. The last three are selected from BIG-Bench (Srivastava et al., 2022), containing diverse task topics including linguistics.

2.1.2 World Knowledge Competency

World knowledge is non-linguistic information that helps a reader or listener interpret the meanings of words and sentences (Ovchinnikova, 2012). It is also referred to as extra-linguistic knowledge. In this paper, we categorize world knowledge into general knowledge and domain knowledge. The general knowledge includes commonsense knowledge (Davis, 2014) and prevalent knowledge. The commonsense knowledge consists of world facts, such as "Lemons are sour", or "Cows say moo", that most humans are expected to know. The prevalent knowledge exists at a particular time or place. For example, "Chinese people are used to drinking boiled water." is only known by a part of human beings; "There were eight planets in the solar system" is prevalent knowledge until it is overthrown. The domain knowledge (Alexander, 1992) is of a specific, specialized discipline or field, in contrast to general or domain-independent knowledge. People who have domain knowledge, are often considered specialists or experts in the field.

The bottom group of Table 1 shows some task examples that are used for testing world knowledge. For example, the LexGLUE (Chalkidis et al., 2022) tests whether LLMs perform well in the legal domain; WikiFact (Yasunaga et al., 2022) is a fact completion scenario that tests language models' factual knowledge based on Wikipedia. The input will be a partial sentence such as "The capital of France is _", and the output will be the continuation of the sentence such as "Paris"; TruthfulQA (Lin et al., 2022b) comprises questions spanning numerous categories including economics, science, and law. The questions are strategically chosen so humans may also incorrectly answer them based on misconceptions and biases; language models should ideally return accurate and truthful responses; HellaSwag (Zellers et al., 2019) tests commonsense inference and was created through adversarial filtering to synthesize wrong answers. The World knowledge competency, along with linguistic knowledge, serves as the foundation for solving different NLP tasks and is one of the core competencies of LLMs.

2.2 Reasoning

Reasoning competency is a crucial skill for LLMs to solve complex problems. What's more, from the perspective of intelligent agents, reasoning ability is also one of the core capabilities towards achieving

[2] https://github.com/google/BIG-bench/blob/main/bigbench/benchmark_tasks/linguistic_mappings
[3] https://github.com/google/BIG-bench/blob/main/bigbench/benchmark_tasks/minute_mysteries_qa
[4] https://github.com/google/BIG-bench/tree/main/bigbench/benchmark_tasks/metaphor_boolean

Dataset	Reasoning Competency	LLM evaluated	Task Format	Lang
COPA	Causal/Commonsense*	UL2;Deberta;GLaM;GPT3;PaLM;etc.	Classification	En
Mathematical Induction	Induction/Mathematical*	Gopher;Chinchilla;FLAN-T5;GLM;etc.	Generation	En
Synthetic Reasoning	Abduction/Deduction	HELM	Multiple choice	En
SAT Analogy	Analogical	GPT-3	Multiple choice	En
StrategyQA	Multi-hop/Commonsense*	Gopher;Chinchilla;FLAN-T5;GLM;etc.	Classification	En
GSM8K	Mathematical*	BLOOM;LLaMA;GPT-4;MT-NLG	Generation	En
ToTTo	Structured Data*	UL2	Generation	En

Table 2: Datasets that are used to evaluate the reasoning competency of LLMs. * represents a specific reasoning scenario.

AGI (Bubeck et al., 2023; Qiao et al., 2022). However, there remains no consensus whether LLMs can really reason, or just simply produce a larger context that increases the likelihood of correctly predicting the missing tokens (Mialon et al., 2023). Although ”reasoning” itself may currently be an excuse of language, we can still objectively verify the reasoning performance of LLMs through various reasoning competencies. Previous methods mainly focus on the division of reasoning tasks. Yu et al. (2023) divides existing evaluation tasks into three major categories, namely knowledge reasoning, symbolic reasoning, and mathematical reasoning, based on the type of logic and evidence involved in the reasoning process. Zhao et al. (2023) divides reasoning tasks into deductive reasoning and defeasible reasoning according to the reasoning form. In this section, we decompose the reasoning competency into 6 sub-parts from the perspective of model competency, providing a comprehensive overview of existing research efforts and suggesting potential future directions. And Table 2 presents some datasets for evaluating LLM's reasoning competency using this categorization approach.

2.2.1 Causal Reasoning Competency

Causal reasoning competency is a highly significant cognitive ability aimed at inferring causality through the observation of cause-effect relationships (Vowels et al., 2023; Dündar-Coecke, 2022; Chan et al., 2023). It enables us to comprehend and explain the relationships between events, variables, and actions, ultimately empowering us to make informed predictions and decisions (Gao et al., 2023).

The benchmarks Causal-TimeBank (Mirza et al., 2014), StoryLine (Caselli and Vossen, 2017), and MAVEN-ERE (Wang et al., 2022c) aim to test the existence of causal relationships between two events in sentences. COPA (Gordon et al., 2012) and XCOPA (Ponti et al., 2020) are evaluation benchmarks for extracting causal relationships in sentences, consisting of a set of premises and possible causes or effects. Tested systems are required to apply commonsense knowledge to identify the correct answers. e-CARE (Du et al., 2022) and CALM-Bench (Dalal et al., 2023) introduce a set of causal querying tasks to evaluate models, which include a cause and several potential effect sentences. Additionally, an annotated and interpretable causal reasoning dataset is provided for these tasks.

2.2.2 Deduction Reasoning Competency

In the era of Large Language Models (LLMs), deductive reasoning abilities serve as the foundational skills for logical reasoning (Evans, 2002). Unlike traditional rule-based deductive reasoning systems, it involves deriving specific conclusions or answers from general and universally applicable premises using given rules and logic. Specifically, it manifests as a process of Zero-Shot Chain-of-Thought utilizing given rules (Lyu et al., 2023; Kojima et al., 2022). For instance, (Kojima et al., 2022) introduced the “Let's think step by step” prompt technique to better evaluate the Deduction Reasoning Competency.

Current testing of this ability often intertwines with other skills and still lacks an independent evaluation on typical text (Clark et al., 2020) and symbol-related (Wu et al., 2021) deductive datasets. However, in general, almost all QA tasks can be explicitly evaluated for Deduction Reasoning using the Chain-of-Thought (CoT) approach. Therefore, the effectiveness of models' Deduction Reasoning Competency can be to some extent reflected by evaluating the performance of QA tasks after applying the CoT method.

2.2.3 Induction Reasoning Competency

In contrast to deductive reasoning, inductive reasoning aims to derive conclusions from specific observations to general principles (Yang et al., 2022; Olsson et al., 2022). In recent years, a new paradigm

of Induction Reasoning has been proposed by (Cheng et al., 2023), which requires models to generate general-purpose program code to solve a class of problems based on given contextual questions and a specific question. For example, Cheng et al. (2023), Jiang et al. (2023) and Surís et al. (2023) induced general principle-based solutions by generalizing each question into a universal executable language.

Therefore, for competency evaluation, while DEER (Yang et al., 2022) and Mathematical Induction (BIGBench Split (Srivastava et al., 2022)) took the first step in inductive reasoning, we still hope to establish a more systematic and comprehensive benchmark for evaluating this capability. Recently, Bills et al. (2023) has tested the inductive ability of GPT-4 (OpenAI, 2023) to evaluate its effectiveness in inducing patterns that are difficult for humans to express clearly. Intriguingly, Mankowitz et al. (2023) used some techniques to evaluate the extent to which LLM can mine previously unknown patterns.

2.2.4 Abduction Reasoning Competency

Abduction Reasoning Competency encompasses the task of providing explanations for the output generated based on given inputs (Kakas and Michael, 2020). This form of reasoning is particularly critical in scenarios where uncertainty or incomplete information exists, enabling systems to generate hypotheses and make informed decisions based on the available evidence. Notably, the research conducted by LIREx (Zhao and Vydiswaran, 2021) and STaR (Zelikman et al., 2022) delved into the Abduction Reasoning Competency of models and demonstrated the effectiveness of rationales provided during the Abduction Reasoning process in facilitating improved learning in downstream models.

In terms of datasets within the LLM setting, the benchmarks HUMMINGBIRD (Mathew et al., 2021) and HateXplain (Hayati et al., 2021) require models to output word-level textual segments as explanations for sentiment classification results. On the other hand, benchmarks such as WikiQA (Yang et al., 2015), HotpotQA (Yang et al., 2018), and SciFact (Wadden et al., 2020) provide sentence-level coarse-grained textual segments as explanations for model classification results. ERASER (DeYoung et al., 2020) and FineIEB (Wang et al., 2022b) provide benchmarks for evaluating Abduction Reasoning with diverse granularity explanations. Based on previous research, Synthetic Reasoning (Liang et al., 2022) provides a comprehensive evaluation of both Deduction Reasoning and Abduction Reasoning Competency. Moreover, Hessel et al. (2022) introduced the first comprehensive multimodal benchmark for testing Abduction Reasoning capabilities, providing a solid foundation for future advancements in this domain. Recently, Bills et al. (2023) evaluate GPT-4 by observing the activation of neurons in GPT-2 and offering explanations for the GPT-2's outputs. This research avenue also presents a novel approach for exploring the future evaluation of Abduction Reasoning Competency.

2.2.5 Analogical Reasoning Competency

Analogy reasoning competency encompasses the ability of reasoning by identifying and applying similarities between diverse situations or domains. It is based on the assumption that similar cases or objects tend to exhibit common attributes or behaviors. By recognizing these similarities, analogy reasoning enables systems to transfer knowledge or experience from one context to another (Sinha et al., 2019; Wei et al., 2022b). This type of reasoning plays a vital role in problem-solving, decision-making, and learning from past experiences. A typical example is In-Context-Learning (Dong et al., 2023), where the model is required to perform analogical reasoning based on given contexts, which are evaluated based on the final analogical results.

For a better assessment and understanding of the model's analogical reasoning ability, Brown et al. (2020) introduces SAT Analogies as a test to evaluate LLM's analogical reasoning capabilities. In recent years, Authorship Verification and ARC datasets (Srivastava et al., 2022) have also proposed evaluation benchmark that involve presenting contextual examples and requiring the model to produce induced pattern-compliant results. However, it should be noted that In-Context Learning (ICL) can be utilized for almost all tasks, enabling the evaluation of models' Analogical Reasoning Competency to some extent through the assessment of their performance after undergoing ICL.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 88-109, Harbin, China, August 3 - 5, 2023.
Volume 2: Frotier Forum

2.2.6 Multi-hop Reasoning Competency

Multi-hop reasoning refers to the ability to combine and integrate information from multiple sources or contexts to arrive at logical conclusions. This competency of reasoning enables systems to retrieve coherent and comprehensive answers by traversing multiple pieces of information, thus performing complex tasks of information retrieval, comprehension, and reasoning (Wang et al., 2022a; Qiu et al., 2019).

Currently, HotpotQA (Yang et al., 2018) serves as a commonly used dataset for multi-hop question answering tasks. Expanding on this, Ye and Durrett (2022) introduced a new and demanding subset that aimed to achieve a balance between accurate and inaccurate predictions using their model. Similarly, StrategyQA (Geva et al., 2021) is another widely used benchmark for multi-hop question answering (Wei et al., 2022b), where the required reasoning steps are implicit in the questions and should be inferred using strategies.

2.2.7 Reasoning in Scenarios

Commonsense Reasoning Commonsense reasoning is crucial for machines to achieve human-like understanding and interaction with the world in the field of machine intelligence (Storks et al., 2019; Bhargava and Ng, 2022). The ability to comprehend and apply commonsense knowledge enables machines to make accurate predictions, engage in logical reasoning, and navigate complex social situations.

OpenBookQA (Mihaylov et al., 2018) provides a foundational test for evaluating Commonsense Reasoning abilities in the form of an open-book exam. Building upon this, CommonsenseQA (Talmor et al., 2019) requires models to employ rich world knowledge for reasoning tasks. PIQA (Bisk et al., 2020) introduces a dataset for testing models' understanding of physical world commonsense reasoning. StrategyQA (Geva et al., 2021) presents a complex benchmark that requires commonsense-based multi-step/multi-hop reasoning, enabling a better exploration of the upper limits of models' Commonsense Reasoning Competency. Currently, due to early research on LLM (Wei et al., 2022b), CommonsenseQA (Talmor et al., 2019) remains the most widely used benchmark for commonsense reasoning.

Mathematical Reasoning Mathematical reasoning competency is crucial for general intelligent systems. It empowers intelligent systems with the capability of logical reasoning, problem-solving, and data manipulation and analysis, thereby facilitating the development and application of intelligent systems (Qiao et al., 2022; Mishra et al., 2022b; Mishra et al., 2022a).

Early evluation studies focused on small datasets of elementary-level mathematical word problems (MWPs) (Hosseini et al., 2014), but subsequent research aimed to increase complexity and scale (Srivastava et al., 2022; Brown et al., 2020). Furthermore, recent benchmarks (Mishra et al., 2022b; Mishra et al., 2022a) have provided comprehensive evaluation platforms and benchmarks for mathematical reasoning abilities. GSM8K (Cobbe et al., 2021) aims to evaluate elementary school MWPs. Currently, due to early research efforts on LLMs (Wei et al., 2022b), it remains the most widely used benchmark for mathematical reasoning in the LLM evaluation. Moreover, There have been recent advancements in evaluation research that explore mathematical reasoning competency integrating external knowledge, leveraging language diversity for multilingual evaluation (Shi et al., 2023), and testing mathematical reasoning on multi-modal setting (Lindström and Abraham, 2022), aiming to judge the broader data reasoning capabilities of large language models (LLMs).

Structured Data Reasoning Structured data reasoning involves the ability to reason and derive insights and answers from structured data sources, such as structured tabular data (Qiao et al., 2022; Li et al., 2023b; Xie et al., 2022).

WikiSQL (Zhong et al., 2017) and WikiTQ (Pasupat and Liang, 2015) provide tables as input and answer questions based on the additional input of questions. HybridQA (Chen et al., 2020b) and MultiModalQA (Talmor et al., 2021) propose benchmarks for hybrid Structure Reasoning by combining structured table inputs with text (and even other modalities). Similarly, MultiWoZ (Budzianowski et al., 2018), KVRET (Eric et al., 2017) and SQA (Iyyer et al., 2017) integrate table data into task-oriented dialogue systems to generate more complex structures and output dialog-related classifications. Unlike traditional QA, FeTaQA (Nan et al., 2021) requires free-form answers instead of extracting answer spans from passages. ToTTo (Parikh et al., 2020) introduces an open-domain English table-to-text dataset for

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 88-109, Harbin, China, August 3 - 5, 2023.
Volume 2: Frotier Forum

Structured Data Reasoning. Additionally, benchmarks such as TabFact (Chen et al., 2020a) and FEVEROUS (Aly et al., 2021) evaluate whether model statements are consistent with facts mentioned in structured data. In recent years, with a deeper focus on testing models' mathematical abilities, TabMWP (Lu et al., 2023) introduces a grade-level dataset of table-based mathematical word problems that require mathematical reasoning using both text and table data.

2.3 Reliability

Reliability measures to what extent a human can trust the contents generated by a LLM. It is of vital importance for the deployment and usability of the LLM, and attracts tons of concerns along with the rapid and astonishing development of recent LLMs (Weidinger et al., 2021; Wang et al., 2022d; Ji et al., 2023; Zhuo et al., 2023). Lots of concepts are closely related to reliability under the context of LLM, including but not limited to hallucination, truthfulness, factuality, honesty, calibration, robustness, interpretability (Lee et al., 2018; Belinkov et al., 2020; Evans et al., 2021; Mielke et al., 2022; Lin et al., 2022b). Reliability also overlaps with the safety and generalization of a LLM (Weidinger et al., 2021). In this section, we will give an overview of two most concerned directions: Hallucination, Uncertainty and Calibration.

2.3.1 Hallucination

Hallucination is a term often used to describe LLM's falsehoods, which is the opposite side of truthfulness or factuality (Ji et al., 2023; OpenAI, 2023; Bubeck et al., 2023). Hallucination is always categorized into intrinsic (close domain) hallucination and extrinsic (open domain) hallucination (Ji et al., 2023; OpenAI, 2023). Intrinsic hallucination refers to the unfaithfulness of the model output to a given context, while extrinsic hallucination refers to the untruthful contents about the world generated by the model without reference to a given source.

Early research on hallucination mainly focused on the intrinsic hallucination and lots of interesting metrics were proposed to evaluate the intrinsic hallucination level of a PTM (Ji et al., 2023). However, Bang et al. (2023) claimed that intrinsic hallucination was barely found after conducting a comprehensive analysis of ChatGPT's responses. Hence for LLM, the extrinsic hallucination is of the greatest concern. To evaluate the extrinsic hallucination potential of a LLM, a common practice is to leverage knowledge-intensive tasks such as Factual Question Answering (Joshi et al., 2017; Zheng et al., 2023) or Knowledge-grounded Dialogue (Dinan et al., 2019b; Das et al., 2022). TruthfulQA (Lin et al., 2022b) is the most popular dataset used to quantify hallucination level of a LLM. This dataset is adversarially constructed to exploit the weakness of LLM, which contained 817 questions that span 38 categories. OpenAI (2023) leveraged real-world data flagged as non-factual to construct an adversarial dataset to test GPT-4's hallucination potential. BIG-bench (Srivastava et al., 2022), a famous benchmark to evaluate LLM's capabilities, also contains many sub-tasks on factual correctness including TruthfulQA. Although most of these tasks are multiple choices or classification in a fact verification(Thorne et al., 2018) manner, they are closely associated with truthfulness and can be regarded as a generalized hallucination evaluation.

2.3.2 Uncertainty and Calibration

A reliable and trustworthy Language model must have the capability to accurately articulate its level of confidence over its response, which requires the model to be aware of its uncertainty. A model that can precisely measure its own uncertainty is sometimes called self-aware, honesty or known-unknown (Kadavath et al., 2022; Yin et al., 2023). In general deep learning applications, calibration concerns about the uncertainty estimation of a classifier. Output probability from a well-calibrated classifier are supposed to be consistent with the empirical accuracy in real world (Vaicenavicius et al., 2019). HELM (Liang et al., 2022) treated calibration as one of general metrics and comprehensively evaluated the calibration degree of many prevailing models on multiple choice and classification tasks. (OpenAI, 2023) also showed that GPT-4 before RLHF was well-calibrated on multiple choice tasks, although the decent calibration degree was compromised significantly by post-training.

when it comes to free-form generation, it's a different story. Kuhn et al. (2023) pointed out that semantic nature of language and intractable output space guaranteed the uniqueness of free-form generation.

Dataset	Safety Category	LLM evaluated	Task Format	Lang
RealToxicityPrompts	Harmful Contents	InstructGPT;LLaMA;Flan-PaLM;GPT-4;BLOOM	Generation	En
BAD	Harmful Contents	-	Generation	En
CrowS-Pairs	Social Bias	LLaMA;MT-NLG;InstructGPT;Pythia	Generatio	En
French CrowS-Pairs	Social Bias	MT-NLG	Generation	Fr
StereoSet	Social Bias	-	Multiple choice	En

Table 3: Datasets used to evaluate the safety competency of LLMs.

They proposed an algorithm to cluster model outputs and then estimate the model uncertainty. Mielke et al. (2022) claimed that models always express confidence over incorrect answers and proposed the notion of linguistic calibration, which teached models to verbally express uncertainty rather than estimating a probability. Lin et al. (2022a) trained models to directly generate predicted uncertainty probability in natural language. Yin et al. (2023) proposed the SelfAware dataset which contains unanswerable questions and used the accuracy of model rejection as a measure of uncertainty.

2.4 Safety

As the LLMs rapidly penetrate into the manufactural and interactive activities of human society, such as LLM-based poem-template generators and chatting robots, the safety concerns for LLMs gain much attention nowadays. The rationales of LLMs are statistics-based, and this inherent stochasticity brings limitations and underlying risks, which deeply affect the real-world deployment of LLMs. Some datasets are proposed to evaluated the safety of LLMs (Table 3), however, the corresponding validity and authority of the safety judgement are inadequate as the current evaluative dimensions are not sufficient (Waseem et al., 2017; Weidinger et al., 2021) and the perception of safety is highly subjective (Kocoń et al., 2021; Weidinger et al., 2021). To this end, based on our survey on relevant papers, we propose a comprehensive perspective on the safety competency of LLMs, ranging from harmful contents to the ethical consideration, to inspire the further developments towards the techniques and evaluations of LLMs safety.

2.4.1 Harmfulness

The harmful contents include the offensive language or others that have the explicit harm towards the specific object, such content that has been widely discussed. However, there is not a unified definition of the constitution of harmful contents, based on our surveys, we conclude the relevant themes into five aspects, including offensiveness, violence, crime, sexual-explicit, and unauthorized expertise. Many researches focus on the language detection for the outputs of LLMs to ensure the harmlessness (Wulczyn et al., 2017; Davidson et al., 2017; Zampieri et al., 2019; Dinan et al., 2019a), while other techniques are proposed to stimulate LLMs to generate safe outputs directly (Krause et al., 2021; Atwell et al., 2022). For the unauthorized expertise, a general LLM should avoid any unauthorized expertise before the establishment of accountability system (Sun et al., 2022), which involves the psychological orientation and any medical advice. Besides, the impact of conversation context on safety gains more attention recently, as a results, detective and generative algorithms base on the context are proposed successively (Dinan et al., 2019a; Baheti et al., 2021; Dinan et al., 2022). RealToxicityPrompts (Gehman et al., 2020) is a dataset derived from English web texts, where prompts are automatically truncated from sentences classified as toxicity from a widely-used toxicity classifier. RealToxicityPrompts consists of 100K natural prompts, with average 11.7 tokens in length. BAD (Xu et al., 2021) is a dataset collected by the human-in-the-loop strategy, where crowdworkers are ask to prob harmful model outputs. BAD consist of 5k conversations with around 70k utterances in total, which could be used in both non-adversarially and adversarially testing the model weakness.

2.4.2 Unfairness and Social Bias

Unfairness and social bias present more covertly and widely for LLMs. Following the previous studies, we conclude that social bias is an inherent characteristic of a LLM, which mainly embody in the distribution difference of a LLM in language selection based on different demographic groups. Compared to the social bias, unfairness is the external form, which reflected in the output performance of specific tasks, for example, the African American English (AAE) is frequently mis-classified as the offensive lan-

guage by some language detector (Lwowski et al., 2022). However, issues of unfairness and social bias are inevitable as they are widely distributed in human languages, and LLMs are required to memorize language as accurately as possible in the training stage (Weidinger et al., 2021). With respect to evaluate this important aspect, CrowS-Pairs (Nangia et al., 2020) is benchmark proposed to evaluating social bias. There are 1508 examples in CrowS-Pairs that involves nine types of social bias, like gender, race, and Nationality. StereoSet (Nadeem et al., 2021) is a dataset that could be used to evaluate social bias level in both word-level and sentence level, which examples are in four domains: race, gender,religion, and profession. For the StereoSet, the bias level is computed by the difference between model generation probabilities of biased and anti-biased sentence.

2.4.3 Others

As current algorithms for model safety based on the human perception, there is still no golden standardized judgement for LLMs to refer to, especially when a judgement is highly various across societies. It is necessary to align LLMs with the morality, ethics, and values of human society. More and more works focus on reifying this abstract concept into textual data recently, for example, Sap et al. (2020) proposal an implicit reasoning frame to explain the underlying harm of the target language. Besides, other works leverage rule-of-thumb (RoT) annotations of texts to support the judgement (Forbes et al., 2020; Ziems et al., 2022). However, current works in this area are neonatal, and we could expect more related works in the future.

Besides, we are also concerned about the privacy and political risks of LLMs. Since the LLMs are trained on vast corpus collected from books, conversations, web texts and so on, the privacy safety of LLMs arouses people's concern. These training texts might contain the private or sensitive information such as personal physical information, home address, etc. Many studies indicate LLMs are brittle under attacks, leaking the sensitive information unintentionally (Carlini et al., 2020; Li et al., 2022). Therefore, it is essential to test the privacy protection ability of a LLM. Moreover, the politics ignorance is also intractable for a LLM. The politics-related risk mainly stems from the composition of the training corpus. Texts in the corpus are derived from different language and social environments (usually the larger the more diversified), and different countries have different political prudence and stance, which brings additional risks to the wide deployment of a LM.

3 Future Directions

In this section, we outline some other competencies that are important for evaluating LLMs.

3.1 Sentiment

It is crucial to equip LLMs with the ability to understand and generate sentiments. As an indispensable factor in human life, sentiments are widely present in daily chats, social media posts, customer reviews, and news articles (Liu, 2015). Through the comprehensive research and high-level summary of the literature related to sentiments, we introduce the sentiment competency of LLMs in two aspects: sentiment understand and sentiment generation.

3.1.1 Sentiment Understand

Sentiment understand mainly involves the understanding of opinions, sentiments and emotions in the text (Liu, 2015). Representative tasks that reflect this competency include sentiment classification (SC), aspect-based sentiment analysis (ABSA), and multifaceted analysis of subjective texts (MAST). SC aims at assigning pre-defined sentiment classes to given texts. The typical datasets include IMDB (Maas et al., 2011), SST (Socher et al., 2013), Twitter (Rosenthal et al., 2017), Yelp (Zhang et al., 2015). ABSA focuses on identifying the sentiments of specific aspects in a sentence (Zhang et al., 2022), and the most widely used datasets are the SemEval series (Pontiki et al., 2014; Pontiki et al., 2015; Pontiki et al., 2016). MAST are tasks that involve the finer-grained and broader range of human subjective feelings (emotions (Sailunaz et al., 2018), stance (Küçük and Can, 2021), hate (Schmidt and Wiegand, 2017), irony (Zeng and Li, 2022), offensive (Pradhan et al., 2020), etc.) (Poria et al., 2023). Given that MAST includes a wide range of tasks, the datasets are not listed here in detail. Among them, the

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 88-109, Harbin, China, August 3 - 5, 2023.
Volume 2: Frotier Forum

commonly used evaluation metrics for the above tasks are accuracy and F1 score (micro or macro). Some preliminary empirical studies (Zhang et al., 2023; Wang et al., 2023) indicate that LLMs can significantly improve performance on these tasks in few-shot learning settings. LLMs have the potential to be a general solution without designing different models for various tasks. Therefore, the sentiment understand competency of different LLMs deserves comprehensive exploration and empirical evaluation. To evaluate the performance of this competency, we can utilize multiple domain-specific datasets or choose the comprehensive benchmark (Srivastava et al., 2022; Liang et al., 2022).

3.1.2 Sentiment Generation

We categorize sentiment generation into two manifestations. One is to generate text that contains sentiments, and the other is to generate text that elicits sentiments. The former requires specifying the desired sentiment, and the latter requires a combination of commonsense knowledge (Speer et al., 2017; Hwang et al., 2021) or theory of mind (Sodian and Kristen, 2010). A classic application scenario is in open-domain dialogue, specifically, emotional dialogue (Zhou et al., 2018), empathetic dialogue (Rashkin et al., 2019), and emotional support conversation (Liu et al., 2021). To measure the quality of the generated text, it is necessary to employ both automatic metrics (such as sentiment accuracy, BLEU (Papineni et al., 2002), perplexity) and human evaluations (human ratings or preference tests). Currently, no work has comprehensively explored this aspect, but it is an essential path towards artificial general intelligence (AGI) (Bubeck et al., 2023).

3.2 Planning

Planning is the thinking before the actions take place. Given a specific goal, planning is the process to decide the means to achieve the goal. There're few works (Valmeekam et al., 2023; Valmeekam et al., 2022; Pallagani et al., 2023; Huang et al., 2022) that look at the planning ability of LLMs. Some of them focus on commonsense areas (Huang et al., 2022) like wedding or menu planning. Others adopted automated planning problems, formal language translators, and verifiers to automatically evaluate LLMs' competency(Valmeekam et al., 2023). With PDDL [5] represented problem descriptions and the translation of such problems into text and back, LLMs can thus sequence a series of actions to reach the planning goal. Whether the planning purpose is achieved can be easily verified via automatic verifiers. Possessing web-scale knowledge, LLMs have great potential for executing planning tasks or assisting planners.

3.3 Code

Coding competency is one of the advanced abilities of LLMs. LLMs with this competency can not only perform program synthesis but also possess the potential of self-evolving. Technically, all of the tasks involved with code like code generation and code understanding need this competency. In oracle manual evaluation, prominent LLMs like ChatGPT are capable of up to 15 ubiquitous software engineering tasks and perform well in most of them (Sridhara et al., 2023). The most explored evaluation task in coding competency would be program synthesis, where program description and function signature are given for its code implementation. One of the most pioneering benchmarks in program synthesis, HUMANEVAL (Chen et al., 2021), consists of 164 pairs of human-generated docstrings and the associated unit tests to test the functional correctness of model generation. However, with the worry of insufficient testing and the imprecise problem description (Liu et al., 2023), existing LLM-for-code benchmarks still have lots of room for improvement.

4 Conclusion

This survey provides a comprehensive review of various literature for the evaluation of LLMs. We aggregate different works with their intended competencies. Some of the competencies(reasoning, knowledge) already have holistic evaluation benchmarks, while others(planning, coding) still face disparate challenges. The goal of this paper is to comb the numerous work concerning LLMs' evaluation through the lens of the core competencies test. Lighten the cognitive load for assimilating numerous evaluation

[5]Planning Domain Definition Language, a formal language used to describe classical planning problems.

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 88-109, Harbin, China, August 3 - 5, 2023.
Volume 2: Frotier Forum

works due to the various functions of LLMs. In doing so, we have also identified the challenge faced by each competency, looking forward to alleviating it in the future.

Acknowledgements

We want to thank Yuanxing Liu, Xuesong Wang, Mengzhou Sun, Runze Liu, Yuhang Gou, Shuhan Zhou, Yifan Chen, Ruiyu Xiao, Xinyu Li, Yuchi Zhang, Yang Wang, Jiahang Han, Wenqi Ding, and Xinpeng Liu for their priceless help with the initial dataset investigation process.

References

Patricia A Alexander. 1992. Domain knowledge: Evolving themes and emerging concerns. *Educational psychologist*, 27(1):33–51.

Rami Aly, Zhijiang Guo, M. 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. *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*.

Katherine Atwell, Sabit Hassan, and Malihe Alikhani. 2022. APPDIA: A discourse-aware transformer-based style transfer model for offensive social media conversations. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6063–6074, Gyeongju, Republic of Korea, October. International Committee on Computational Linguistics.

John Langshaw Austin. 1975. *How to do things with words*, volume 88. Oxford university press.

Ashutosh Baheti, Maarten Sap, Alan Ritter, and Mark Riedl. 2021. Just say no: Analyzing the stance of neural dialogue generation in offensive contexts. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4846–4862, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *CoRR*, abs/2302.04023.

Yonatan Belinkov, Sebastian Gehrmann, and Ellie Pavlick. 2020. Interpretability and analysis in neural NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 1–5, Online, July. Association for Computational Linguistics.

Prajjwal Bhargava and Vincent Ng. 2022. Commonsense knowledge reasoning and generation with pre-trained language models: A survey. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 12317–12325. AAAI Press.

Steven Bills, Nick Cammarata, Dan Mossing, Henk Tillman, Leo Gao, Gabriel Goh, Ilya Sutskever, Jan Leike, Jeff Wu, and William Saunders. 2023. Language models can explain neurons in language models. https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. PIQA: reasoning about physical commonsense in natural language. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7432–7439. AAAI Press.

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. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott M. Lundberg, Harsha Nori, Hamid Palangi, Marco Túlio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with GPT-4. *CoRR*, abs/2303.12712.

Pawel Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gasic. 2018. Multiwoz - A large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 5016–5026. Association for Computational Linguistics.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom B. Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. 2020. Extracting training data from large language models. *CoRR*, abs/2012.07805.

Tommaso Caselli and Piek Vossen. 2017. The event storyline corpus: A new benchmark for causal and temporal relation extraction. In Tommaso Caselli, Ben Miller, Marieke van Erp, Piek Vossen, Martha Palmer, Eduard H. Hovy, Teruko Mitamura, and David Caswell, editors, *Proceedings of the Events and Stories in the News Workshop@ACL 2017, Vancouver, Canada, August 4, 2017*, pages 77–86. Association for Computational Linguistics.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael J. Bommarito II, Ion Androutsopoulos, Daniel Martin Katz, and Nikolaos Aletras. 2022. Lexglue: A benchmark dataset for legal language understanding in english. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 4310–4330. Association for Computational Linguistics.

Chunkit Chan, Jiayang Cheng, Weiqi Wang, Yuxin Jiang, Tianqing Fang, Xin Liu, and Yangqiu Song. 2023. Chatgpt evaluation on sentence level relations: A focus on temporal, causal, and discourse relations. *CoRR*, abs/2304.14827.

Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. 2020a. Tabfact: A large-scale dataset for table-based fact verification. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Wenhu Chen, Hanwen Zha, Zhiyu Chen, Wenhan Xiong, Hong Wang, and William Yang Wang. 2020b. Hybridqa: A dataset of multi-hop question answering over tabular and textual data. In Trevor Cohn, Yulan He, and Yang Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 1026–1036. Association for Computational Linguistics.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Pondé de Oliveira Pinto, Jared Kaplan, Harrison 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, Joshua 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. 2021. Evaluating large language models trained on code. *CoRR*, abs/2107.03374.

Zhoujun Cheng, Tianbao Xie, Peng Shi, Chengzu Li, Rahul Nadkarni, Yushi Hu, Caiming Xiong, Dragomir Radev, Mari Ostendorf, Luke Zettlemoyer, Noah A. Smith, and Tao Yu. 2023. Binding language models in symbolic languages. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Peter Clark, Oyvind Tafjord, and Kyle Richardson. 2020. Transformers as soft reasoners over language. In Christian Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 3882–3890. ijcai.org.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *CoRR*, abs/2110.14168.

Dhairya Dalal, Paul Buitelaar, and Mihael Arcan. 2023. Calm-bench: A multi-task benchmark for evaluating causality-aware language models. In Andreas Vlachos and Isabelle Augenstein, editors, *Findings of the Association for Computational Linguistics: EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, pages 296–311. Association for Computational Linguistics.

Souvik Das, Sougata Saha, and Rohini K. Srihari. 2022. Diving deep into modes of fact hallucinations in dialogue systems. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 684–699. Association for Computational Linguistics.

Thomas Davidson, Dana Warmsley, Michael W. Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *International Conference on Web and Social Media*.

Ernest Davis. 2014. *Representations of commonsense knowledge*. Morgan Kaufmann.

Richard R Day, Julian Bamford, Willy A Renandya, George M Jacobs, and Vivienne Wai-Sze Yu. 1998. Extensive reading in the second language classroom. *RELC Journal*, 29(2):187–191.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A benchmark to evaluate rationalized NLP models. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 4443–4458. Association for Computational Linguistics.

Emily Dinan, Samuel Humeau, Bharath Chintagunta, and Jason Weston. 2019a. Build it break it fix it for dialogue safety: Robustness from adversarial human attack. 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 4537–4546, Hong Kong, China, November. Association for Computational Linguistics.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019b. Wizard of wikipedia: Knowledge-powered conversational agents. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Emily Dinan, Gavin Abercrombie, A. Bergman, Shannon Spruit, Dirk Hovy, Y-Lan Boureau, and Verena Rieser. 2022. SafetyKit: First aid for measuring safety in open-domain conversational systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4113–4133, Dublin, Ireland, May. 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 for in-context learning. *CoRR*, abs/2301.00234.

Li Du, Xiao Ding, Kai Xiong, Ting Liu, and Bing Qin. 2022. e-care: a new dataset for exploring explainable causal reasoning. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 432–446. Association for Computational Linguistics.

Selma Dündar-Coecke. 2022. To what extent is general intelligence relevant to causal reasoning? a developmental study. *Frontiers in Psychology*, 13.

Mihail Eric, Lakshmi Krishnan, François Charette, and Christopher D. Manning. 2017. Key-value retrieval networks for task-oriented dialogue. In Kristiina Jokinen, Manfred Stede, David DeVault, and Annie Louis, editors, *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue, Saarbrücken, Germany, August 15-17, 2017*, pages 37–49. Association for Computational Linguistics.

Owain Evans, Owen Cotton-Barratt, Lukas Finnveden, Adam Bales, Avital Balwit, Peter Wills, Luca Righetti, and William Saunders. 2021. Truthful AI: developing and governing AI that does not lie. *CoRR*, abs/2110.06674.

Jonathan Evans. 2002. Logic and human reasoning: an assessment of the deduction paradigm. *Psychological bulletin*, 128 6:978–96.

Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020. Social chemistry 101: Learning to reason about social and moral norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 653–670, Online, November. Association for Computational Linguistics.

Victoria Fromkin, Robert Rodman, and Nina Hyams. 2018. *An Introduction to Language (w/MLA9E Updates)*. Cengage Learning.

Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, Jason Phang, Laria Reynolds, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2021. A framework for few-shot language model evaluation, September.

Jinglong Gao, Xiao Ding, Bing Qin, and Ting Liu. 2023. Is chatgpt a good causal reasoner? A comprehensive evaluation. *CoRR*, abs/2305.07375.

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, November. Association for Computational Linguistics.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? A question answering benchmark with implicit reasoning strategies. *Trans. Assoc. Comput. Linguistics*, 9:346–361.

Andrew S. Gordon, Zornitsa Kozareva, and Melissa Roemmele. 2012. Semeval-2012 task 7: Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In Eneko Agirre, Johan Bos, and Mona T. Diab, editors, *Proceedings of the 6th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2012, Montréal, Canada, June 7-8, 2012*, pages 394–398. The Association for Computer Linguistics.

Shirley Anugrah Hayati, Dongyeop Kang, and Lyle Ungar. 2021. Does BERT learn as humans perceive? understanding linguistic styles through lexica. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 6323–6331. Association for Computational Linguistics.

Jack Hessel, Jena D. Hwang, Jae Sung Park, Rowan Zellers, Chandra Bhagavatula, Anna Rohrbach, Kate Saenko, and Yejin Choi. 2022. The abduction of sherlock holmes: A dataset for visual abductive reasoning. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXVI*, volume 13696 of *Lecture Notes in Computer Science*, pages 558–575. Springer.

Terrence Hoffmann. 1999. The meanings of competency. *Journal of european industrial training*, 23(6):275–286.

Mohammad Javad Hosseini, Hannaneh Hajishirzi, Oren Etzioni, and Nate Kushman. 2014. Learning to solve arithmetic word problems with verb categorization. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 523–533. ACL.

Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. 2022. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato, editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 9118–9147. PMLR.

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. 2023. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *CoRR*, abs/2305.08322.

Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021. (comet-) atomic 2020: On symbolic and neural commonsense knowledge graphs. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 6384–6392. AAAI Press.

Mohit Iyyer, Wen-tau Yih, and Ming-Wei Chang. 2017. Search-based neural structured learning for sequential question answering. In Regina Barzilay and Min-Yen Kan, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1821–1831. Association for Computational Linguistics.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12):248:1–248:38.

Jinhao Jiang, Kun Zhou, Zican Dong, Keming Ye, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Structgpt: A general framework for large language model to reason over structured data. *CoRR*, abs/2305.09645.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In Regina Barzilay and Min-Yen Kan, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1601–1611. 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. *CoRR*, abs/2207.05221.

Antonis C. Kakas and Loizos Michael. 2020. Abduction and argumentation for explainable machine learning: A position survey. *CoRR*, abs/2010.12896.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *CoRR*, abs/2001.08361.

Jan Kocoń, Alicja Figas, Marcin Gruza, Daria Puchalska, Tomasz Kajdanowicz, and Przemysław Kazienko. 2021. Offensive, aggressive, and hate speech analysis: From data-centric to human-centered approach. *Information Processing Management*, 58(5):102643.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *NeurIPS*.

Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq Joty, Richard Socher, and Nazneen Fatema Rajani. 2021. GeDi: Generative discriminator guided sequence generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4929–4952, Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Dilek Küçük and Fazli Can. 2021. Stance detection: A survey. *ACM Comput. Surv.*, 53(1):12:1–12:37.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Katherine Lee, Orhan Firat, Ashish Agarwal, Clara Fannjiang, and David Sussillo. 2018. Hallucinations in neural machine translation.

Haoran Li, Yangqiu Song, and Lixin Fan. 2022. You don't know my favorite color: Preventing dialogue representations from revealing speakers' private personas. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5858–5870, Seattle, United States, July. Association for Computational Linguistics.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2023a. CMMLU: measuring massive multitask language understanding in chinese. *CoRR*, abs/2306.09212.

Xingxuan Li, Ruochen Zhao, Yew Ken Chia, Bosheng Ding, Lidong Bing, Shafiq R. Joty, and Soujanya Poria. 2023b. Chain of knowledge: A framework for grounding large language models with structured knowledge bases. *CoRR*, abs/2305.13269.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel J. Orr, Lucia Zheng, Mert Yüksekgönül, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2022. Holistic evaluation of language models. *CoRR*, abs/2211.09110.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022a. Teaching models to express their uncertainty in words. *Trans. Mach. Learn. Res.*, 2022.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022b. Truthfulqa: Measuring how models mimic human falsehoods. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 3214–3252. Association for Computational Linguistics.

Adam Dahlgren Lindström and Savitha Sam Abraham. 2022. Clevr-math: A dataset for compositional language, visual and mathematical reasoning. In Artur S. d'Avila Garcez and Ernesto Jiménez-Ruiz, editors, *Proceedings of the 16th International Workshop on Neural-Symbolic Learning and Reasoning as part of the 2nd International Joint Conference on Learning & Reasoning (IJCLR 2022), Cumberland Lodge, Windsor Great Park, UK, September 28-30, 2022*, volume 3212 of *CEUR Workshop Proceedings*, pages 155–170. CEUR-WS.org.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 3469–3483. Association for Computational Linguistics.

Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. 2023. Is your code generated by chatgpt really correct? rigorous evaluation of large language models for code generation. *CoRR*, abs/2305.01210.

Bing Liu. 2015. *Sentiment Analysis - Mining Opinions, Sentiments, and Emotions*. Cambridge University Press.

Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. 2023. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Brandon Lwowski, Paul Rad, and Anthony Rios. 2022. Measuring geographic performance disparities of offensive language classifiers. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6600–6616, Gyeongju, Republic of Korea, October. International Committee on Computational Linguistics.

Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. 2023. Faithful chain-of-thought reasoning. *CoRR*, abs/2301.13379.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In Dekang Lin, Yuji Matsumoto, and Rada Mihalcea, editors, *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA*, pages 142–150. The Association for Computer Linguistics.

Daniel Jaymin Mankowitz, Andrea Michi, Anton Zhernov, Marco Gelmi, Marco Selvi, Cosmin Paduraru, Edouard Leurent, Shariq Iqbal, Jean-Baptiste Lespiau, Alex Ahern, Thomas Köppe, Kevin Millikin, Stephen Gaffney, Sophie Elster, Jackson Broshear, Chris Gamble, Kieran Milan, Robert Tung, Minjae Hwang, taylan. cemgil, Mohammadamin Barekatain, Yujia Li, Amol Mandhane, Thomas Hubert, Julian Schrittwieser, Demis Hassabis, Pushmeet Kohli, Martin A. Riedmiller, Oriol Vinyals, and David Silver. 2023. Faster sorting algorithms discovered using deep reinforcement learning. *Nature*, 618:257 – 263.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 14867–14875. AAAI Press.

Grégoire Mialon, Roberto Dessì, Maria Lomeli, Christoforos Nalmpantis, Ramakanth Pasunuru, Roberta Raileanu, Baptiste Rozière, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, Edouard Grave, Yann LeCun, and Thomas Scialom. 2023. Augmented language models: a survey. *CoRR*, abs/2302.07842.

Sabrina J. Mielke, Arthur Szlam, Emily Dinan, and Y-Lan Boureau. 2022. Reducing conversational agents' overconfidence through linguistic calibration. *Trans. Assoc. Comput. Linguistics*, 10:857–872.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? A new dataset for open book question answering. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2381–2391. Association for Computational Linguistics.

Paramita Mirza, R. Sprugnoli, Sara Tonelli, and Manuela Speranza. 2014. Annotating causality in the tempeval-3 corpus. In *Conference of the European Chapter of the Association for Computational Linguistics*.

Swaroop Mishra, Matthew Finlayson, Pan Lu, Leonard Tang, Sean Welleck, Chitta Baral, Tanmay Rajpurohit, Oyvind Tafjord, Ashish Sabharwal, Peter Clark, and Ashwin Kalyan. 2022a. LILA: A unified benchmark for mathematical reasoning. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 5807–5832. Association for Computational Linguistics.

Swaroop Mishra, Arindam Mitra, Neeraj Varshney, Bhavdeep Singh Sachdeva, Peter Clark, Chitta Baral, and Ashwin Kalyan. 2022b. Numglue: A suite of fundamental yet challenging mathematical reasoning tasks. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 3505–3523. Association for Computational Linguistics.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. 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 5356–5371, Online, August. Association for Computational Linguistics.

Linyong Nan, Chia-Hsuan Hsieh, Ziming Mao, Xi Victoria Lin, Neha Verma, Rui Zhang, Wojciech Kryscinski, Nick Schoelkopf, Riley Kong, Xiangru Tang, Murori Mutuma, Benjamin Rosand, Isabel Trindade, Renusree Bandaru, Jacob Cunningham, Caiming Xiong, and Dragomir R. Radev. 2021. Fetaqa: Free-form table question answering. *Transactions of the Association for Computational Linguistics*, 10:35–49.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online, November. Association for Computational Linguistics.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. 2022. In-context learning and induction heads. *CoRR*, abs/2209.11895.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

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

Ekaterina Ovchinnikova. 2012. *Integration of World Knowledge for Natural Language Understanding*, volume 3 of *Atlantis Thinking Machines*. Atlantis Press.

Vishal Pallagani, Bharath Muppasani, Keerthiram Murugesan, Francesca Rossi, Biplav Srivastava, Lior Horesh, Francesco Fabiano, and Andrea Loreggia. 2023. Understanding the capabilities of large language models for automated planning. *CoRR*, abs/2305.16151.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

Ankur P. Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. 2020. Totto: A controlled table-to-text generation dataset. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 1173–1186. Association for Computational Linguistics.

Panupong Pasupat and Percy Liang. 2015. Compositional semantic parsing on semi-structured tables. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, pages 1470–1480. The Association for Computer Linguistics.

Edoardo Maria Ponti, Goran Glavas, Olga Majewska, Qianchu Liu, Ivan Vulic, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 2362–2376. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. Semeval-2014 task 4: Aspect based sentiment analysis. In Preslav Nakov and Torsten Zesch, editors, *Proceedings of the 8th International Workshop on Semantic Evaluation, SemEval@COLING 2014, Dublin, Ireland, August 23-24, 2014*, pages 27–35. The Association for Computer Linguistics.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. Semeval-2015 task 12: Aspect based sentiment analysis. In Daniel M. Cer, David Jurgens, Preslav Nakov, and Torsten Zesch, editors, *Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2015, Denver, Colorado, USA, June 4-5, 2015*, pages 486–495. The Association for Computer Linguistics.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Nuria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit. 2016. SemEval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 19–30, San Diego, California, June. Association for Computational Linguistics.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, and Rada Mihalcea. 2023. Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research. *IEEE Trans. Affect. Comput.*, 14(1):108–132.

Rahul Pradhan, Ankur Chaturvedi, Aprna Tripathi, and Dilip Kumar Sharma. 2020. A review on offensive language detection. In Mohan L. Kolhe, Shailesh Tiwari, Munesh C. Trivedi, and Krishn K. Mishra, editors, *Advances in Data and Information Sciences*, pages 433–439, Singapore. Springer Singapore.

Shuofei Qiao, Yixin Ou, Ningyu Zhang, Xiang Chen, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, and Huajun Chen. 2022. Reasoning with language model prompting: A survey. *CoRR*, abs/2212.09597.

Lin Qiu, Yunxuan Xiao, Yanru Qu, Hao Zhou, Lei Li, Weinan Zhang, and Yong Yu. 2019. Dynamically fused graph network for multi-hop reasoning. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 6140–6150. Association for Computational Linguistics.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5370–5381. Association for Computational Linguistics.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. Semeval-2017 task 4: Sentiment analysis in twitter. In Steven Bethard, Marine Carpuat, Marianna Apidianaki, Saif M. Mohammad, Daniel M. Cer, and David Jurgens, editors, *Proceedings of the 11th International Workshop on Semantic Evaluation, SemEval@ACL 2017, Vancouver, Canada, August 3-4, 2017*, pages 502–518. Association for Computational Linguistics.

Kashfia Sailunaz, Manmeet Dhaliwal, Jon G. Rokne, and Reda Alhajj. 2018. Emotion detection from text and speech: a survey. *Soc. Netw. Anal. Min.*, 8(1):28:1–28:26.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online, July. Association for Computational Linguistics.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In Lun-Wei Ku and Cheng-Te Li, editors, *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, SocialNLP@EACL 2017, Valencia, Spain, April 3, 2017*, pages 1–10. Association for Computational Linguistics.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2023. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Koustuv Sinha, Shagun Sodhani, Jin Dong, Joelle Pineau, and William L. Hamilton. 2019. CLUTRR: A diagnostic benchmark for inductive reasoning from text. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 4505–4514. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA, October. Association for Computational Linguistics.

Beate Sodian and Susanne Kristen, 2010. *Theory of Mind*, pages 189–201. Springer Berlin Heidelberg, Berlin, Heidelberg.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In Satinder Singh and Shaul Markovitch, editors, *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 4444–4451. AAAI Press.

Giriprasad Sridhara, Ranjani H. G., and Sourav Mazumdar. 2023. Chatgpt: A study on its utility for ubiquitous software engineering tasks. *CoRR*, abs/2305.16837.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Santilli, Andreas Stuhlmüller, Andrew M. Dai, Andrew La, Andrew K. Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakas, and et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *CoRR*, abs/2206.04615.

Shane Storks, Qiaozi Gao, and Joyce Y. Chai. 2019. Commonsense reasoning for natural language understanding: A survey of benchmarks, resources, and approaches. *CoRR*, abs/1904.01172.

Hao Sun, Guangxuan Xu, Jiawen Deng, Jiale Cheng, Chujie Zheng, Hao Zhou, Nanyun Peng, Xiaoyan Zhu, and Minlie Huang. 2022. On the safety of conversational models: Taxonomy, dataset, and benchmark. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3906–3923, Dublin, Ireland, May. Association for Computational Linguistics.

Dídac Surís, Sachit Menon, and Carl Vondrick. 2023. Vipergpt: Visual inference via python execution for reasoning. *CoRR*, abs/2303.08128.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. 2014. Intriguing properties of neural networks. In Yoshua Bengio and Yann LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4149–4158. Association for Computational Linguistics.

Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hannaneh Hajishirzi, and Jonathan Berant. 2021. Multimodalqa: complex question answering over text, tables and images. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

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

Juozas Vaicenavicius, David Widmann, Carl R. Andersson, Fredrik Lindsten, Jacob Roll, and Thomas B. Schön. 2019. Evaluating model calibration in classification. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan*, volume 89 of *Proceedings of Machine Learning Research*, pages 3459–3467. PMLR.

Karthik Valmeekam, Alberto Olmo Hernandez, Sarath Sreedharan, and Subbarao Kambhampati. 2022. Large language models still can't plan (A benchmark for llms on planning and reasoning about change). *CoRR*, abs/2206.10498.

Karthik Valmeekam, Matthew Marquez, Sarath Sreedharan, and Subbarao Kambhampati. 2023. On the planning abilities of large language models - A critical investigation. *CoRR*, abs/2305.15771.

Matthew J. Vowels, Necati Cihan Camgöz, and Richard Bowden. 2023. D'ya like dags? A survey on structure learning and causal discovery. *ACM Comput. Surv.*, 55(4):82:1–82:36.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 7534–7550. Association for Computational Linguistics.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019a. Superglue: A stickier benchmark for general-purpose language understanding systems. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 3261–3275.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019b. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Dingzirui Wang, Longxu Dou, and Wanxiang Che. 2022a. A survey on table-and-text hybridqa: Concepts, methods, challenges and future directions. *CoRR*, abs/2212.13465.

Lijie Wang, Yaozong Shen, Shuyuan Peng, Shuai Zhang, Xinyan Xiao, Hao Liu, Hongxuan Tang, Ying Chen, Hua Wu, and Haifeng Wang. 2022b. A fine-grained interpretability evaluation benchmark for neural NLP. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 70–84, Abu Dhabi, United Arab Emirates (Hybrid), December. Association for Computational Linguistics.

Xiaozhi Wang, Yulin Chen, Ning Ding, Hao Peng, Zimu Wang, Yankai Lin, Xu Han, Lei Hou, Juanzi Li, Zhiyuan Liu, Peng Li, and Jie Zhou. 2022c. MAVEN-ERE: A unified large-scale dataset for event coreference, temporal, causal, and subevent relation extraction. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 926–941. Association for Computational Linguistics.

Xuezhi Wang, Haohan Wang, and Diyi Yang. 2022d. Measure and improve robustness in NLP models: A survey. In Marine Carpuat, Marie-Catherine de Marneffe, and Iván Vladimir Meza Ruíz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 4569–4586. Association for Computational Linguistics.

Zengzhi Wang, Qiming Xie, Zixiang Ding, Yi Feng, and Rui Xia. 2023. Is chatgpt a good sentiment analyzer? A preliminary study. *CoRR*, abs/2304.04339.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. Blimp: The benchmark of linguistic minimal pairs for english. *Trans. Assoc. Comput. Linguistics*, 8:377–392.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84, Vancouver, BC, Canada, August. Association for Computational Linguistics.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022a. Emergent abilities of large language models. *Trans. Mach. Learn. Res.*, 2022.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022b. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2021. Ethical and social risks of harm from language models. *CoRR*, abs/2112.04359.

Yuhuai Wu, Markus N. Rabe, Wenda Li, Jimmy Ba, Roger B. Grosse, and Christian Szegedy. 2021. LIME: learning inductive bias for primitives of mathematical reasoning. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 11251–11262. PMLR.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, page 1391–1399, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Tianbao Xie, Chen Henry Wu, Peng Shi, Ruiqi Zhong, Torsten Scholak, Michihiro Yasunaga, Chien-Sheng Wu, Ming Zhong, Pengcheng Yin, Sida I. Wang, Victor Zhong, Bailin Wang, Chengzu Li, Connor Boyle, Ansong Ni, Ziyu Yao, Dragomir Radev, Caiming Xiong, Lingpeng Kong, Rui Zhang, Noah A. Smith, Luke Zettlemoyer, and Tao Yu. 2022. Unifiedskg: Unifying and multi-tasking structured knowledge grounding with text-to-text language models. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 602–631. Association for Computational Linguistics.

Jing Xu, Da Ju, Margaret Li, Y-Lan Boureau, Jason Weston, and Emily Dinan. 2021. Bot-adversarial dialogue for safe conversational agents. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2950–2968, Online, June. Association for Computational Linguistics.

Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. Wikiqa: A challenge dataset for open-domain question answering. In Lluís Màrquez, Chris Callison-Burch, Jian Su, Daniele Pighin, and Yuval Marton, editors, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 2013–2018. The Association for Computational Linguistics.

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. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2369–2380. Association for Computational Linguistics.

Zonglin Yang, Li Dong, Xinya Du, Hao Cheng, Erik Cambria, Xiaodong Liu, Jianfeng Gao, and Furu Wei. 2022. Language models as inductive reasoners. *CoRR*, abs/2212.10923.

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2022. Linkbert: Pretraining language models with document links. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 8003–8016. Association for Computational Linguistics.

Xi Ye and Greg Durrett. 2022. The unreliability of explanations in few-shot prompting for textual reasoning. In *NeurIPS*.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. 2023. Do large language models know what they don't know? *CoRR*, abs/2305.18153.

Fei Yu, Hongbo Zhang, Prayag Tiwari, and Benyou Wang. 2023. Natural language reasoning, a survey.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. 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 1415–1420, Minneapolis, Minnesota, June. Association for Computational Linguistics.

E. Zelikman, Yuhuai Wu, and Noah D. Goodman. 2022. Star: Bootstrapping reasoning with reasoning. *ArXiv*, abs/2203.14465.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4791–4800. Association for Computational Linguistics.

Qingcheng Zeng and An-Ran Li. 2022. A survey in automatic irony processing: Linguistic, cognitive, and multi-x perspectives. In Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na, editors, *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, pages 824–836. International Committee on Computational Linguistics.

Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 649–657.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2022. A survey on aspect-based sentiment analysis: Tasks, methods, and challenges. *CoRR*, abs/2203.01054.

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Jialin Pan, and Lidong Bing. 2023. Sentiment analysis in the era of large language models: A reality check. *CoRR*, abs/2305.15005.

Xinyan Zhao and V. G. Vinod Vydiswaran. 2021. Lirex: Augmenting language inference with relevant explanations. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 14532–14539. AAAI Press.

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

Shen Zheng, Jie Huang, and Kevin Chen-Chuan Chang. 2023. Why does chatgpt fall short in providing truthful answers?

Victor Zhong, Caiming Xiong, and Richard Socher. 2017. Seq2sql: Generating structured queries from natural language using reinforcement learning. *CoRR*, abs/1709.00103.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2023. Agieval: A human-centric benchmark for evaluating foundation models. *CoRR*, abs/2304.06364.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018. Emotional chatting machine: Emotional conversation generation with internal and external memory. In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 730–739. AAAI Press.

Terry Yue Zhuo, Yujin Huang, Chunyang Chen, and Zhenchang Xing. 2023. Red teaming chatgpt via jailbreaking: Bias, robustness, reliability and toxicity.

Caleb Ziems, Jane Yu, Yi-Chia Wang, Alon Halevy, and Diyi Yang. 2022. The moral integrity corpus: A benchmark for ethical dialogue systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3755–3773, Dublin, Ireland, May. Association for Computational Linguistics.