TOWARDS UNDERSTANDING FACTUAL KNOWLEDGE OF LARGE LANGUAGE MODELS

Xuming Hu[1,2*], Junzhe Chen[1*], Xiaochuan Li[1*], Yufei Guo[1], Lijie Wen[1†], Philip S. Yu[3], Zhijiang Guo[4†]
[1] Tsinghua University [2] The Hong Kong University of Science and Technology (Guangzhou)
[3] University of Illinois at Chicago [4] University of Cambridge
xuminghu@hkust-gz.edu.cn, wenlj@tsinghua.edu.cn, zg283@cam.ac.uk

ABSTRACT

Large language models (LLMs) have recently driven striking performance improvements across a range of natural language processing tasks. The factual knowledge acquired during pretraining and instruction tuning can be useful in various downstream tasks, such as question answering, and language generation. Unlike conventional Knowledge Bases (KBs) that explicitly store factual knowledge, LLMs implicitly store facts in their parameters. Content generated by the LLMs can often exhibit inaccuracies or deviations from the truth, due to facts that can be incorrectly induced or become obsolete over time. To this end, we aim to explore the extent and scope of factual knowledge within LLMs by designing the benchmark Pinocchio. Pinocchio contains 20K diverse factual questions that span different sources, timelines, domains, regions, and languages. Furthermore, we investigate whether LLMs can compose multiple facts, update factual knowledge temporally, reason over multiple pieces of facts, identify subtle factual differences, and resist adversarial examples. Extensive experiments on different sizes and types of LLMs show that existing LLMs still lack factual knowledge and suffer from various spurious correlations. We believe this is a critical bottleneck for realizing trustworthy artificial intelligence. The dataset Pinocchio and our codes are publicly available at: https://github.com/THU-BPM/Pinocchio.

1 INTRODUCTION

Large language models (LLMs) have revolutionized natural language processing (NLP) in recent years since they have significantly improved performance on various downstream tasks (Brown et al., 2020; Chowdhery et al., 2022; Ouyang et al., 2022; Touvron et al., 2023a;b; OpenAI, 2022; 2023). Prior efforts have shown that language models can store factual knowledge and act as knowledge bases (Petroni et al., 2019; Jiang et al., 2020c). Factual knowledge in language models acquired during pretraining can benefit knowledge-intensive downstream tasks such as question answering and fact checking (Roberts et al., 2020; Yu et al., 2023a; Pan et al., 2023).

Despite advancements in LLMs, they still struggle with generating content that exhibits inaccuracies or deviations from the facts and making reasoning errors (Lin et al., 2022; Bubeck et al., 2023). These factual errors can be difficult to identify since LLMs implicitly memorize facts through their parameters rather than explicitly store factual knowledge as traditional Knowledge Bases. Accessing and interpreting the computations and memories of these models can be challenging (Ribeiro et al., 2016; Belinkov & Glass, 2019), especially when APIs are the only means of interaction and many interpretation methods rely on weights and representations (Cao et al., 2021b). The presence of errors in stored factual knowledge or the incorrect induction and obsolescence of certain facts over time may be contributing factors to this limitation, which in turn affects the performance of LLMs (Elazar et al., 2021; Cao et al., 2021a). This limitation restricts the application of LLMs in some high-stakes areas, such as healthcare, finance, and law (Dong et al., 2022). Hence, exploring the degree to which LLMs hold factual information and their ability to reason with such knowledge is vital.

[*] Equal Contribution.
[†] Corresponding authors.

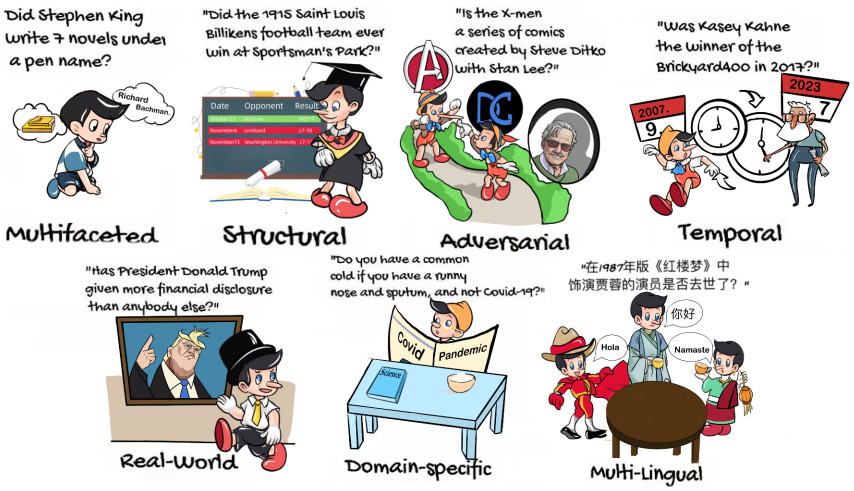


Figure 1: Pinocchio is a comprehensive dataset that tackles 7 distinct tasks related to factual knowledge and reasoning. It consists of 20,713 multiple-choice questions that have been sourced from various reliable and diverse channels.

To this end, we propose the Pinocchio, a testbed aimed at understanding factuality and reasoning for LLMs. It contains 20K diverse factual questions that span different sources, timelines, domains, regions, and languages. Furthermore, we investigate whether LLMs are able to recognize the combination of multiple facts, reason over structured and unstructured evidence, realize facts change over time, identify subtle factual differences, and resist adversarial examples based on the dataset. We control for problem difficulty in each distinct reasoning task to enable fine-grained analysis.

With the Pinocchio benchmark, we explore whether various LLMs (Scao et al., 2022b; Zhang et al., 2022; Ouyang et al., 2022; Chung et al., 2022; Touvron et al., 2023a; Chiang et al., 2023) could store factual knowledge and perform reasoning based on it. We envision Pinocchio as a suite of benchmarks, subsets of which could be separately utilized to assess certain model abilities of interest and analyze important strengths and limitations of LLMs. For instance, in temporal tasks, we find that LLMs lack factual knowledge for up-to-date questions; in complex factual tasks that require multi-hop reasoning, LLMs still have limitations, even when various prompting strategies are employed. We hope Pinocchio can serve as the initial step towards understanding the abilities of LLMs from multiple dimensions and facilitate the development of LLMs.

2 Dataset Construction

2.1 Tasks

Aiming to systematically evaluate the factual knowledge and related reasoning abilities of LLMs, we raise seven research questions, then carefully select factual statements from different sources summarized in Table 1.

- **Task 1: Multifaceted** Previous research (Petroni et al., 2019) has shown that small language models like BERT have the ability to retain relational knowledge from training data and answer "fill-in-the-blank" cloze statements. This raises the question of *whether LLMs can also store and reason over multiple pieces of facts obtained during pretraining*. It is not just important for LLMs to memorize individual facts accurately, but to also recognize and generate new combinations of facts from different sources. To investigate this issue, we have selected claims from the FEVER dataset (Thorne et al., 2018), which were written by human annotators based on information from Wikipedia articles. These claims are either supported or refuted by multiple facts from (the same or several) Wikipedia articles, or there is insufficient information available to verify them. To assess the performance of language models in handling various combinations of facts, we have sampled statements that require different numbers of evidence, ranging from one to many, enabling fine-grained analysis.
- **Task 2: Structural** In addition to unstructured text, factual knowledge is also commonly stored in a structured format, such as tables, lists, or databases (Bhagavatula et al., 2013). However,

Table 1: Pinocchio Dataset Sources, Descriptions, and Data Distribution.

Domain	Description	Sources	Distribution			
			Fact.	Non-Fact.	NEI	ALL
Multifaceted	Contain multiple facts	FEVER	1,111	1,111	1,110	3,332
Structural	Contain structured and unstructured facts	FEVEROUS	1,741	1,953	250	3,944
Adversarial	Contain facts edited by adversarial methods	Symmetric, FM2	815	921	-	1,736
Temporal	Contain facts that change over time	VitaminC	1,898	1,043	355	3,296
Real-World	Contain factual statements spread online	PolitiFact	986	1,987	609	3,582
Domain-Specific	Contain facts from health and science domains	PubHealth, SciFact	1,156	715	737	2,608
Multi-Lingual	Contain facts in different languages	XFact, CHEF	820	848	547	2,215

current LLMs are primarily trained on unstructured text using next word prediction loss (Brown et al., 2020; Touvron et al., 2023a). In order to process structured data, it is often converted into text strings using various methods, such as linearizing tables. This raises the question of *whether LLMs are capable of effectively memorizing and reasoning over facts from structured sources, similar to their performance with unstructured text*. To investigate this question, we sample factual statements from the FEVEROUS dataset (Aly et al., 2021), which is constructed in a similar manner to FEVER but includes evidence in the form of tables, sentences, or both.

- **Task 3: Adversarial** Language models are known to be vulnerable to adversarial examples that are strategically modified to deceive even advanced models with hardly noticeable changes (Shen et al., 2023). Given this knowledge, it is important to examine *whether LLMs can withstand adversarial examples in the context of factuality*. To investigate this, we utilize two datasets, namely Symmetric (Schuster et al., 2019) and FM2 (Eisenschlos et al., 2021). These datasets consist of adversarial examples that have been crafted using various strategies, including temporal inference and diverting to unrelated facts.
- **Task 4: Temporal** Facts are not static but rather possess a dynamic nature. With the vast amount of new information constantly emerging, facts often undergo changes, additions, or alterations. It raises the question of *whether LLMs are able to adapt to these factual changes over time*. In particular, we wonder if LLMs are capable of discerning factual knowledge from different time periods, since the pretraining corpus may not be processed and organized chronologically. To explore this, we utilize the VitaminC (Schuster et al., 2021) dataset, which consists of claims based on modifications made to factual content in Wikipedia articles. Claims can be either refuted by outdated facts or supported by updated facts.
- **Task 5: Real-World** In contrast to other tasks that assume Wikipedia has all the essential factual information, verifying viral claims on the internet often requires not only factual knowledge from various sources but also common sense and worldly knowledge. An important query we have is *whether LLMs can effectively integrate diverse types and sources of knowledge acquired during training*. To address this, we select claims from the FactCheck (Misra, 2022) dataset, which consists of claims spread over the Internet and subsequently verified by journalists.
- **Task 6: Domain-Specific** In addition to the tasks mentioned earlier, which primarily focus on factual knowledge in general domains, we are also interested in exploring *how LLMs possess the capability to access domain-specific factual knowledge*. The domain-specific setting presents unique challenges. Take the science domain as an example, LLMs need to acquire background knowledge, handle quantitative reasoning, and comprehend specialized statistical language. To investigate this further, we sample claims from PubHealth (Kotonya & Toni, 2020) in the public health domain and SciFact (Wadden et al., 2022) in the science domain.
- **Task 7: Multi-Lingual** Existing LLMs are mainly trained on English corpus because of their abundance and quality (Chowdhery et al., 2022; Touvron et al., 2023a). However, the scarcity of training data in other languages raises the question of *whether LLMs can transfer the factual knowledge acquired in English to other languages*. To investigate this, we collected claims from various languages including French, Chinese, and more, using the XFACT dataset (Gupta & Srikumar, 2021) and the CHEF dataset (Hu et al., 2022b) in a total of 27 different languages.

2.2 Annotation and Quality Control

Multiple-choice questions offer a practical approach to assess the complex capabilities of LLMs, of which GPT-4 is a prime example (OpenAI, 2023). Key benchmarks such as MMLU (Hendrycks et al., 2021b), HellaSwag (Zellers et al., 2019), ARC (Clark et al., 2018a), and TruthfulQA (Lin et al., 2022), all of which utilize multi-choice formats, serve distinct purposes in evaluating various aspects of GPT-4's proficiency. Specifically, the MMLU gauges an LLM's knowledge breadth and depth.

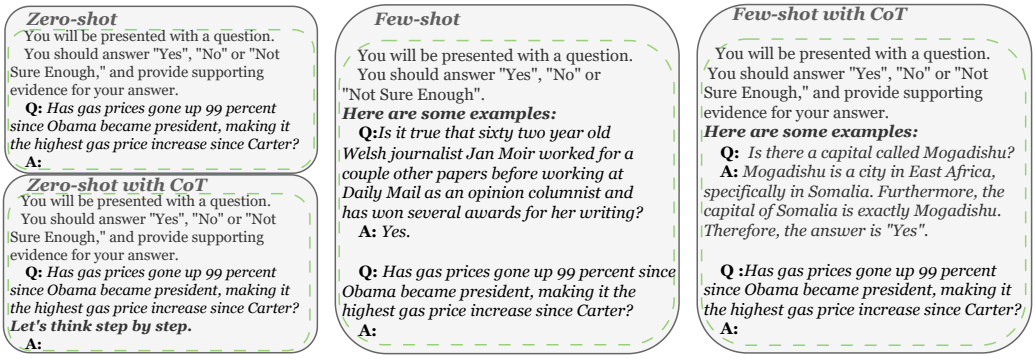


Figure 2: Illustration of prompts using different settings.

HellaSwag tests commonsense reasoning, and ARC focuses on challenging questions. TruthfulQA measures how LLMs mimic human falsehoods. Furthermore, the evaluation of language generation brings its own set of challenges, as a universal metric for measurement is currently lacking (Sai et al., 2023), which multiple-choice questions help to mitigate by offering straightforward classification accuracy for assessment (Hendrycks et al., 2021b). Also, prior studies (Kadavath et al., 2022) underscore that LLMs demonstrate reliable calibration on multiple-choice scenarios. Therefore, we also used the multi-choice questions as a simple but good proxy to evaluate the abilities of LLMs.

For data annotation, we hired 10 undergraduate students, all with good English proficiency. We asked the students to rewrite the original claims into questions without distorting factuality while providing factuality labels for the questions. By transforming declarative statements into questions, using a Question-Answering approach can more effectively elicit factual knowledge from LLMs (Kadavath et al., 2022; Lin et al., 2022), and we also illustrate through experiments in Sec. 4.2. Note that claims in the original datasets are usually labeled based on given evidence, e.g. evidence supports or refutes the claim, but in Pinocchio, we only need to judge the factuality of the question. So we use unified labels: Yes, No, Not Sure Enough. The three labels correspond respectively to Factual, Non-Factual, and Not Enough Information for factual questions. Considering that all fact-checking datasets use a three-label system (Guo et al., 2022), we did not modify the number of labels to maintain consistency in labeling. When dealing with factuality questions in low-resource languages, for Chinese, the 5 undergraduate students we hired are native Chinese speakers. For other low-resource languages, we first use Google Translate to translate them into English and generate factuality questions, then translate the English questions back to the corresponding languages. The label distribution is shown in Table 1. We paid the annotators accordingly based on the quantity and quality of the annotations.

We ensure the quality of the annotated factuality questions in two ways. The two authors of this paper served as meta-reviewers, sampling 10 questions from each of the three categories across the seven domains in Pinocchio. The meta-reviewers judged if the factuality labels were correct. For the 210 factuality questions, the average label accuracy was 92.4%. We divided the 10 students into two groups and had each group re-annotate a random 200 questions annotated by the other group, then calculated inter-annotator agreement (IAA). The final IAA was 85.6%. Based on meta-reviewer results and IAA, the factuality labels in Pinocchio are of good quality.

3 METHODOLOGY

3.1 MODELS

To give a comprehensive view of the status of LLMs in a factual context, we evaluate 10 accessible LLMs, undergone different training stages including pretraining, instruction tuning, and reinforcement learning from human feedback (Ouyang et al., 2022), covering diverse organizations and varying in size. A detailed description can be found in Appendix A.2.

3.2 PROMPT STRATEGY

As illustrated in Figure 2, we employ 4 types of prompts to elicit desired responses from LLMs, namely: Zero-shot, Zero-shot with CoT (Kojima et al., 2022), Few-shot, and Few-shot with CoT (Wei et al., 2022). Specifically, we begin by providing the model with task instruction, denoted as Z: "You

Table 2: Results obtained using different forms of prompts on 10 accessible LLMs.

Methods	Zero-shot w/o CoT		Zero-shot w/ CoT		Few-shot w/o CoT		Few-shot w/ CoT		Overall Performance	
	Accuracy	F1	Accuracy	F1	Accuracy	F1	Accuracy	F1	Accuracy	F1
OPT-6.7B	—	—	—	—	36.9	27.9	37.9	28.5	18.8	14.3
BLOOM-7B	29.7	26.2	14.8	18.1	29.7	28.1	6.6	12.2	20.2	21.2
LLaMA-7B	31.8	29.6	22.3	24.9	36.8	28.6	35.3	31.4	31.6	28.6
Alpaca-7B	40.2	23.7	33.7	24.4	37.9	24.9	39.4	26.2	37.8	24.8
Vicuna-7B	33.2	33.6	34.2	32.9	35.5	34.8	48.5	40.6	37.9	34.9
Vicuna-13B	42.6	35.6	44.0	36.9	47.0	38.6	47.0	42.5	45.2	38.4
ChatGLM-6B	37.4	31.0	36.5	31.7	41.6	37.9	42.9	37.5	39.6	34.5
Flan-T5-11B	24.6	21.5	29.9	29.3	25.9	23.7	38.4	38.4	29.7	26.9
Text-Davinci-002	45.2	36.2	45.7	37.3	46.6	40.4	46.2	42.5	45.9	39.1
Text-Davinci-003	42.8	41.4	43.1	42.1	**48.8**	43.2	46.9	43.4	45.5	42.5
GPT-3.5-Turbo	**46.9**	**44.3**	**46.8**	**44.4**	47.2	**44.7**	47.1	45.7	**47.0**	**44.8**

will be given a question. You should answer whether it is Yes, No, or Not Sure Enough and show your evidence". This instruction informs the LLMs about the expected input and output. Subsequently, for any given input Q, we anticipate obtaining an output label Y from the LLMs f: $Y = f(Q, Z)$.

Zero-Shot Prompt In the zero-shot setting, the LLMs are expected to provide answers based on the Question Q and the task instruction Z. We anticipate that the LLMs can directly generate the factual answer "No" when presented with Q: "Has gas prices gone up 99 percent since Obama became president, making it the highest gas price increase since Carter?" The zero-shot with CoT setting extends the question Q by adding a two-stage prompt (Kojima et al., 2022): "Let's think step by step", designed to encourage the LLMs to contemplate the process of determining the factual label Y.

Few-Shot Prompt In the few-shot setting, we employ three shots for model input (Q). Detailed examples of the prompts in Figure 2 are presented in Appendix A.4. In the few-shot with CoT setting, we provide potential reasoning instructions to the LLMs before presenting the factual label (Y). As shown in Figure 2, for the Q: "Is there a capital called Mogadish?" Our reasoning approach entails first explaining the noun phrase in the Q (the subject and object), and subsequently elaborating on modifying phrases such as predicates or adjectives. Regarding the subject "Mogadish", we begin by furnishing a detailed definition: "Mogadishu is a city in East Africa, specifically in Somalia." Following this, we proceed to reason about the relation between "Mogadish" and "capital": "Furthermore, the capital of Somalia is indeed Mogadishu." Consequently, we arrive at the ultimate factual label: "Therefore, the answer is Yes."

4 EXPERIMENTS

In an effort to take the initial step in understanding the capabilities of LLMs, we undertake a comprehensive analysis of various LLMs on Pinocchio, under different conditions and tasks.

4.1 MAIN RESULTS

In Table 2, we present the average results of 10 accessible LLMs operating under varying settings on Pinocchio, run three times each. From Table 2, we draw the following conclusions:

- Regarding overall performance, we observe that, on average, LLMs without instruction tuning underperform those with instruction tuning by 16.0%. GPT family LLMs undergoing RLHF exhibit superior results, indicating that instruction tuning and RLHF optimize alignment with human knowledge, thereby improving factual question response accuracy.
- Results obtained using the Few-shot setting significantly outperform those obtained when simply asking factual questions to LLMs in the Zero-shot setting, especially for models without RLHF, exhibiting an average improvement of 7.3%. This highlights the capability of some sample prompts to better extract the inherent factual knowledge of LLMs.
- Using the CoT method, we observed a relative boost in performance in LLMs subjected to instruction tuning and RLHF, improving by an average of 2.1%. Notably, the factual accuracy of LLMs like OPT, BLOOM, and LLaMA was mostly stable or even decreased. A review of outputs from these untuned LLMs revealed that, post-CoT application, LLMs tend to produce related

Table 3: Results of different LLMs using Few-shot w/ CoT prompts across different tasks.

Task	Multifaceted		Structural		Adversarial		Temporal		Real-World		Domain Specific		Multi-lingual	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
OPT-6.7B	34.5	24.1	45.5	30.9	51.8	51.7	30.0	18.0	**53.7**	27.5	28.2	28.3	16.2	17.7
BLOOM-7B	10.7	13.5	0.8	3.5	2.0	3.7	3.7	7.7	5.4	8.5	11.8	15.6	9.8	15.9
LLaMA-7B	38.3	33.9	44.1	32.1	43.2	46.1	41.6	30.0	26.4	26.3	23.6	25.0	27.8	27.7
Alpaca-7B	38.6	28.8	48.0	23.6	46.4	35.1	49.6	26.1	24.5	19.9	42.9	26.8	24.2	17.7
Vicuna-7B	44.2	36.0	49.7	36.3	59.0	59.2	**50.1**	37.6	49.0	41.8	44.3	38.6	**46.7**	43.1
Vicuna-13B	49.9	45.3	48.1	37.9	58.9	60.0	45.4	37.8	47.7	42.7	43.5	40.4	37.8	37.9
ChatGLM-6B	41.0	36.0	46.8	35.7	51.5	48.6	39.4	32.4	48.9	34.8	35.2	35.0	37.1	35.3
Flan-T5-11B	49.2	49.4	43.5	33.7	54.7	56.6	31.6	30.6	31.1	29.4	35.6	34.6	25.3	14.4
Text-Davinci-002	47.7	47.7	**50.8**	38.4	64.2	64.3	33.9	31.1	51.7	41.4	36.4	36.1	43.1	39.5
Text-Davinci-003	51.1	47.8	44.3	33.7	64.1	63.7	41.4	35.1	48.0	42.8	40.4	41.4	43.7	**43.6**
GPT-3.5-Turbo	53.6	**53.1**	44.8	37.8	67.4	67.4	37.4	33.9	50.4	43.1	38.7	40.3	41.3	41.1

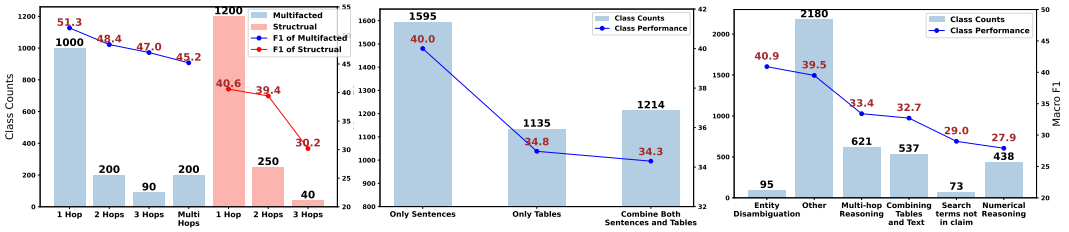
content considerations, and extensive considerations often overshadow factual discernment tasks, causing incorrect factual label outputs. In contrast, for instruction-tuned LLMs, the CoT method facilitates enhanced exploration of factual entity relations in questions, resulting in accurate factual labels. See Appendix A.5 for detailed case analyses.

- The OPT model, without being tuned to instructions, struggles significantly to output correct factual labels under the settings of Zero-shot and Zero-shot CoT, often resulting in either a repetition of the original question or a refusal to output any content at all. This issue is somewhat alleviated under the settings of Few-shot and Few-shot CoT.
- Additionally, we studied the hyperparameters of LLMs. Due to limited computing resources, we only explored Vicuna-7B and Vicuna-13B. We found that as model parameters increase, performance on factual questions improves correspondingly, with an average increase of 5.4%. This indicates that LLMs with more parameters can store more world knowledge and have stronger factual knowledge recognition capabilities.

In Table 3, we present the factual performance of LLMs in various tasks under the Few-shot CoT setting. This reveals the relative difficulty LLMs have in understanding and responding to factual questions in different tasks, providing insights for future training of factual knowledge in LLMs. From Table 3, it is observed that LLMs exhibit relatively poorer performance on factual questions related to the real-world, domain-specific knowledge, and multilingualism, being on average 6.4% lower compared to the other four tasks. This is attributed to the fact that the training data for LLMs typically come from general domains and are not up-to-date, which indirectly inspires the exploration of retrieval-augmented LLMs (Ram et al., 2023). We analyze the LLMs in different tasks in Sec. 4.2.

4.2 Analysis

In this section, we explore LLMs' capabilities focusing on key areas like handling of multi-hop factual questions, proficiency in diverse prompt strategies, and tackling challenges like numerical reasoning and entity ambiguity. We also examine their performance on time-sensitive factual questions, against adversarial attacks, with fine-grained labels and prompts in multiple languages.



(a) Multi-hop Reasoning Analysis (b) Structural Knowledge Analysis (c) Challenges of Different Questions

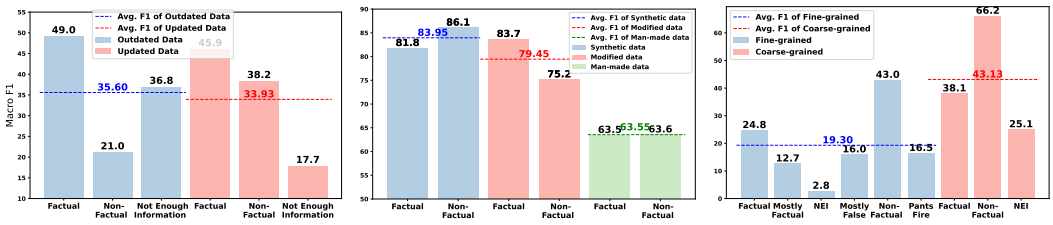
Figure 3: GPT-3.5-Turbo's outcomes across three distinct tasks under Few-shot CoT setting.

Multi-hop Factual Question Analysis To analyze the performance of LLMs when faced with factual questions based on multiple pieces of facts that require complex logical reasoning, we categorize multifaced and structural factual questions into distinct subsets, depending on the number of "hops" necessary to validate each factual question. To maintain fairness, we randomly sampled 1,490 data pieces from each of the two datasets for verification. Figure 3(a) illustrates the data

counts and Macro F1 scores of GPT-3.5-Turbo for each respective subset. The figure reveals a clear pattern: as the number of "hops" increases, the reasoning chain for deriving conclusions from existing factual knowledge extends, necessitating heightened logical reasoning capabilities from the LLMs. Consequently, the performance of the LLMs exhibits diminishing trends.

Structural Knowledge Analysis in LLMs To investigate whether LLMs can effectively memorize factual knowledge from structured data, we divided the structural task questions into three subsets according to evidence distribution: evidence in unstructured data (Only text), structured data (Only tables), or both (Combine text and tables). Figure 3(b) shows a notable decline (Avg. -5.5%) in GPT-3.5-Turbo's performance when evidence involves structured data, indicating LLMs' limited ability in extracting knowledge from structured tables. The LLMs also perform less effectively when handling questions requiring the combination of both evidence types, reflecting their incapacity to integrate diverse structured evidence effectively.

Analysis of Different Factual Questions Poses Challenges To assess the capabilities of LLMs in addressing various challenges, we partitioned each factual question within the structural task into six distinct challenges: 1) Entity disambiguation, 2) Other, 3) Multi-hop reasoning, 4) Combining tables and text, 5) Search terms not in claim, 6) Numerical reasoning, each centered around the most critical difficulty encountered during verification. Figure 3(c) illustrates GPT-3.5-Turbo's performance and data distribution across challenges. The extensive training and large-scale parameters enhance LLMs' performance in handling entity ambiguity. Longer reasoning chains and various forms of evidence challenge LLMs' factual abilities. When correct inference involves unmentioned entities, LLMs may lack necessary hints from factual questions, posing significant challenges. LLMs also exhibit deficiencies in precise numerical calculations due to the inherent hallucination phenomenon, resulting in subpar performance when numerical reasoning is needed for verification.



(a) Temporal Questions Verification (b) Adversarial Attacks Resilience (c) Label Granularity Variations

Figure 4: Results of GPT-3.5-Turbo in three different tasks under Few-shot CoT setting.

Temporal Analysis As time progresses, the factuality of questions may undergo changes. This task encompasses such data, and we leverage this task to explore the ability of LLMs to adapt to factual changes. Figure 4(a) illustrates that GPT-3.5-Turbo exhibits a modest yet noticeable performance difference when dealing with outdated data as compared to updated data. This discrepancy arises from the fact that LLMs are pretrained on a corpus of text prior to a specific temporal point. Consequently, LLMs lack the capability to acquire real-time, up-to-date knowledge, rendering them unable to validate questions that hinge on the most recent information for accurate assessments.

Adversarial Analysis To evaluate the robustness of LLMs to adversarial attacks, we divide the adversarial questions into three subsets: auto-generated questions from the corpus, manually modified synthesized questions yielding adversarial ones, and artificially created adversarial questions. Figure 3(b) presents the performance of GPT-3.5-Turbo on these three subsets. It is evident that following adversarial attacks, LLMs exhibit a substantial decrease in performance. Furthermore, factual questions that have undergone manual modifications or were artificially created prove to be more challenging compared to those that are automatically generated (Shen et al., 2023). This disparity could be attributed to the fact that automatically synthesized factual questions often contain explicit positive or negative words that hint at the outcome, and the exceptional comprehension abilities of LLMs enable them to accurately discern and provide the correct response in such cases.

Label Granularity Analysis To assess the effect of different label granularities on LLMs' performance, we conducted a manual re-labeling of the real-world task questions. Per the settings of Misra (2022), besides labeling as "Factual", "Non-Factual", and "Not Enough Information", we also require them to annotate the dataset with six factual labels: "Factual", "Mostly Factual", "Mostly

False", "Non-Factual", "Pants-Fire", and "Not Enough Information". We also modified the prompt for GPT-3.5-Turbo for more intricate factual responses to test its competency with nuanced labels. Results in Figure 4(c) disclosed: 1) The results show that, in general, there is a significant decrease in performance (-23.83%) when transitioning from coarse-grained justification to fine-grained justification. With finer granularity, LLMs are not only required to assess the authenticity of each question but also to judiciously employ their knowledge base to precisely gauge the credibility of each factual questions. 2) When comparing the performance of coarse-grained labels with fine-grained labels, we observe significant drops in the three categories: "Factual" by 13.3%, "Non-Factual" by 23.2%, and "Not Enough Information" by 22.3%. This indicates that finer-grained labels introduce additional options that can potentially disrupt the original judgment of the LLMs. A potential remedy could be the aggregation of multiple judgments through voting (Wang et al., 2023a).

Multilingual Task with Chinese and English Prompts To investigate the influence of prompts in different languages on LLMs, we extracted Chinese factual questions from the multilingual tasks to create a subset. We then evaluated the LLMs' performance when using both Chinese and English prompts, both of which are depicted in Appendix A.4. Table 4 illustrates the results, indicating that the LLMs perform better when using a Chinese prompt. This underscores the notion that employing prompts in the same language as the questions can enhance the transfer capabilities from English factual knowledge to other languages of LLMs.

Language	English	Chinese
Factual	41.7	**55.5**
Non-Factual	47.9	**49.7**
NEI	**43.8**	35.5
Overall	44.5	**46.9**

Table 4: Macro F1 over Chinese and English prompts.

Table 5: Results in different domains obtained on the Pinocchio-Lite using different prompts.

Task	Multifaceted		Structural		Adversarial		Temporal		Real-World		Domain Specific		Multi-lingual		Overall	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
1 shot	56.0	50.9	37.0	35.7	50.5	56.6	39.5	39.5	43.0	42.7	40.0	40.1	42.0	38.7	44.0	43.7
2 shots	56.0	**53.4**	41.0	42.3	47.5	56.2	41.0	42.0	40.5	41.7	**42.5**	**43.5**	36.5	34.8	43.6	43.7
3 shot	54.5	50.0	38.0	36.8	49.0	54.9	40.0	39.0	39.5	38.1	41.5	41.7	40.5	39.2	43.3	43.9
6 shots	54.5	51.7	38.5	38.3	49.0	55.8	42.0	41.5	42.5	41.6	39.0	39.5	41.0	38.4	43.8	43.8
9 shots	**57.5**	53.3	38.0	37.8	52.0	57.3	43.0	42.2	42.5	39.8	37.5	36.7	37.5	35.0	44.0	44.0
12 shots	55.5	52.0	38.5	38.6	53.0	58.8	**47.0**	**46.9**	**46.0**	**44.7**	34.0	34.5	39.0	37.1	44.7	44.8
Complex CoT	51.0	50.2	38.5	35.0	37.5	47.2	39.0	39.0	39.5	36.8	36.0	35.7	38.5	31.7	40.0	39.7
Self-Consistency	55.5	51.2	43.0	42.6	49.5	54.8	43.0	41.6	43.0	41.9	42.0	42.4	39.5	36.8	45.1	45.0
Self-Refinement	55.0	52.1	**44.5**	**44.0**	**53.5**	**59.2**	42.5	42.2	41.5	40.3	42.0	43.4	**43.0**	**39.9**	**46.0**	**46.2**
Declarative Claim	52.0	51.1	39.0	35.1	45.5	49.3	40.5	40.7	40.0	37.9	41.0	40.6	38.5	36.3	42.3	41.6

Prompt Strategy Analysis In prior research, various CoT methods have been employed to enhance the performance of LLMs. These methods include 1) augmenting the number of in-context learning examples, 2) implementing self-consistency mechanisms, which alleviates the hallucination phenomenon through majority voting after multiple judgments of LLMs (Wang et al., 2023a), 3) incorporating complex instances as demos to steer the cognitive processes of LLMs (Fu et al., 2022), and 4) employing self-refinement strategies, which refines LLMs' answers through continuous feedback of another LLM on responses to achieve better results (Madaan et al., 2023) and so forth. Additionally, we examined the influence of utilizing declarative claims as instances of in-context learning. We randomly sampled 200 factual questions from each task of the Pinocchio, totaling 1400 questions, to compose Pinocchio-Lite with the aim of speeding up the testing of different prompt strategies. The performance results of various CoT methods are presented in Table 5. To maintain fairness, three in-context learning examples are employed in the complex CoT, self-consistency, self-refinement, and declarative claim methods. Different types of CoT prompts are shown in Appendix A.4.

It is worth noting that 1) when the number of in-context learning examples is limited, the incremental improvement in performance is marginal upon increasing the number of examples. However, beyond a specific threshold, the addition of more examples gains more performance improvement. This could be due to the inability of LLMs to fully encapsulate the correct reasoning with fewer examples. 2) Concurrently, a fascinating observation is that the LLM's performance substantially deteriorates as the complexity of the CoT increases. This could stem from the difficulty LLMs have in extracting a generalized reasoning pattern from complex, multi-stage thinking processes with limited examples. 3) The self-consistency method markedly boosts performance by mitigating the hallucination issue in LLMs through consistency voting, enhancing their response accuracy. 4) In the self-refinement approach, the model might initially provide an incorrect response, but it can amend its mistakes through feedback and refine its answers. In the end, when no additional refinement is needed, the model often reaches the correct conclusion, achieving optimal performance. 5) Compared to the 3

shots method, the declarative claims method saw a 2.3% performance drop, illustrating that using questions as inputs better elicits factual knowledge than the original claim in the datasets.

5 Related Work

Factual Knowledge in Language Models Previous research shows that LLMs can retain and utilize factual knowledge, effectively acting as knowledge bases (Petroni et al., 2019; 2020; Heinzerling & Inui, 2021). This acquired factual knowledge in language models during pretraining can be advantageous for knowledge-intensive tasks like question answering and fact checking (Roberts et al., 2020; Yu et al., 2023a; Pan et al., 2023). To evaluate the factual knowledge stored in language models, Petroni et al. (2019) employed cloze tests consisting of triples and prompts specifically designed to simulate missing objects. Jiang et al. (2020b) explored the role of prompts in retrieving factual information from language models and devised improved prompts for probing. However, Elazar et al. (2021) demonstrated the unreliability of rank-based probing methods with paraphrased context, leading to inconsistent findings. Cao et al. (2021b) contended that biased prompts and leakage of golden answers often lead to overestimations of LLMs' knowledge storage capability. Our method is more in line with Kadavath et al. (2022) and Lin et al. (2022), employing self-evaluation by querying the models to assess response accuracy regarding factual knowledge.

More recent studies have directed their focus towards the detection of hallucinations—factually incorrect statements—in the responses generated by LLMs. For instance, the SelfCheckGPT (Manakul et al., 2023) uses a sampling method to detect inconsistencies in LLM responses, identifying hallucinated claims. Alternatively, FactScore (Min et al., 2023) approaches the challenge by deconstructing generations into atomic facts—concise statements—and assigning binary labels to assess their veracity. Furthermore, Chern et al. (2023) introduced a tool-enhanced framework for hallucination detection encompassing five core components: claim extraction, query formulation, tool-based querying, evidence gathering, and validation of consistency. However, these contributions primarily target the identification of factual inaccuracies in the models' output. In contrast, our benchmark is primarily designed to evaluate the breadth and depth of factual knowledge within LLMs.

Benchmarks for Large Language Models The advent of LLMs has underscored the importance of exhaustive benchmarks for effective capability assessment. Presently, there are predominantly two types of existing benchmarks. One evaluates the general knowledge and reasoning capacities of LLMs, exemplified by the MMLU (Hendrycks et al., 2021a), a multi-choice benchmark that measures tasks from real-world tests and literature, spanning diverse subjects like elementary math, US history, computer science, and law. Moreover, benchmarks also exist for non-English languages (Huang et al., 2023) or in a bilingual context (Zhong et al., 2023). BIG-bench (Srivastava et al., 2022) is a collaborative benchmark examining LLMs' capabilities across 204 diverse tasks from various fields like linguistics, childhood development, software development, and more. HELM (Liang et al., 2022) employs 7 metrics over 42 tasks to assess LLMs, focusing on aspects from accuracy to robustness. Specific benchmarks like GSM8K (Cobbe et al., 2021) and MATH (Hendrycks et al., 2021a) target mathematical problem-solving, presenting elementary to competition-level problems. In program synthesis, HumanEval (Chen et al., 2021a) and MBPP (Austin et al., 2021) evaluate functional correctness through program synthesis from docstrings. Additional benchmarks address instruction following (Dubois et al., 2023), tool usage (Xu et al., 2023), and decision making (Liu et al., 2023). Our benchmark mainly focuses on factual knowledge, differing from ones like TruthfulQA (Lin et al., 2022), which specifically tests truthfulness in LLMs' generated responses, with questions structured to provoke imitative falsehoods over truthful answers.

6 Conclusion

In this work, our primary focus is the development of the Pinocchio benchmark, an extensive test bed encompassing 20,713 questions across seven varying complexity tasks, as a tool to investigate whether LLMs are capable of memorizing factual knowledge and reasoning on the basis of it. Upon applying the Pinocchio benchmark, we observe that various types of LLMs using different prompting strategies such as self-refine and self-consistency still have challenges in optimal performance on factual tasks. It is our hope that this novel benchmark will shed light on this area and act as a foundation for further improvements in LLMs' factual knowledge and reasoning abilities.

ACKNOWLEDGEMENT

This work is supported in part by NSF under grant III-2106758. Additionally, Junzhe Chen and Xiaochuan Li are supported by Beijing Natural Science Foundation under grant number QY23115 and QY23116.

REFERENCES

Firoj Alam, Shaden Shaar, Fahim Dalvi, Hassan Sajjad, Alex Nikolov, Hamdy Mubarak, Giovanni Da San Martino, Ahmed Abdelali, Nadir Durrani, Kareem Darwish, Abdulaziz Al-Homaid, Wajdi Zaghouani, Tommaso Caselli, Gijs Danoe, Friso Stolk, Britt Bruntink, and Preslav Nakov. Fighting the COVID-19 infodemic: Modeling the perspective of journalists, fact-checkers, social media platforms, policy makers, and the society. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 611–649, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.56. URL https://aclanthology.org/2021.findings-emnlp.56.

Rami Aly, Zhijiang Guo, Michael Sejr Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. FEVEROUS: fact extraction and verification over unstructured and structured information. In Joaquin Vanschoren and Sai-Kit Yeung (eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/68d30a9594728bc39aa24be94b319d21-Abstract-round1.html.

Fatma Arslan, Naeemul Hassan, Chengkai Li, and Mark Tremayne. A benchmark dataset of check-worthy factual claims. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pp. 821–829, 2020.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. On the cross-lingual transferability of monolingual representations. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pp. 4623–4637. Association for Computational Linguistics, 2020. doi: 10.18653/V1/2020.ACL-MAIN.421. URL https://doi.org/10.18653/v1/2020.acl-main.421.

Akari Asai and Eunsol Choi. Challenges in information-seeking QA: unanswerable questions and paragraph retrieval. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *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*, pp. 1492–1504. Association for Computational Linguistics, 2021. doi: 10.18653/V1/2021.ACL-LONG.118. URL https://doi.org/10.18653/v1/2021.acl-long.118.

Akari Asai, Akiko Eriguchi, Kazuma Hashimoto, and Yoshimasa Tsuruoka. Multilingual extractive reading comprehension by runtime machine translation. *CoRR*, abs/1809.03275, 2018. URL http://arxiv.org/abs/1809.03275.

Akari Asai, Jungo Kasai, Jonathan H. Clark, Kenton Lee, Eunsol Choi, and Hannaneh Hajishirzi. XOR QA: cross-lingual open-retrieval question answering. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tür, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pp. 547–564. Association for Computational Linguistics, 2021. doi: 10.18653/V1/2021.NAACL-MAIN.46. URL https://doi.org/10.18653/v1/2021.naacl-main.46.

Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. Multifc: A real-world multi-domain dataset for evidence-based fact checking of claims. *arXiv preprint arXiv:1909.03242*, 2019.

Jacob Austin, Augustus Odena, Maxwell I. Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie J. Cai, Michael Terry, Quoc V. Le, and Charles Sutton. Program synthesis with large language models. *CoRR*, abs/2108.07732, 2021. URL https://arxiv.org/abs/2108.07732.

Lucas Azevedo, Mathieu D'aquin, Brian Davis, and Manel Zarrouk. Lux (linguistic aspects under examination): Discourse analysis for automatic fake news classification. In *The Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021)*, pp. 41–56. Association for Computational Linguistics, 2021.

Yonatan Belinkov and James R. Glass. Analysis methods in neural language processing: A survey. *Trans. Assoc. Comput. Linguistics*, 7:49–72, 2019. doi: 10.1162/tacl_a_00254. URL https://doi.org/10.1162/tacl_a_00254.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. Semantic parsing on freebase from question-answer pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pp. 1533–1544. ACL, 2013. URL https://aclanthology.org/D13-1160/.

Chandra Sekhar Bhagavatula, Thanapon Noraset, and Doug Downey. Methods for exploring and mining tables on wikipedia. In Duen Horng Chau, Jilles Vreeken, Matthijs van Leeuwen, and Christos Faloutsos (eds.), *Proceedings of the ACM SIGKDD Workshop on Interactive Data Exploration and Analytics, IDEA@KDD 2013, Chicago, Illinois, USA, August 11, 2013*, pp. 18–26. ACM, 2013. doi: 10.1145/2501511.2501516. URL https://doi.org/10.1145/2501511.2501516.

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. Language models are few-shot learners. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

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. Sparks of artificial general intelligence: Early experiments with GPT-4. *CoRR*, abs/2303.12712, 2023. doi: 10.48550/arXiv.2303.12712. URL https://doi.org/10.48550/arXiv.2303.12712.

Boxi Cao, Hongyu Lin, Xianpei Han, Le Sun, Lingyong Yan, Meng Liao, Tong Xue, and Jin Xu. Knowledgeable or educated guess? revisiting language models as knowledge bases. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *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*, pp. 1860–1874. Association for Computational Linguistics, 2021a. doi: 10.18653/v1/2021.acl-long.146. URL https://doi.org/10.18653/v1/2021.acl-long.146.

Nicola De Cao, Wilker Aziz, and Ivan Titov. Editing factual knowledge in language models. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pp. 6491–6506. Association for Computational Linguistics, 2021b. doi: 10.18653/v1/2021.emnlp-main.522. URL https://doi.org/10.18653/v1/2021.emnlp-main.522.

Jifan Chen and Greg Durrett. Understanding dataset design choices for multi-hop reasoning. *arXiv preprint arXiv:1904.12106*, 2019.

Jifan Chen, Aniruddh Sriram, Eunsol Choi, and Greg Durrett. Generating literal and implied subquestions to fact-check complex claims. *arXiv preprint arXiv:2205.06938*, 2022.

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. Evaluating large language models trained on code. *CoRR*, abs/2107.03374, 2021a. URL https://arxiv.org/abs/2107.03374.

Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. Tabfact: A large-scale dataset for table-based fact verification. *arXiv preprint arXiv:1909.02164*, 2019.

Wenhu Chen, Hanwen Zha, Zhiyu Chen, Wenhan Xiong, Hong Wang, and William Yang Wang. HybridQA: A dataset of multi-hop question answering over tabular and textual data. In Trevor Cohn, Yulan He, and Yang Liu (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1026–1036, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.91. URL https://aclanthology.org/2020.findings-emnlp.91.

Wenhu Chen, Ming-Wei Chang, Eva Schlinger, William Yang Wang, and William W. Cohen. Open question answering over tables and text. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021b. URL https://openreview.net/forum?id=MmCRswl1UYl.

Wenhu Chen, Xinyi Wang, and William Yang Wang. A dataset for answering time-sensitive questions. In Joaquin Vanschoren and Sai-Kit Yeung (eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*, 2021c. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/1f0e3dad99908345f7439f8ffabdffc4-Abstract-round2.html.

I-Chun Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, and Pengfei Liu. Factool: Factuality detection in generative AI - A tool augmented framework for multi-task and multi-domain scenarios. *CoRR*, abs/2307.13528, 2023. doi: 10.48550/ARXIV.2307.13528. URL https://doi.org/10.48550/arXiv.2307.13528.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. URL https://lmsys.org/blog/2023-03-30-vicuna/.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern,

Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways. *CoRR*, abs/2204.02311, 2022. doi: 10.48550/arXiv.2204.02311. URL https://doi.org/10.48550/arXiv.2204.02311.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models. *CoRR*, abs/2210.11416, 2022. doi: 10.48550/arXiv.2210.11416. URL https://doi.org/10.48550/arXiv.2210.11416.

Jonathan H. Clark, Jennimaria Palomaki, Vitaly Nikolaev, Eunsol Choi, Dan Garrette, Michael Collins, and Tom Kwiatkowski. Tydi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Trans. Assoc. Comput. Linguistics*, 8:454–470, 2020. doi: 10.1162/TACL_A_00317. URL https://doi.org/10.1162/tacl_a_00317.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the AI2 reasoning challenge. *CoRR*, abs/1803.05457, 2018a. URL http://arxiv.org/abs/1803.05457.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018b.

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. Training verifiers to solve math word problems. *CoRR*, abs/2110.14168, 2021. URL https://arxiv.org/abs/2110.14168.

Danilo Croce, Alexandra Zelenanska, and Roberto Basili. Enabling deep learning for large scale question answering in italian. *Intelligenza Artificiale*, 13(1):49–61, 2019. doi: 10.3233/IA-190018. URL https://doi.org/10.3233/IA-190018.

Jeff Da, Maxwell Forbes, Rowan Zellers, Anthony Zheng, Jena D Hwang, Antoine Bosselut, and Yejin Choi. Edited media understanding frames: Reasoning about the intent and implications of visual misinformation. 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)*, pp. 2026–2039, 2021.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. A dataset of information-seeking questions and answers anchored in research papers. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tür, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pp. 4599–4610. Association for Computational Linguistics, 2021. doi: 10.18653/V1/2021.NAACL-MAIN.365. URL https://doi.org/10.18653/v1/2021.naacl-main.365.

Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. Climate-fever: A dataset for verification of real-world climate claims, 2020.

Qingxiu Dong, Damai Dai, Yifan Song, Jingjing Xu, Zhifang Sui, and Lei Li. Calibrating factual knowledge in pretrained language models. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pp. 5937–5947. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.findings-emnlp.438. URL https://doi.org/10.18653/v1/2022.findings-emnlp.438.

Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Alpacafarm: A simulation framework for methods that learn from human feedback. *CoRR*, abs/2305.14387, 2023. doi: 10.48550/arXiv.2305.14387. URL https://doi.org/10.48550/arXiv.2305.14387.

Matthew Dunn, Levent Sagun, Mike Higgins, V. Ugur Güney, Volkan Cirik, and Kyunghyun Cho. Searchqa: A new q&a dataset augmented with context from a search engine. *CoRR*, abs/1704.05179, 2017. URL http://arxiv.org/abs/1704.05179.

Julian Eisenschlos, Bhuwan Dhingra, Jannis Bulian, Benjamin Börschinger, and Jordan L. Boyd-Graber. Fool me twice: Entailment from wikipedia gamification. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tür, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pp. 352–365. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.naacl-main.32. URL https://doi.org/10.18653/v1/2021.naacl-main.32.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard H. Hovy, Hinrich Schütze, and Yoav Goldberg. Measuring and improving consistency in pretrained language models. *Trans. Assoc. Comput. Linguistics*, 9:1012–1031, 2021. doi: 10.1162/tacl_a_00410. URL https://doi.org/10.1162/tacl_a_00410.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. ELI5: long form question answering. In Anna Korhonen, David R. Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pp. 3558–3567. Association for Computational Linguistics, 2019. doi: 10.18653/V1/P19-1346. URL https://doi.org/10.18653/v1/p19-1346.

Lorenzo Jaime Yu Flores and Yiding Hao. An adversarial benchmark for fake news detection models. *arXiv preprint arXiv:2201.00912*, 2022.

Yao Fu, Hao Peng, Ashish Sabharwal, Peter Clark, and Tushar Khot. Complexity-based prompting for multi-step reasoning. *arXiv preprint arXiv:2210.00720*, 2022.

Yi Fung, Christopher Thomas, Revanth Gangi Reddy, Sandeep Polisetty, Heng Ji, Shih-Fu Chang, Kathleen McKeown, Mohit Bansal, and Avirup Sil. Infosurgeon: Cross-media fine-grained information consistency checking for fake news detection. 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)*, pp. 1683–1698, 2021.

Haoyuan Gao, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, and Wei Xu. Are you talking to a machine? dataset and methods for multilingual image question answering. *CoRR*, abs/1505.05612, 2015. URL http://arxiv.org/abs/1505.05612.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. Did aristotle use a laptop? A question answering benchmark with implicit reasoning strategies. *Trans. Assoc. Comput. Linguistics*, 9:346–361, 2021. doi: 10.1162/TACL_A_00370. URL https://doi.org/10.1162/tacl_a_00370.

Genevieve Gorrell, Elena Kochkina, Maria Liakata, Ahmet Aker, Arkaitz Zubiaga, Kalina Bontcheva, and Leon Derczynski. Semeval-2019 task 7: Rumoureval 2019: Determining rumour veracity and support for rumours. In *Proceedings of the 13th International Workshop on Semantic Evaluation: NAACL HLT 2019*, pp. 845–854. Association for Computational Linguistics, 2019.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206, 2022. doi: 10.1162/tacl_a_00454. URL https://aclanthology.org/2022.tacl-1.11.

Ashim Gupta and Vivek Srikumar. X-fact: A new benchmark dataset for multilingual fact checking. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 675–682, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-short.86. URL https://aclanthology.org/2021.acl-short.86.

Deepak Gupta, Surabhi Kumari, Asif Ekbal, and Pushpak Bhattacharyya. MMQA: A multi-domain multi-lingual question-answering framework for english and hindi. In Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Kôiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga (eds.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018*. European Language Resources Association (ELRA), 2018. URL http://www.lrec-conf.org/proceedings/lrec2018/summaries/826.html.

Prakhar Gupta, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. Dialfact: A benchmark for fact-checking in dialogue. *arXiv preprint arXiv:2110.08222*, 2021.

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

Salim Hafid, Sebastian Schellhammer, Sandra Bringay, Konstantin Todorov, and Stefan Dietze. Scitweets-a dataset and annotation framework for detecting scientific online discourse. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pp. 3988–3992, 2022.

Momchil Hardalov, Todor Mihaylov, Dimitrina Zlatkova, Yoan Dinkov, Ivan Koychev, and Preslav Nakov. EXAMS: A multi-subject high school examinations dataset for cross-lingual and multilingual question answering. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pp. 5427–5444. Association for Computational Linguistics, 2020. doi: 10.18653/V1/2020.EMNLP-MAIN.438. URL https://doi.org/10.18653/v1/2020.emnlp-main.438.

Wei He, Kai Liu, Jing Liu, Yajuan Lyu, Shiqi Zhao, Xinyan Xiao, Yuan Liu, Yizhong Wang, Hua Wu, Qiaoqiao She, Xuan Liu, Tian Wu, and Haifeng Wang. Dureader: a chinese machine reading comprehension dataset from real-world applications. In Eunsol Choi, Minjoon Seo, Danqi Chen, Robin Jia, and Jonathan Berant (eds.), *Proceedings of the Workshop on Machine Reading for Question Answering@ACL 2018, Melbourne, Australia, July 19, 2018*, pp. 37–46. Association for Computational Linguistics, 2018. doi: 10.18653/V1/W18-2605. URL https://aclanthology.org/W18-2605/.

Benjamin Heinzerling and Kentaro Inui. Language models as knowledge bases: On entity representations, storage capacity, and paraphrased queries. In Paola Merlo, Jörg Tiedemann, and Reut Tsarfaty (eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pp. 1772–1791. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.eacl-main.153. URL https://doi.org/10.18653/v1/2021.eacl-main.153.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021a. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In Joaquin Vanschoren and Sai-Kit Yeung (eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*, 2021b. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/be83ab3ecd0db773eb2dc1b0a17836a1-Abstract-round2.html.

Jonathan Herzig, Thomas Müller, Syrine Krichene, and Julian Martin Eisenschlos. Open domain question answering over tables via dense retrieval. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tür, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pp. 512–519. Association for Computational Linguistics,

2021. doi: 10.18653/V1/2021.NAACL-MAIN.43. URL https://doi.org/10.18653/v1/2021.naacl-main.43.

Christopher Hidey, Tuhin Chakrabarty, Tariq Alhindi, Siddharth Varia, Kriste Krstovski, Mona Diab, and Smaranda Muresan. Deseption: Dual sequence prediction and adversarial examples for improved fact-checking. *arXiv preprint arXiv:2004.12864*, 2020.

Xuming Hu, Zhijiang Guo, GuanYu Wu, Aiwei Liu, Lijie Wen, and S Yu Philip. Chef: A pilot chinese dataset for evidence-based fact-checking. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3362–3376, 2022a.

Xuming Hu, Zhijiang Guo, GuanYu Wu, Aiwei Liu, Lijie Wen, and Philip Yu. CHEF: A pilot Chinese dataset for evidence-based fact-checking. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3362–3376, Seattle, United States, July 2022b. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.246. URL https://aclanthology.org/2022.naacl-main.246.

Xuming Hu, Zhijiang Guo, Junzhe Chen, Lijie Wen, and Philip S Yu. Mr2: A benchmark for multimodal retrieval-augmented rumor detection in social media. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2901–2912, 2023.

Kung-Hsiang Huang, Kathleen McKeown, Preslav Nakov, Yejin Choi, and Heng Ji. Faking fake news for real fake news detection: Propaganda-loaded training data generation. *arXiv preprint arXiv:2203.05386*, 2022.

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. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *CoRR*, abs/2305.08322, 2023. doi: 10.48550/arXiv.2305.08322. URL https://doi.org/10.48550/arXiv.2305.08322.

Zhen Jia, Abdalghani Abujabal, Rishiraj Saha Roy, Jannik Strötgen, and Gerhard Weikum. Tempquestions: A benchmark for temporal question answering. In Pierre-Antoine Champin, Fabien Gandon, Mounia Lalmas, and Panagiotis G. Ipeirotis (eds.), *Companion of the The Web Conference 2018 on The Web Conference 2018, WWW 2018, Lyon , France, April 23-27, 2018*, pp. 1057–1062. ACM, 2018. doi: 10.1145/3184558.3191536. URL https://doi.org/10.1145/3184558.3191536.

Yichen Jiang, Shikha Bordia, Zheng Zhong, Charles Dognin, Maneesh Singh, and Mohit Bansal. Hover: A dataset for many-hop fact extraction and claim verification. *arXiv preprint arXiv:2011.03088*, 2020a.

Zhengbao Jiang, Antonios Anastasopoulos, Jun Araki, Haibo Ding, and Graham Neubig. X-FACTR: multilingual factual knowledge retrieval from pretrained language models. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pp. 5943–5959. Association for Computational Linguistics, 2020b. doi: 10.18653/v1/2020.emnlp-main.479. URL https://doi.org/10.18653/v1/2020.emnlp-main.479.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. How can we know what language models know. *Trans. Assoc. Comput. Linguistics*, 8:423–438, 2020c. doi: 10.1162/tacl_a_00324. URL https://doi.org/10.1162/tacl_a_00324.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In Regina Barzilay and Min-Yen Kan (eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pp. 1601–1611. Association for Computational Linguistics, 2017. doi: 10.18653/V1/P17-1147. URL https://doi.org/10.18653/v1/P17-1147.

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. Language models (mostly) know what they know. *CoRR*, abs/2207.05221, 2022. doi: 10.48550/arXiv.2207.05221. URL https://doi.org/10.48550/arXiv.2207.05221.

Jungo Kasai, Keisuke Sakaguchi, Yoichi Takahashi, Ronan Le Bras, Akari Asai, Xinyan Yu, Dragomir R. Radev, Noah A. Smith, Yejin Choi, and Kentaro Inui. Realtime QA: what's the answer right now? *CoRR*, abs/2207.13332, 2022a. doi: 10.48550/ARXIV.2207.13332. URL https://doi.org/10.48550/arXiv.2207.13332.

Jungo Kasai, Keisuke Sakaguchi, Yoichi Takahashi, Ronan Le Bras, Akari Asai, Xinyan Yu, Dragomir R. Radev, Noah A. Smith, Yejin Choi, and Kentaro Inui. Realtime QA: what's the answer right now? *CoRR*, abs/2207.13332, 2022b. doi: 10.48550/ARXIV.2207.13332. URL https://doi.org/10.48550/arXiv.2207.13332.

Kashif Khan, Ruizhe Wang, and Pascal Poupart. Watclaimcheck: A new dataset for claim entailment and inference. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1293–1304, 2022.

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In Marilyn A. Walker, Heng Ji, and Amanda Stent (eds.), *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)*, pp. 252–262. Association for Computational Linguistics, 2018. doi: 10.18653/V1/N18-1023. URL https://doi.org/10.18653/v1/n18-1023.

Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. QASC: A dataset for question answering via sentence composition. 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*, pp. 8082–8090. AAAI Press, 2020. doi: 10.1609/AAAI.V34I05.6319. URL https://doi.org/10.1609/aaai.v34i05.6319.

Jiho Kim, Sungjin Park, Yeonsu Kwon, Yohan Jo, James Thorne, and Edward Choi. Factkg: Fact verification via reasoning on knowledge graphs. *arXiv preprint arXiv:2305.06590*, 2023a.

Najoung Kim, Ellie Pavlick, Burcu Karagol Ayan, and Deepak Ramachandran. Which linguist invented the lightbulb? presupposition verification for question-answering. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *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*, pp. 3932–3945. Association for Computational Linguistics, 2021. doi: 10.18653/V1/2021.ACL-LONG.304. URL https://doi.org/10.18653/v1/2021.acl-long.304.

Najoung Kim, Phu Mon Htut, Samuel R. Bowman, and Jackson Petty. $(QA)^2$: Question answering with questionable assumptions. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8466–8487, Toronto, Canada, July 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.472. URL https://aclanthology.org/2023.acl-long.472.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In *NeurIPS*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/8bb0d291acd4acf06ef112099c16f326-Abstract-Conference.html.

Neema Kotonya and Francesca Toni. Explainable automated fact-checking for public health claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7740–7754, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.623. URL https://aclanthology.org/2020.emnlp-main.623.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur P. Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: a benchmark for question answering research. *Trans. Assoc. Comput. Linguistics*, 7:452–466, 2019. doi: 10.1162/TACL_A_00276. URL https://doi.org/10.1162/tacl_a_00276.

Patrick S. H. Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. MLQA: evaluating cross-lingual extractive question answering. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pp. 7315–7330. Association for Computational Linguistics, 2020. doi: 10.18653/V1/2020.ACL-MAIN.653. URL https://doi.org/10.18653/v1/2020.acl-main.653.

Xiao Li, Yawei Sun, and Gong Cheng. Tsqa: Tabular scenario based question answering. *ArXiv*, abs/2101.11429, 2021. URL https://api.semanticscholar.org/CorpusID:231719096.

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. Holistic evaluation of language models. *CoRR*, abs/2211.09110, 2022. doi: 10.48550/arXiv.2211.09110. URL https://doi.org/10.48550/arXiv.2211.09110.

Anders Edelbo Lillie, Emil Refsgaard Middelboe, and Leon Derczynski. Joint rumour stance and veracity prediction. In *Nordic Conference of Computational Linguistics (2019)*, pp. 208–221. Linköping University Electronic Press, 2019.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*, 2021.

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words. *Trans. Mach. Learn. Res.*, 2022, 2022. URL https://openreview.net/forum?id=8s8K2UZGTZ.

Adam Liska, Tomás Kociský, Elena Gribovskaya, Tayfun Terzi, Eren Sezener, Devang Agrawal, Cyprien de Masson d'Autume, Tim Scholtes, Manzil Zaheer, Susannah Young, Ellen Gilsenan-McMahon, Sophia Austin, Phil Blunsom, and Angeliki Lazaridou. Streamingqa: A benchmark for adaptation to new knowledge over time in question answering models. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 13604–13622. PMLR, 2022. URL https://proceedings.mlr.press/v162/liska22a.html.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. Agentbench: Evaluating llms as agents. *arXiv preprint arXiv: 2308.03688*, 2023.

Shayne Longpre, Yi Lu, and Joachim Daiber. MKQA: A linguistically diverse benchmark for multilingual open domain question answering. *Trans. Assoc. Comput. Linguistics*, 9:1389–1406, 2021. doi: 10.1162/TACL_A_00433. URL https://doi.org/10.1162/tacl_a_00433.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *NeurIPS*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/11332b6b6cf4485b84afadb1352d3a9a-Abstract-Conference.html.

Grace Luo, Trevor Darrell, and Anna Rohrbach. Newsclippings: Automatic generation of out-of-context multimodal media. *arXiv preprint arXiv:2104.05893*, 2021.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *arXiv preprint arXiv:2303.17651*, 2023.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *CoRR*, abs/2303.08896, 2023. doi: 10.48550/ARXIV.2303.08896. URL https://doi.org/10.48550/arXiv.2303.08896.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 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 (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pp. 2381–2391. Association for Computational Linguistics, 2018. doi: 10.18653/V1/D18-1260. URL https://doi.org/10.18653/v1/d18-1260.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *CoRR*, abs/2305.14251, 2023. doi: 10.48550/ARXIV.2305.14251. URL https://doi.org/10.48550/arXiv.2305.14251.

Rishabh Misra. Kaggle politifact fact-checking dataset, 2022. URL https://www.kaggle.com/datasets/rmisra/politifact-fact-check-dataset.

Isabelle Mohr, Amelie Wührl, and Roman Klinger. Covert: A corpus of fact-checked biomedical covid-19 tweets. *arXiv preprint arXiv:2204.12164*, 2022.

Hussein Mozannar, Elie Maamary, Karl El Hajal, and Hazem M. Hajj. Neural arabic question answering. In Wassim El-Hajj, Lamia Hadrich Belguith, Fethi Bougares, Walid Magdy, and Imed Zitouni (eds.), *Proceedings of the Fourth Arabic Natural Language Processing Workshop, WANLP@ACL 2019, Florence, Italy, August 1, 2019*, pp. 108–118. Association for Computational Linguistics, 2019. doi: 10.18653/V1/W19-4612. URL https://doi.org/10.18653/v1/w19-4612.

Kai Nakamura, Sharon Levy, and William Yang Wang. r/fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. *arXiv preprint arXiv:1911.03854*, 2019.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. MS MARCO: A human generated machine reading comprehension dataset. In Tarek Richard Besold, Antoine Bordes, Artur S. d'Avila Garcez, and Greg Wayne (eds.), *Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, December 9, 2016*, volume 1773 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2016. URL https://ceur-ws.org/Vol-1773/CoCoNIPS_2016_paper9.pdf.

Dan S Nielsen and Ryan McConville. Mumin: A large-scale multilingual multimodal fact-checked misinformation social network dataset. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 3141–3153, 2022.

Jeppe Nørregaard and Leon Derczynski. Danfever: claim verification dataset for danish. In *Proceedings of the 23rd Nordic conference on computational linguistics (NoDaLiDa)*, pp. 422–428, 2021.

Yasumasa Onoe, Michael J. Q. Zhang, Eunsol Choi, and Greg Durrett. Entity cloze by date: What lms know about unseen entities. In Marine Carpuat, Marie-Catherine de Marneffe, and Iván Vladimir Meza Ruíz (eds.), *Findings of the Association for Computational Linguistics: NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pp. 693–702. Association for Computational Linguistics, 2022. doi: 10.18653/V1/2022.FINDINGS-NAACL.52. URL https://doi.org/10.18653/v1/2022.findings-naacl.52.

Yasumasa Onoe, Michael J. Q. Zhang, Shankar Padmanabhan, Greg Durrett, and Eunsol Choi. Can lms learn new entities from descriptions? challenges in propagating injected knowledge. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pp. 5469–5485. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.ACL-LONG.300. URL https://doi.org/10.18653/v1/2023.acl-long.300.

OpenAI. Chatgpt url, 2022. URL https://chat.openai.com.

OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023. doi: 10.48550/arXiv.2303.08774. URL https://doi.org/10.48550/arXiv.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. Training language models to follow instructions with human feedback. In *NeurIPS*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In Gerardo Flores, George H. Chen, Tom J. Pollard, Joyce C. Ho, and Tristan Naumann (eds.), *Conference on Health, Inference, and Learning, CHIL 2022, 7-8 April 2022, Virtual Event*, volume 174 of *Proceedings of Machine Learning Research*, pp. 248–260. PMLR, 2022. URL https://proceedings.mlr.press/v174/pal22a.html.

Liangming Pan, Xiaobao Wu, Xinyuan Lu, Anh Tuan Luu, William Yang Wang, Min-Yen Kan, and Preslav Nakov. Fact-checking complex claims with program-guided reasoning. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pp. 6981–7004. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.386. URL https://doi.org/10.18653/v1/2023.acl-long.386.

Chan Young Park, Julia Mendelsohn, Anjalie Field, and Yulia Tsvetkov. Challenges and opportunities in information manipulation detection: An examination of wartime russian media. *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 5209–5235, 2022.

Hao Peng, Xiaozhi Wang, Shengding Hu, Hailong Jin, Lei Hou, Juanzi Li, Zhiyuan Liu, and Qun Liu. COPEN: probing conceptual knowledge in pre-trained language models. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pp. 5015–5035. Association for Computational Linguistics, 2022. doi: 10.18653/V1/2022.EMNLP-MAIN.335. URL https://doi.org/10.18653/v1/2022.emnlp-main.335.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick S. H. Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander H. Miller. Language models as knowledge bases? In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *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*, pp. 2463–2473. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1250. URL https://doi.org/10.18653/v1/D19-1250.

Fabio Petroni, Patrick S. H. Lewis, Aleksandra Piktus, Tim Rocktäschel, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. How context affects language models' factual predictions. In Dipanjan Das, Hannaneh Hajishirzi, Andrew McCallum, and Sameer Singh (eds.), *Conference on Automated Knowledge Base Construction, AKBC 2020, Virtual, June 22-24, 2020*, 2020. doi: 10.24432/C5201W. URL https://doi.org/10.24432/C5201W.

Peng Qi, Yuyan Bu, Juan Cao, Wei Ji, Ruihao Shui, Junbin Xiao, Danding Wang, and Tat-Seng Chua. Fakesv: A multimodal benchmark with rich social context for fake news detection on short video platforms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 14444–14452, 2023.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. Is chatgpt a general-purpose natural language processing task solver? *CoRR*, abs/2302.06476, 2023. doi: 10.48550/arXiv.2302.06476. URL https://doi.org/10.48550/arXiv.2302.06476.

Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for squad. In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers*, pp. 784–789. Association for Computational Linguistics, 2018. doi: 10.18653/V1/P18-2124. URL https://aclanthology.org/P18-2124/.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. In-context retrieval-augmented language models. *arXiv preprint arXiv:2302.00083*, 2023.

Dongning Rao, Xin Miao, Zhihua Jiang, and Ran Li. Stanker: Stacking network based on level-grained attention-masked bert for rumor detection on social media. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 3347–3363, 2021.

Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. Model-agnostic interpretability of machine learning. *CoRR*, abs/1606.05386, 2016. URL http://arxiv.org/abs/1606.05386.

Adam Roberts, Colin Raffel, and Noam Shazeer. How much knowledge can you pack into the parameters of a language model? In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pp. 5418–5426. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.437. URL https://doi.org/10.18653/v1/2020.emnlp-main.437.

Arkadiy Saakyan, Tuhin Chakrabarty, and Smaranda Muresan. Covid-fact: Fact extraction and verification of real-world claims on covid-19 pandemic. *arXiv preprint arXiv:2106.03794*, 2021.

Ananya B. Sai, Akash Kumar Mohankumar, and Mitesh M. Khapra. A survey of evaluation metrics used for NLG systems. *ACM Comput. Surv.*, 55(2):26:1–26:39, 2023. doi: 10.1145/3485766. URL https://doi.org/10.1145/3485766.

Mourad Sarrouti, Asma Ben Abacha, Yassine M'rabet, and Dina Demner-Fushman. Evidence-based fact-checking of health-related claims. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 3499–3512, 2021.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilic, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, and et al. BLOOM: A 176b-parameter open-access multilingual language model. *CoRR*, abs/2211.05100, 2022a. doi: 10.48550/arXiv.2211.05100. URL https://doi.org/10.48550/arXiv.2211.05100.

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

Michael Sejr Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. Averitec: A dataset for real-world claim verification with evidence from the web. *CoRR*, abs/2305.13117, 2023. doi: 10.48550/ARXIV.2305.13117. URL https://doi.org/10.48550/arXiv.2305.13117.

Tal Schuster, Darsh J. Shah, Yun Jie Serene Yeo, Daniel Filizzola, Enrico Santus, and Regina Barzilay. Towards debiasing fact verification models. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *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*, pp. 3417–3423. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1341. URL https://doi.org/10.18653/v1/D19-1341.

Tal Schuster, Adam Fisch, and Regina Barzilay. Get your vitamin C! robust fact verification with contrastive evidence. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 624–643, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.52. URL https://aclanthology.org/2021.naacl-main.52.

Gautam Kishore Shahi and Durgesh Nandini. Fakecovid–a multilingual cross-domain fact check news dataset for covid-19. *arXiv preprint arXiv:2006.11343*, 2020.

ShareGPT. Sharegpt url, 2023. URL https://sharegpt.com/.

Xinyue Shen, Zeyuan Chen, Michael Backes, and Yang Zhang. In chatgpt we trust? measuring and characterizing the reliability of chatgpt. *CoRR*, abs/2304.08979, 2023. doi: 10.48550/arXiv.2304.08979. URL https://doi.org/10.48550/arXiv.2304.08979.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big data*, 8(3):171–188, 2020.

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. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *CoRR*, abs/2206.04615, 2022. doi: 10.48550/arXiv.2206.04615. URL https://doi.org/10.48550/arXiv.2206.04615.

StanfordCRFM. Alpaca url, 2023. URL https://crfm.stanford.edu/2023/03/13/alpaca.html.

Megha Sundriyal, Atharva Kulkarni, Vaibhav Pulastya, Md Shad Akhtar, and Tanmoy Chakraborty. Empowering the fact-checkers! automatic identification of claim spans on twitter. *arXiv preprint arXiv:2210.04710*, 2022.

Alon Talmor and Jonathan Berant. The web as a knowledge-base for answering complex questions. In Marilyn A. Walker, Heng Ji, and Amanda Stent (eds.), *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)*, pp. 641–651. Association for Computational Linguistics, 2018. doi: 10.18653/V1/N18-1059. URL https://doi.org/10.18653/v1/n18-1059.

Reuben Tan, Bryan A Plummer, and Kate Saenko. Detecting cross-modal inconsistency to defend against neural fake news. *arXiv preprint arXiv:2009.07698*, 2020.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a large-scale dataset for fact extraction and verification. In Marilyn A. Walker, Heng Ji, and Amanda Stent (eds.), *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)*, pp. 809–819. Association for Computational Linguistics, 2018. doi: 10.18653/v1/n18-1074. URL https://doi.org/10.18653/v1/n18-1074.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971, 2023a. doi: 10.48550/arXiv.2302.13971. URL https://doi.org/10.48550/arXiv.2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288, 2023b. doi: 10.48550/arXiv.2307.09288. URL https://doi.org/10.48550/arXiv.2307.09288.

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry W. Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc V. Le, and Thang Luong. Freshllms: Refreshing large language models with search engine augmentation. *CoRR*, abs/2310.03214, 2023. doi: 10.48550/ARXIV.2310.03214. URL https://doi.org/10.48550/arXiv.2310.03214.

David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Iz Beltagy, Lucy Lu Wang, and Hannaneh Hajishirzi. SciFact-open: Towards open-domain scientific claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 4719–4734, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.findings-emnlp.347.

Nancy XR Wang, Diwakar Mahajan, Marina Danilevsky, and Sara Rosenthal. Semeval-2021 task 9: Fact verification and evidence finding for tabular data in scientific documents (sem-tab-facts). *arXiv preprint arXiv:2105.13995*, 2021.

William Yang Wang. " liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*, 2017.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023a.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pp. 13484–13508. Association for Computational Linguistics, 2023b. doi: 10.18653/v1/2023.acl-long.754. URL https://doi.org/10.18653/v1/2023.acl-long.754.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.

Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. Constructing datasets for multi-hop reading comprehension across documents. *Trans. Assoc. Comput. Linguistics*, 6:287–302, 2018a. doi: 10.1162/TACL_A_00021. URL https://doi.org/10.1162/tacl_a_00021.

Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. Constructing datasets for multi-hop reading comprehension across documents. *Transactions of the Association for Computational Linguistics*, 6:287–302, 2018b.

Dustin Wright, Jiaxin Pei, David Jurgens, and Isabelle Augenstein. Modeling information change in science communication with semantically matched paraphrases. *arXiv preprint arXiv:2210.13001*, 2022.

Qiantong Xu, Fenglu Hong, Bo Li, Changran Hu, Zhengyu Chen, and Jian Zhang. On the tool manipulation capability of open-source large language models, 2023.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pp. 2369–2380. Association for Computational Linguistics, 2018. doi: 10.18653/V1/D18-1259. URL https://doi.org/10.18653/v1/d18-1259.

Wenhao Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, Soumya Sanyal, Chenguang Zhu, Michael Zeng, and Meng Jiang. Generate rather than retrieve: Large language models are strong context generators. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023a. URL https://openreview.net/pdf?id=fB0hRu9GZUS.

Xinyan Yu, Sewon Min, Luke Zettlemoyer, and Hannaneh Hajishirzi. CREPE: Open-domain question answering with false presuppositions. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10457–10480, Toronto, Canada, July 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.583. URL https://aclanthology.org/2023.acl-long.583.

Majid Zarharan, Mahsa Ghaderan, Amin Pourdabiri, Zahra Sayedi, Behrouz Minaei-Bidgoli, Sauleh Eetemadi, and Mohammad Taher Pilehvar. Parsfever: a dataset for farsi fact extraction and verification. In *Proceedings of* SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, pp. 99–104, 2021.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In Anna Korhonen, David R. Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pp. 4791–4800. Association for Computational Linguistics, 2019. doi: 10.18653/V1/P19-1472. URL https://doi.org/10.18653/v1/p19-1472.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Zhiyuan Liu, Peng Zhang, Yuxiao Dong, and Jie Tang. GLM-130B: an open bilingual pre-trained model. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/pdf?id=-Aw0rrrPUF.

Michael J. Q. Zhang and Eunsol Choi. Situatedqa: Incorporating extra-linguistic contexts into QA. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing,*

EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021, pp. 7371–7387. Association for Computational Linguistics, 2021. doi: 10.18653/V1/2021.EMNLP-MAIN.586. URL https://doi.org/10.18653/v1/2021.emnlp-main.586.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona T. Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. OPT: open pre-trained transformer language models. *CoRR*, abs/2205.01068, 2022. doi: 10.48550/arXiv.2205.01068. URL https://doi.org/10.48550/arXiv.2205.01068.

Wenxuan Zhang, Yang Deng, Jing Ma, and Wai Lam. Answerfact: Fact checking in product question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2407–2417, 2020.

Jonathan Zheng, Ashutosh Baheti, Tarek Naous, Wei Xu, and Alan Ritter. Stanceosaurus: Classifying stance towards multicultural misinformation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 2132–2151, 2022.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. Agieval: A human-centric benchmark for evaluating foundation models. *CoRR*, abs/2304.06364, 2023. doi: 10.48550/arXiv.2304.06364. URL https://doi.org/10.48550/arXiv.2304.06364.

Chen Zhu, Ankit Singh Rawat, Manzil Zaheer, Srinadh Bhojanapalli, Daliang Li, Felix X. Yu, and Sanjiv Kumar. Modifying memories in transformer models. *CoRR*, abs/2012.00363, 2020a. URL https://arxiv.org/abs/2012.00363.

Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. TAT-QA: A question answering benchmark on a hybrid of tabular and textual content in finance. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *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)*, pp. 3277–3287, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.254. URL https://aclanthology.org/2021.acl-long.254.

Ming Zhu, Aman Ahuja, Wei Wei, and Chandan K. Reddy. A hierarchical attention retrieval model for healthcare question answering. In Ling Liu, Ryen W. White, Amin Mantrach, Fabrizio Silvestri, Julian J. McAuley, Ricardo Baeza-Yates, and Leila Zia (eds.), *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, pp. 2472–2482. ACM, 2019. doi: 10.1145/3308558.3313699. URL https://doi.org/10.1145/3308558.3313699.

Ming Zhu, Aman Ahuja, Da-Cheng Juan, Wei Wei, and Chandan K. Reddy. Question answering with long multiple-span answers. In Trevor Cohn, Yulan He, and Yang Liu (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pp. 3840–3849. Association for Computational Linguistics, 2020b. doi: 10.18653/V1/2020.FINDINGS-EMNLP.342. URL https://doi.org/10.18653/v1/2020.findings-emnlp.342.

A Appendix

A.1 Ethical Statement

Pinocchio primarily serves to assess LLMs' responses to questions concerning factual knowledge. If a model performs effectively, it would be imprudent to infer that its reliability will uniformly translate to diverse task domains (even if some degree of transfer learning is anticipated). For instance, Pinocchio does not encompass long-form generation, such as news articles, or interactive settings, such as extended dialogues with adversarial entities. Furthermore, although the questions within Pinocchio parallel real-world inquiries, they originate not from a deployed system, thus posing a potential risk of over- or under-estimating the factuality of such a system.

We postulate that Pinocchio is unlikely to prove advantageous for those intending to fabricate deceptive models with malicious intent. To effectuate deception, a model must generate erroneous responses relatively infrequently, lest humans swiftly discern its unreliability. However, acquiring a low score on Pinocchio necessitates the provision of incorrect answers to virtually all questions. To be instrumental for malevolent purposes, a model must generate highly specific false statements, such as assertions concerning a maliciously targeted victim or a particular governmental policy. Yet, Pinocchio lacks coverage of highly specific subjects, offering instead a superficial overview of general factual topics.

While Wikipedia and some news websites are exemplary collaborative resources, they inherently contain inaccuracies and noise, akin to any encyclopedia or knowledge repository. Consequently, we advise users of Pinocchio against making absolute assertions about the validated claims and discourage its utilization for the development of truth-revealing models. We refrained from collecting participants' personal data in any form. Participants accessed our online tool exclusively using an identification number. Generated assertions must solely incorporate information deemed as general world knowledge or sourced from Wikipedia, thereby excluding any personally identifiable information or offensive content.

A.2 The detailed introduction to the LLMs

For pretraining models, OPT (Zhang et al., 2022) is an open-sourced large causal language model which perform similar in performance to GPT-3 (Brown et al., 2020). BLOOM (Scao et al., 2022a) is an open-access multilingual large language model that is suitable for non-English facts. LLaMA (Touvron et al., 2023a) is probably the best open-weight foundation model so far that achieves the highest accuracy on various English benchmarks (e.g. MMLU (Hendrycks et al., 2021a)) within open-weight models. For instruction-tuned models, Alpaca (StanfordCRFM, 2023) is fine-tuned from the LLaMA model on 52K self-instructed demonstrations (Wang et al., 2023b). Alpaca behaves qualitatively similarly to OpenAI's Text-Davinci-003 on evaluation of single-turn instruction following. Vicuna is an open-source chatbot trained by fine-tuning LLaMA on user-shared conversations collected from ShareGPT (ShareGPT, 2023). Flan -T5 (Chung et al., 2022) is an enhanced version of T5 that has been instruction fine-tuned in a mixture of tasks. ChatGLM is an open bilingual language model based on the General Language Model (Zeng et al., 2023). ChatGLM is trained on Chinese and English corpus, supplemented by instruction tuning, feedback bootstrap, and reinforcement learning with human feedback (RLHF; Ouyang et al. 2022). ChatGPT (OpenAI, 2022) from OpenAI that has undergone pretraining, instruction tuning, and RLHF. ChatGPT has been observed to have impressive capabilities in various aspects favoring reasoning capabilities (Qin et al., 2023).

A.3 Task Results

In this section, we present the results of all LLMs across different tasks under three different settings: Zero-shot w/o CoT, Zero-shot w/ CoT, and Few-shot w/o CoT.

Table 6: Results of different LLMs using Zero-shot w/o CoT prompts across different domains.

Task	Multifaceted		Structural		Adversarial		Temporal		Real-World		Domain Specific		Multi-lingual	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
OPT-6.7B	-	-	-	-	-	-	-	-	-	-	-	-	-	-
BLOOM-7B	21.9	17.8	24.9	17.9	32.4	36.3	17.6	14.2	52.1	23.8	30.1	29.9	29.0	30.4
LLaMA-7B	30.7	28.8	38.3	29.3	30.8	35.6	37.9	26.0	35.1	32.4	27.1	29.1	13.9	17.2
Alpaca-7B	34.8	21.6	47.9	23.7	47.7	35.7	**52.9**	26.8	28.1	19.0	43.1	24.2	26.4	19.5
Vicuna-7B	38.6	35.4	19.4	16.8	50.8	53.9	37.9	**42.0**	29.8	30.1	33.6	30.4	34.8	34.4
Vicuna-13B	45.0	41.1	43.9	31.0	57.1	56.7	45.9	33.7	32.0	29.0	43.1	32.3	37.3	34.7
ChatGLM-6B	30.6	30.3	45.6	30.8	42.9	46.4	28.0	24.1	45.9	31.9	34.1	30.2	32.9	28.5
Flan-T5-11B	39.2	29.6	11.2	10.2	56.2	49.9	12.9	10.5	17.4	10.6	28.8	16.5	25.4	14.7
Text-Davinci-002	44.7	38.4	**49.2**	**37.8**	57.2	56.1	36.2	27.8	**53.2**	32.7	31.3	30.1	42.2	32.5
Text-Davinci-003	50.9	48.9	36.4	29.5	58.7	57.9	51.7	36.6	40.4	37.0	41.3	33.3	42.7	43.1
GPT-3.5-Turbo	53.2	50.1	43.1	35.8	62.3	61.8	43.4	35.9	46.1	**42.1**	42.5	35.6	**45.0**	**45.7**

Table 7: Results of different LLMs using Zero-shot w/ CoT prompts across different domains.

Task	Multifaceted		Structural		Adversarial		Temporal		Real-World		Domain Specific		Multi-lingual	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
OPT-6.7B	-	-	-	-	-	-	-	-	-	-	-	-	-	-
BLOOM-7B	17.0	20.2	10.1	12.6	12.0	19.2	6.9	9.4	15.5	16.5	27.3	23.4	17.9	19.3
LLaMA-7B	20.3	23.5	29.5	26.4	18.3	26.2	25.7	26.3	22.9	24.9	20.0	23.0	12.2	16.9
Alpaca-7B	38.3	28.9	42.7	22.4	38.6	36.1	38.0	23.0	29.7	23.1	28.5	21.7	13.5	15.2
Vicuna-7B	29.4	35.8	45.7	31.6	4.4	8.3	**49.0**	36.6	15.1	19.6	**47.4**	**39.6**	37.9	33.9
Vicuna-13B	46.7	42.8	46.2	32.7	58.8	58.6	47.3	34.6	34.1	31.1	43.6	33.6	36.0	33.2
ChatGLM-6B	34.0	33.0	40.5	29.8	46.3	46.6	27.3	24.7	44.9	30.7	32.2	30.1	30.2	30.4
Flan-T5-11B	49.6	49.1	19.2	16.8	58.2	58.2	21.7	21.8	20.4	17.1	30.3	20.8	25.8	15.6
Text-Davinci-002	47.2	40.1	**51.7**	**38.0**	59.9	58.2	37.2	30.8	**52.7**	34.4	29.9	30.3	42.5	36.6
Text-Davinci-003	52.7	51.1	37.5	31.3	61.0	59.5	40.8	36.7	38.8	36.2	41.4	33.0	42.2	42.4
GPT-3.5-Turbo	53.3	52.1	43.1	35.5	59.8	61.6	42.2	**37.7**	44.8	**43.3**	41.4	36.0	**43.4**	**45.3**

Table 8: Results of different LLMs using Few-shot w/o CoT prompts across different domains.

Task	Multifaceted		Structural		Adversarial		Temporal		Real-World		Domain Specific		Multi-lingual	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
OPT-6.7B	38.1	30.1	45.9	27.1	46.8	32.4	28.7	20.0	51.1	25.5	37.0	29.6	-	-
BLOOM-7B	32.7	22.5	8.8	9.0	43.5	32.6	23.8	21.1	**53.3**	31.4	29.3	28.4	22.3	19.3
LLaMA-7B	34.8	21.9	40.5	27.0	47.4	38.4	45.5	26.9	22.4	22.0	39.3	34.3	32.6	27.0
Alpaca-7B	34.9	25.4	48.0	22.6	43.4	32.5	48.0	25.8	24.0	19.4	42.6	27.0	21.8	17.4
Vicuna-7B	34.5	27.6	40.1	25.4	54.5	53.3	30.1	26.6	36.1	34.0	33.9	27.7	22.8	20.5
Vicuna-13B	47.9	42.5	48.9	31.4	54.7	53.1	**53.4**	**38.6**	39.7	35.2	47.4	34.9	37.7	36.8
ChatGLM-6B	37.9	32.9	44.6	35.4	52.2	46.8	44.9	35.4	38.0	33.9	41.6	38.0	34.5	33.8
Flan-T5-11B	42.3	35.0	12.4	11.7	57.7	53.6	15.1	13.0	17.7	11.4	29.7	19.4	24.9	13.6
Text-Davinci-002	45.4	41.2	**51.4**	**38.4**	61.7	61.8	37.0	31.3	52.0	38.6	33.0	32.6	42.5	40.0
Text-Davinci-003	**59.6**	43.4	48.1	33.7	62.0	**61.8**	46.4	36.3	50.6	43.0	41.7	36.3	**44.2**	**44.4**
GPT-3.5-Turbo	52.1	48.4	42.5	35.4	61.2	61.1	43.7	36.2	48.9	**43.2**	42.0	35.6	42.8	43.0

A.4 PROMPT STRATEGY

In this section, we provide the comprehensive versions of all the prompts utilized in both the main experiments and the subsequent analysis. We engaged native Chinese annotators to rephrase the English prompts while maintaining their semantic integrity, thus yielding Chinese prompts.

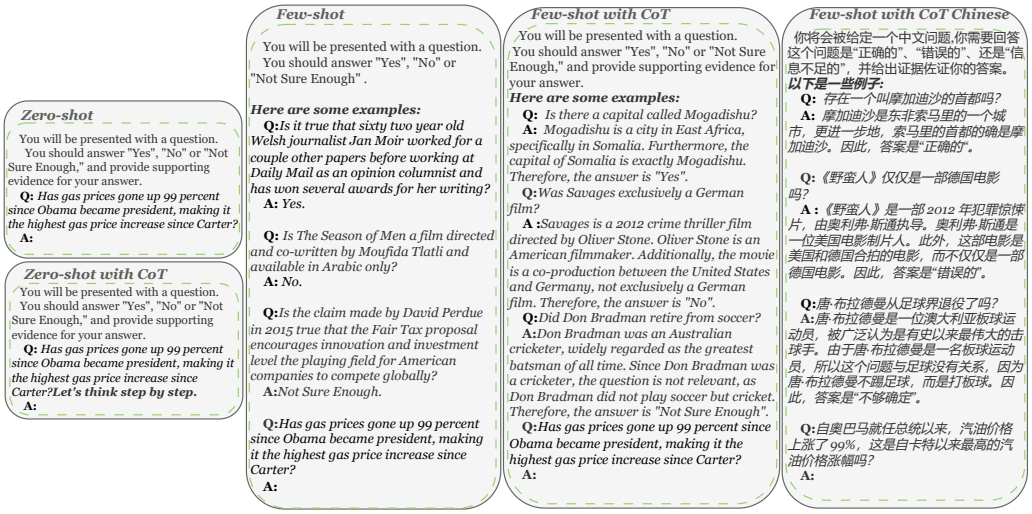


Figure 5: Prompts of four different settings.

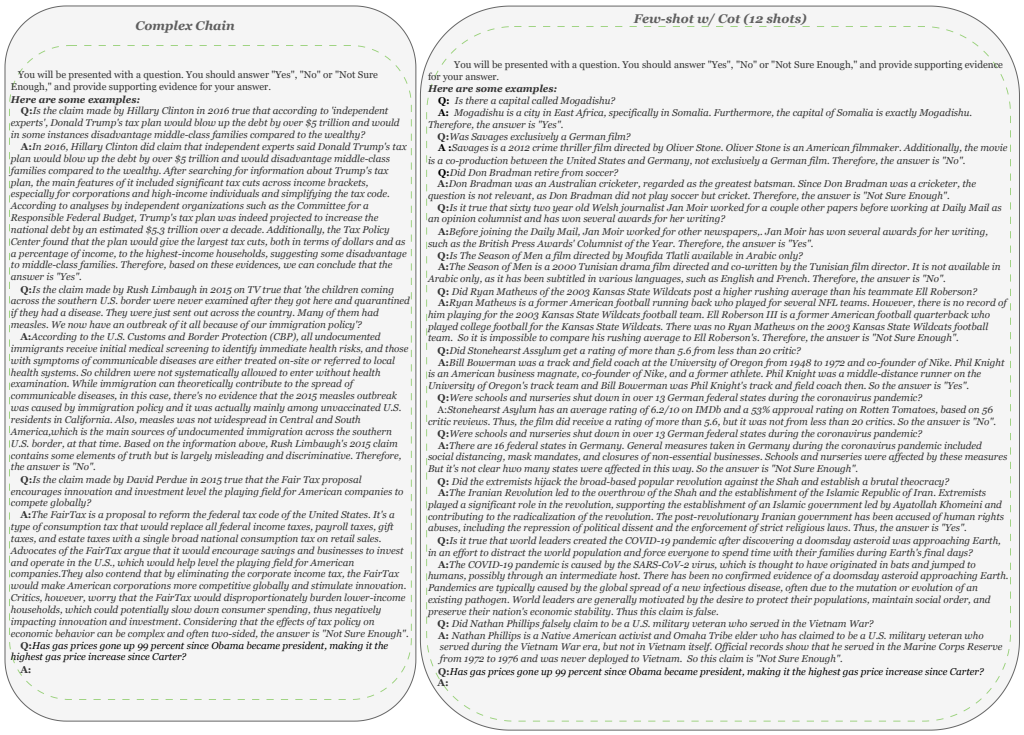


Figure 6: Prompts of complex chain and Few-shot CoT with 12 shots method.

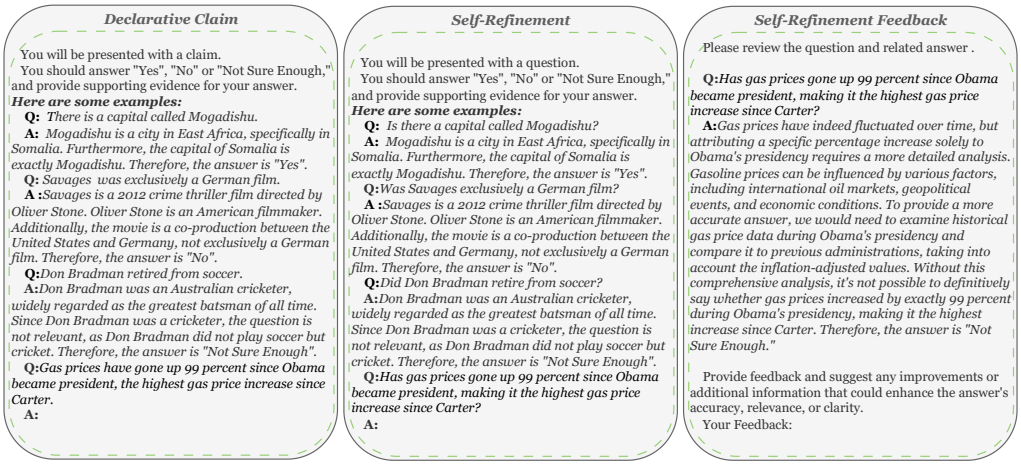


Figure 7: Prompts of self-refinement and declarative claim method.

A.5 CASE STUDY

We have introduced an additional scenario for investigation, which occurs frequently in the output generated by the zero-shot prompt method. We conducted an experiment involving three models: OPT, ChatGLM, and GPT-3.5-Turbo. These models are presented with the same set of questions, and their responses are shown in Figure 8. It is noteworthy that the OPT model, in both questions, reiterated the question itself without providing the corresponding answer. It is essential to mention that the actual output of the OPT model repeats the problem until it reaches the maximum output length (controlled by the "max_length" parameter), and we truncated the repeated portion.

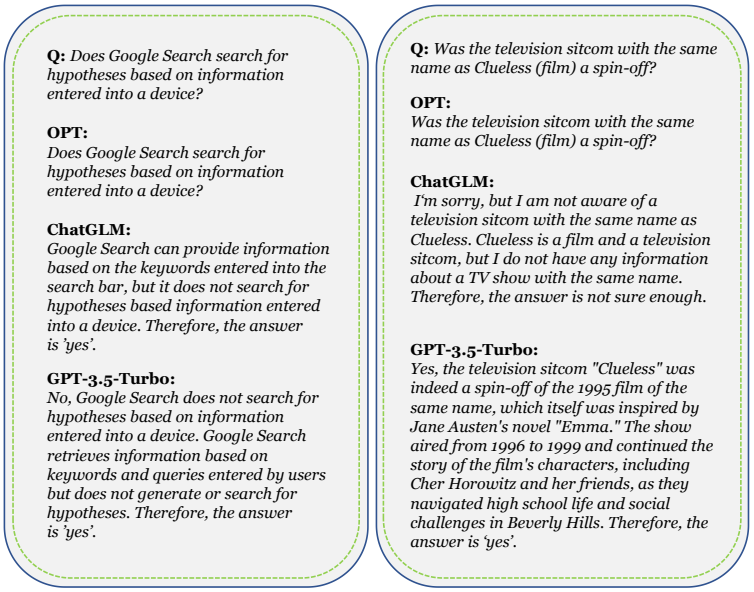


Figure 8: Answers to the same question from different LLMs in the zero-shot setting.

The OPT model even declined to generate any content when presented with the zero-shot prompt, resulting in a significant number of empty responses in the statistical results. In the first question, both ChatGLM and GPT-3.5-Turbo provided correct answers. However, in the second question, when faced with more detailed information inquiries, ChatGLM failed to produce a correct response, while GPT-3.5-Turbo demonstrated proficient reasoning and provided accurate answers. Increasing

the number of shots in a few-shot learning setting also did not yield a substantial improvement in performance.

A.6 FACT-CHECKING DATASETS

In this section, we first undertook a comprehensive survey of numerous existing Fact-Checking datasets, the summary of which is presented in Table 13. Our analysis focused on identifying the various challenges these datasets aim to address regarding factuality. We found that the challenges tackled by these datasets generally align with the seven aspects we have listed in our dataset. These aspects either appear individually or in combination across the surveyed datasets, indicating their relevance and importance in the field of fact-checking. This realization led us to intentionally design our evaluation framework around these seven specific challenges, ensuring that our benchmark is not only comprehensive but also directly addresses the core difficulties encountered in current fact-checking tasks.

Table 9: Domain Distribution of Various Fact-Checking Datasets.

Dataset	Multifaceted	Structural	Adversarial	Temporal	Real-World	Domain-Specific	Multi-Lingual
COVID-19 Disinfo (Alam et al., 2021)	✓		✓				✓
SPICED (Wright et al., 2022)	✓		✓	✓		✓	
EMU (Da et al., 2021)	✓						
NeuralNews (Tan et al., 2020)	✓						
Propa-News (Huang et al., 2022)	✓						
HOVER (Jiang et al., 2020a)	✓						
ParsFEVER (Zarharan et al., 2021)	✓						
MultiFC (Augenstein et al., 2019)	✓				✓		
Fact-KG (Kim et al., 2023a)	✓	✓					
NewsCLIPpings (Luo et al., 2021)		✓				✓	
Semeval 2021 Task9 (Wang et al., 2021)		✓					
Infotabs (Gupta et al., 2020)		✓					
TabFact (Chen et al., 2019)		✓					
InfoSurgeon (Fung et al., 2021)	✓		✓				
DeSePtion (Hidey et al., 2020)	✓		✓				
RumorEval19 (Gorrell et al., 2019)			✓		✓		
AdverBenc (Flores & Hao, 2022)h			✓		✓		
Fakeedit (Nakamura et al., 2019)			✓			✓	
Claimde-Comp (Chen et al., 2022)			✓			✓	
AVeriTec (Schlichtkrull et al., 2023)				✓	✓		
VoynaSlov (Park et al., 2022)				✓			✓
WatClaimCheck (Khan et al., 2022)				✓	✓		
MuMiN (Nielsen & McConville, 2022)		✓		✓	✓		✓
MR^2 (Hu et al., 2023)					✓		✓
FakeSV (Qi et al., 2023)					✓		✓
Weibo20 (Rao et al., 2021)					✓		✓
Rumor Stance (Lillie et al., 2019)					✓		✓
Veritas (Azevedo et al., 2021)					✓		
LIAR (Wang, 2017)					✓	✓	
FakeNewsNet (Shu et al., 2020)					✓	✓	
ClaimBuster (Arslan et al., 2020)					✓	✓	
CURT (Sundriyal et al., 2022)					✓		
Health-VER (Sarrouti et al., 2021)					✓	✓	
Covid-Fact (Saakyan et al., 2021)	✓					✓	
CoVERT (Mohr et al., 2022)	✓					✓	
Answer-Fact (Zhang et al., 2020)	✓					✓	
SciTweets (Hafid et al., 2022)						✓	✓
Dial-Fact (Gupta et al., 2021)						✓	
CHEF (Hu et al., 2022a)					✓		✓
DanFEVER (Nørregaard & Derczynski, 2021)	✓						✓
Stanceosaurus (Zheng et al., 2022)	✓						✓
Climate-FEVER (Diggelmann et al., 2020)							✓
FakeCovid (Shahi & Nandini, 2020)						✓	✓

A.7 EVIDENCE ANALYSIS

In this section, we manually analyze the inference process of the LLMs by utilizing their outputs under the few shots with CoT setting. We manually inspected a sample of 100 instances where LLMs provided reasoning outputs. The key observations from our analysis include:

1. Correct Reasoning Leading to Correct Answers: In 38% of the instances, the LLMs demonstrated a correct reasoning process, which corresponded with the accurate answers they provided. As shown in Figure 9 (a), the LLM accurately processed a query regarding the central parity rate of the RMB against the US dollar on November 30, 2020, resulting in a correct answer classified as "Factual".

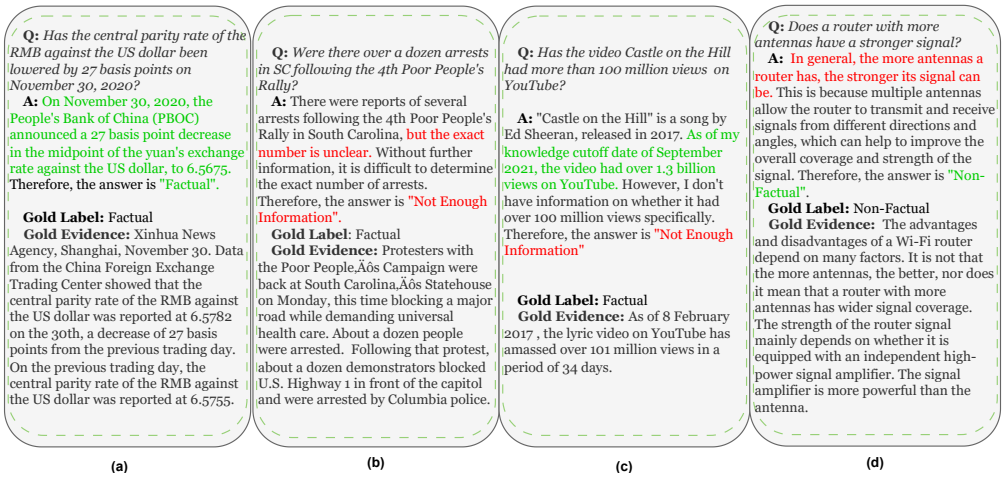


Figure 9: Resoning process from LLMs in the few-shots w/ CoT setting.

2. Flawed Reasoning Leading to I Errors: 51% of the instances exhibited incorrect reasoning processes, which inevitably led to incorrect conclusions. As an example, consider Figure 9 (b): The question posed was whether there were over a dozen arrests in South Carolina following the 4th Poor People's Rally. The LLM responded by stating there were reports of several arrests, but it could not ascertain the exact number, leading to a conclusion of "Not Enough Information." However, the 'Gold Evidence' clearly stated that about a dozen people were arrested following the demonstration, indicating that the correct label should have been "Factual." This instance underscores a scenario where the LLM might not possess the specific numerical details present in the gold evidence, leading to an incomplete and therefore inaccurate conclusion.

3. Discrepancies in Reasoning and Conclusions: Interestingly, in our analysis, we identified cases where the reasoning process did not align with the final conclusion. Specifically, there were instances, comprising 4% of our sample, where the LLMs reached a correct answer despite flawed reasoning. For example, in Figure 9 (c), the question was about whether Ed Sheeran's "Castle on the Hill" had more than 100 million views on YouTube. The LLM correctly concluded that it had over 1.3 billion views by September 2021, but stated "Not Enough Information" for the specific query of 100 million views, despite the fact being factual as per the gold evidence. Conversely, in 6% of the cases, the LLMs displayed sound reasoning but still arrived at an incorrect conclusion. A notable example of this is seen in Figure 9 (d), where the question asked whether a router with more antennas has a stronger signal. The LLM reasoned that a greater number of antennas would generally lead to a stronger signal due to enhanced multi-directional transmission and reception capabilities. Despite this logical reasoning, the LLM concluded the answer as "Non-Factual", which contradicts the reasoning provided.

4. Internal Reasoning Contradictions: We also found an example where the reasoning process of the LLM was contradictory. The question is "*Does a deficiency of vitamin B6 decrease blood levels of homocysteine?*" and LLM answers: *"Vitamin B6 is involved in the metabolism of homocysteine, which is an amino acid that is potentially harmful to blood vessels when present in high levels.* ***A deficiency of vitamin B6 can lead to an accumulation of homocysteine in the blood.*** *Therefore, a deficiency of vitamin B6 can decrease blood levels of homocysteine. The answer is "Factual"."* Initially, the LLM correctly states that a deficiency in vitamin B6 can lead to an accumulation of homocysteine. However, it concludes with a contradictory statement, suggesting that a deficiency in vitamin B6 can decrease homocysteine levels.

A.8 Baseline Analysis

Following Chen & Durrett (2019), we developed the same "no context" baseline to investigate the spurious correlations between questions and labels in our dataset. The results are shown in Table 10:

Table 10: The performance of different models across Pinocchio.

Datasets	Accuracy(%)
No Context	28.3
LLaMA-7B	31.6
Alpaca-7B	37.8
Vicuna-13B	45.2
GPT-3.5	47.0

Our experimental findings show that our dataset does not exhibit the same level of vulnerability to the exploitation of question-label correlations as observed by Chen & Durrett (2019) in the WikiHop dataset. With performance improvements of 16.9 points by Vicuna-13B and 18.7 by GPT-3.5 over the "no context" baseline, our results offer compelling evidence that our dataset is more resilient to such biases, contrary to the reported susceptibilities within WikiHop.

We extended our analysis to include a direct comparison with several established multiple-choice question-answering benchmarks, such as the WikiHop mentioned above, as well as with other prevalent benchmarks like TruthfulQA and ARC utilized in evaluating LLMs. The performance of the "no context" baseline across these benchmarks is displayed in the Table 11:

Table 11: The performance of the "no context" baseline across these benchmarks.

Datasets	Accuracy(%)
WikiHop (Welbl et al., 2018b)	59.7
TruthfulQA (Lin et al., 2021)	34.5
ARC (Clark et al., 2018b)	33.2
Ours	28.3

Evidently, our proposed dataset presented the most challenge to the "no context" baseline, marking the lowest performance compared to other datasets. The notable performance on WikiHop, with a "no context" baseline score of 59.7%, underscores the presence of spurious correlations that facilitate gaming that dataset. On the contrary, the lower baseline performances on TruthfulQA and ARC suggest that such issues are less prevalent. Our dataset, therefore, not only stands out as the least prone "to be gamed" but also underscores its robustness and the high level of rigor needed to tackle it effectively.

A.9 Peer-to-peer Analysis

The comparisons between LLaMA and its instruction-tuned versions, Alpaca and Vicuna, can be found in Table 2. Furthermore, we have conducted extra tests under the few-shots with CoT setting for T5-11B vs. Flan-T5-11B and BLOOM-6.7B vs. BLOOMz-6.7B as shown in Table 12. For T5, the accuracy was 18.6%, and the Macro F1 was 25.2%. In contrast, as shown in Table 2, Flan-T5 achieved an accuracy of 38.4% and a Macro F1 of 38.4%. Similarly, BLOOM's performance was at an accuracy of 6.6% and Macro F1 of 12.2%, whereas BLOOMz showed a marked improvement with an accuracy of 27.5% and a Macro F1 of 27.7%. These peer-by-peer comparisons reveal that, with few exceptions (e.g., LLaMA vs. Alpaca in terms of Macro F1), models that underwent instruction tuning generally outperform their backbone counterparts, achieving an average improvement of 11.3%.

A.10 Related Work: Question Answering Datasets

In this section, we offer a thorough examination of existing question-answering initiatives as they pertain to the seven key dimensions that form the core of our benchmark. These dimensions are multifaceted, structural, adversarial, temporal, real-world, and multilingual.

As detailed in Table 13, we present a comprehensive overview of notable datasets within the realm of question-answering. We categorize these datasets based on several criteria to illuminate their

Table 12: Peer-to-peer comparison between the instruction-tuned models and their backbones.

Models	Accuracy(%)	Macro F1(%)
LLaMA-7B	35.3	31.4
Alpaca-7B	39.4	26.2
Vicuna-7B	48.5	40.6
T5-11B	18.6	25.2
Flan-T5-11B	38.4	38.4
Bloom-6.7B	6.6	12.2
Bloomz-6.7B	27.5	27.7

distinctive challenges and characteristics. First, we identify the "Type" of challenge each dataset presents. Next, "Source" provides the origins of the questions. "Retrieval" indicates the necessity of sourcing external knowledge, such as documents, to formulate an answer. When it comes to the "Answer types", datasets may require various forms of responses ranging from multiple-choice options (A, B, C, etc.), specific text spans (e.g., an entity or a phrase), to Boolean (yes or no) and free-form answers that allow for the generated text of any length. "Domain" captures the field to which the questions belong, encompassing areas like science, biography, or geography.

Interestingly, beyond these seven axes, there exist other datasets that probe the knowledge and reasoning capabilities of large language models (LLMs) from different perspectives. For instance, research centered around knowledge updating, particularly focusing on entities, has been conducted. Onoe et al. (2022) delve into the ability of LLMs to make inferences about newly emerged entities that were not part of the LLMs' pretraining data. Building on this, Onoe et al. (2023) investigated the extent to which LLMs can integrate descriptions of new entities. On the other hand, Peng et al. (2022) have assessed LLMs' conceptual knowledge by crafting three distinct tasks that test whether LLMs are capable of categorizing entities based on conceptual similarities.

Multifaceted Existing efforts in question answering that relate to the multi-faceted nature of our dataset predominantly encompass multi-hop reasoning datasets. These datasets necessitate models to synthesize multiple information snippets to formulate an answer. For instance, WikiHop Welbl et al. (2018a) constructs a bipartite graph from a knowledge base populated with relational triplets. This graph undergoes a breadth-first traversal to yield valid multi-hop reasoning chains. Similarly, HotpotQA (Yang et al., 2018) narrows its focus to 2-hop questions derived from the initial sections of English Wikipedia documents. The selection of two passages to form a reasoning chain is predicated upon one of two conditions: either a hyperlink connects the first document to the second, or the associated entities belong to an identical category. Moving onto a broader spectrum, MultiRC (Khashabi et al., 2018) introduces multi-domain multi-hop questions. It compiles documents from various domains and a multitude of datasets, where the different contexts are all embedded within the same textual passage. As opposed to the questions themselves providing explicit decompositional cues, StrategyQA (Geva et al., 2021) conceals the necessary reasoning steps within the question itself. These steps must be astutely deduced using strategic inference.

Additionally, several datasets intertwine multi-hop reasoning with further complexities. OpenBookQA (Mihaylov et al., 2018) offers a specialized challenge, combining question answering techniques with a compendium of scientific facts to assess knowledge in the scientific domain, supplemented by a broader base of common understanding. In a vein similar to OpenBookQA, QASC (Khot et al., 2020) also revolves around two-hop question answering with a foundation in scientific facts; its methodology for reasoning chain generation closely resembles that of OpenBookQA. Furthermore, datasets like HybridQA (Chen et al., 2020) and OTTQA (Chen et al., 2021b) venture into multi-hop reasoning across both tabular and textual data sources. Vu et al. (2023) introduces FreshQA, incorporating questions that demand multi-hop reasoning where answers may shift over time, as well as tackling premises that are fundamentally flawed.

Structural Structured and semi-structured knowledge are known as unambiguous and compositional. Traditional question answering datasets predominantly cater to uniform types of information, either focusing exclusively on textual data or relying solely on knowledge bases and tables (Berant et al., 2013; Talmor & Berant, 2018). This approach, however, overlooks the complexity of human

knowledge which is inherently diverse and spread across varied formats. Relying solely on homogenous sources may result in limited scope and inadequate coverage of information. Addressing this gap, Chen et al. (2020) proposed HybridQA, a dataset that necessitates reasoning over a blend of heterogeneous information sources. In HybridQA, each question necessitates the integration of information from a Wikipedia table and assorted text corpora tied to the entities mentioned within the table, thereby combining tabular and textual data. Parallel initiatives targeting niche fields also emerged, with Li et al. (2021) focusing on geographical data and Zhu et al. (2021) on financial information. These domain-specific endeavors highlight the growing interest in incorporating structural knowledge. Departing from the provision of pre-selected tables and textual passages, OTTQA (Chen et al., 2021a) and NQ-table (Herzig et al., 2021) propel the question-answering challenge into the open-domain setting. Here, the retrieval of pertinent tables and text from comprehensive sources like Wikipedia becomes an integral part of the task. Our structural task aligns more closely with the objectives of OTTQA and NQ-table, where LLMs are tasked with performing advanced multi-hop inference. This entails navigating through a combination of both structural and unstructured factual knowledge to deduce accurate answers, reflecting a more realistic and complex information processing challenge akin to the ways humans interact with a variety of knowledge types to make informed decisions.

Adversarial Machine learning models have a known susceptibility to adversarial examples—inputs that have been intentionally modified to cause a model to make a mistake. A notable instance within the realm of question answering tasks is the presence of questions based on dubious assumptions, which are typically classified as unanswerable questions (Rajpurkar et al., 2018; Kwiatkowski et al., 2019; Asai & Choi, 2021). More recently, Kim et al. (2021) critiqued the practice of lumping questions with dubious assumptions into the 'unanswerable' category as inadequate. They advocated for employing presuppositions within explanations as a means to more effectively determine their unanswerability. Additionally, their work demonstrates the complexity of verifying assumptions, proving it to be a formidable challenge even in closed-book environments. Building upon this, Kim et al. (2023b) expanded the investigation into open-domain contexts, confirming the inherent difficulties associated with QA that involve problematic assumptions. They discovered that, even when the hurdle of recognizing assumptions is eliminated, the task of factual verification remains unsolved—though, it should be noted, recent enhancements in LLMs have indeed contributed to some progress in verification capabilities. In a related vein, Yu et al. (2023b) presented a new open-domain QA dataset that features a natural distribution of failures due to presuppositions. Their research reveals that the challenges in handling questions with questionable assumptions are consistent, irrespective of the different sources from which the questions are derived, which include search engine prompts as well as Reddit inquiries. This body of work indicates that while strides have been made in addressing some aspects of QA tasks, the nuanced issue of dealing with questionable assumptions persists across various settings and requires further exploration.

Temporal Understanding the temporal evolution of information is a significant area of interest in the field of question answering. Initial research, such as TempQuestions (Jia et al., 2018), investigated temporal aspects of questions that incorporated time specifiers within knowledge bases. Subsequent studies have shifted their focus toward apprehending the nuances of temporal progression in natural language texts. For example, Chen et al. (2021c) introduced TimeQA, a resource constructed by extracting and compiling evolving facts from WikiData alongside corresponding Wikipedia passages, resulting in a dataset of 20,000 timestamped question-answer pairs. Moreover, Zhang & Choi (2021) presented SituatedQA, which includes 9,000 realistically formulated questions from pre-existing open-domain QA datasets, each complemented with temporal contexts, such as specific timestamps. StreamingQA (Liska et al., 2022) is another relevant contribution that encompasses a blend of machine-generated and human-authored questions—altogether totaling 146,000 entries—designed to be answerable using a repository of timestamped news articles. In the same vein, the dynamic RealTimeQA benchmark (Kasai et al., 2022b) poses a challenge for models by offering 30 multiple-choice questions based on recent events curated from news websites, thereby testing their ability to handle fresh content. Adding to these advancements, FreshQA (Vu et al., 2023) brings a new dimension to the table with a static compilation of human-curated open-ended questions. The uniqueness of FreshQA lies in the evolving nature of its answers, which are subject to change in response to ongoing world developments, providing a generative assessment for time-sensitive question answering. This body of work collectively underscores the complexity and dynamism inherent in temporal question answering research.

Domain-Specific While there have been successful developments in question-answering within broad domains, specialized domains such as science and biomedicine remain relatively underexplored and present unique challenges. The limited availability of domain-specific datasets, coupled with the need for an in-depth understanding of specialized knowledge to match that of human experts, marks these areas as fertile ground for ongoing research. In the scientific domain, existing datasets necessitate the use of varied reasoning methods tailored to each specific question (Clark et al., 2018a). For instance, the OpenBookQA dataset (Mihaylov et al., 2018) presents multiple-choice questions that are generated based on a core book of fundamental science facts. Similarly, the QASC dataset (Khot et al., 2020) offers multiple-choice questions on science topics appropriate for elementary and middle school levels, emphasizing the combination of facts. QASC is unique in that it intentionally includes pairs of facts that, according to evaluations by crowd workers, provide enough information to deduce the answer to each question. Shifting the focus to the biomedical field, a range of new datasets have emerged to support question-answering tasks that hinge on domain-specific expertise. These include datasets such as HealthQA (Zhu et al., 2019), MASH-QA (Zhu et al., 2020a), and MedMCQA (Pal et al., 2022), which have been introduced to bolster research in medical question-answering applications. These datasets serve as valuable resources to address the nuanced queries that arise within the complex terrain of biomedical knowledge.

Multi-Lingual Recent effort has been made to create non-English QA datasets to overcome the data scarcity in non-English languages, typically including one or two languages. These include DuReader (He et al., 2018) in Chinese, French/Japanese evaluation sets for SQuAD created via translation (Asai et al., 2018), a semi-automatic Italian translation of SQuAD (Croce et al., 2019), ARCD—an Arabic reading comprehension dataset (Mozannar et al., 2019), a Hindi-English parallel dataset in a SQuAD-like setting (Gupta et al., 2018), and a Chinese–English dataset focused on visual QA (Gao et al., 2015). Recent datasets cover more languages, such as XQuAD (Artetxe et al., 2020) and MLQA (Lewis et al., 2020), which are examples of SQuAD-style extractive datasets, employing human translators to create parallel examples. MLQA and XQuAD ensure that all answers are answerable, and derive answers from provided documents. Instead of extractive answers, Hardalov et al. (2020) introduced EXAMS, a multilingual multiple-choice QA from school exams. TyDiQA (Clark et al., 2020) and MKQA (Longpre et al., 2021), focus on typological diversity in its wide language selection. While TyDiQA offers a more natural distribution of questions, its annotations are based on the retrieval system used by the authors (Google search); hence their answers are actually start and end indices for spans of text within a given passage. Xor QA (Asai et al., 2021) explores cross-lingual subtasks by re-annotating TyDiQA examples, sourcing answers from English documents, and translating them back to the target language. While state-of-the-art models have matched or surpassed human performance in general-purpose monolingual benchmarks, current methods still fall short of human performance on multilingual benchmarks, despite recent gains. Multilingual question answering consequently is at the frontier of such cross-lingual generalization.

Table 13: A comprehensive comparison of question answering datasets.

Dataset	Type	Source	Retrieval	Answer Type	Domain
WikiHop (Welbl et al., 2018a)	Multifaceted	WikiData	✗	Multiple Choice	General
HotpotQA (Yang et al., 2018)	Multifaceted	Wikipedia	✓	Span	General
MultiRC (Khashabi et al., 2018)	Multifaceted	Multiple	✗	Multiple Choice	General
StrategyQA (Geva et al., 2021)	Multifaceted	Wikipedia	✓	Boolean	General
OpenBookQA (Mihaylov et al., 2018)	Multifaceted/Domain-Specific	WorldTree	✓	Multiple Choice	Science
QASC (Khot et al., 2020)	Multifaceted/Domain-Specific	Wikipedia	✓	Multiple Choice	Science
NQ-tables (Herzig et al., 2021)	Structural	Google Queries	✓	Span	General
TAT-QA (Zhu et al., 2021)	Structural/Domain-Specific	Wikipedia	✗	Span	Finance
TSQA (Li et al., 2021)	Structural/Domain-Specific	Exam	✗	Multiple Choice	Geography
HybridQA (Chen et al., 2020)	Structural/Multifaceted	Wikipedia	✗	Span	General
OTTQA (Chen et al., 2021b)	Structural/Multifaceted	Wikipedia	✓	Multiple Choice	General
$(QA)^2$ (Kim et al., 2023b)	Adversarial	Google Queries	✗	Free-form	General
CREPE (Yu et al., 2023b)	Adversarial/Real-World	Reddit	✗	Free-form/Boolean	General
TempQuestions (Jia et al., 2018)	Temporal	Datasets	✗	Free-form/Boolean	General
TimeQA (Chen et al., 2021c)	Temporal	WikiData	✗	Span	General
SituatedQA (Zhang & Choi, 2021)	Temporal	Datasets	✓	Span	Geography
RealTimeQA (Kasai et al., 2022a)	Temporal	News	✓	Multiple-Choice	General
StreamingQA (Liska et al., 2022)	Temporal	News	✓	Free-form	General
FreshQA (Vu et al., 2023)	Temporal/Multifaceted	Manual	✓	Free-form	General
MSMarco (Nguyen et al., 2016)	Real-World	Bing Queries	✓	Free-form	General
SearchQA (Dunn et al., 2017)	Real-World	Google Queries	✓	Span	General
TriviaQA (Joshi et al., 2017)	Real-World	Forum	✓	Span	General
DuReader (He et al., 2018)	Real-World	Baidu Queries	✓	Free-form	General
NQ (Kwiatkowski et al., 2019)	Real-World	Google Queries	✓	Free-form	General
ELI5 (Fan et al., 2019)	Real-World	Reddit	✓	Free-form	General
ARC (Clark et al., 2018a)	Domain-Specific/Multifaceted	Search Queries	✓	Multiple Choice	Science
QASPER (Dasigi et al., 2021)	Domain-Specific	Papers	✓	Span	Science
ScienceQA (Lu et al., 2022)	Domain-Specific	Exams	✗	Multiple Choice	Science
HealthQA (Zhu et al., 2019)	Domain-Specific	Patient	✗	Free-form	BioMed
MedMCQA (Pal et al., 2022)	Domain-Specific	Exams	✗	Multiple Choice	BioMed
MASH-QA (Zhu et al., 2020b)	Domain-Specific	WebMD	✗	Free-form	BioMed
XQuAD (Artetxe et al., 2020)	Multilingual	SQuAD	✗	Span	General
MLQA (Lewis et al., 2020)	Multilingual	Wikipedia	✗	Span	General
EXAMS (Hardalov et al., 2020)	Multilingual	Exam	✗	Multiple Choice	General
TydiQA (Clark et al., 2020)	Multilingual/Real-World	NQ	✗	Span	General
MKQA (Longpre et al., 2021)	Multilingual/Real-World	NQ	✗	Multiple	General
XOR QA (Asai et al., 2021)	Multilingual/Real-World	TydiQA	✓	Span	General