CK12: A Rounded K12 Knowledge Graph Based Benchmark for Chinese Holistic Cognition Evaluation

Weihao You*, Pengcheng Wang*, Changlong Li, Zhilong Ji†, Jinfeng Bai

Tomorrow Advancing Life

{youweihao, wangpengcheng2, lichanglong, jizhilong, baijinfeng1}@tal.com

Abstract

New NLP benchmarks are urgently needed to align with the rapid development of large language models (LLMs). We present a meticulously designed evaluation benchmark that leverages the knowledge graph. This evaluation comprises 584 level-1 knowledge points and 1,989 level-2 knowledge points, thereby encompassing a comprehensive spectrum of the K12 education domain knowledge. The primary objective is to comprehensively assess the high-level comprehension aptitude and reasoning capabilities of LLMs operating within the Chinese context. Our evaluation incorporates five distinct question types with 39,452 questions. We test the current mainstream LLMs by three distinct modes. Firstly, four prompt evaluation modes were employed to assess the fundamental capacity. Additionally, for choice questions, a result-oriented evaluation approach was designed through data augmentation to assess the model's proficiency in advanced knowledge and reasoning. Moreover, a subset with reasoning process is derived, and the process-oriented testing method is used to test the model's interpretability and higher-order reasoning capacity. We further show models' capability in our knowledge points, and anticipate the evaluation can assist in the assessment of the strengths and deficiencies of LLMs on knowledge points, thus fostering their development within the Chinese context. Our Dataset will be publicly available in https://github.com/tal-tech/chinese-k12-evaluation.

Introduction

In recent times, we have witnessed a remarkable surge in the advancement of Large scale Language Models (LLMs), which have found diverse applications across various domains. Notably, ChatGPT and GPT-4 have demonstrated exceptional performance . It is imperative to propose more comprehensive and forward-looking evaluations than ever before to ascertain the extent to which LLMs have genuinely acquired knowledge.

The most important aspect of the field of educational knowledge is cultivating a broad range of cognitive abilities, including memory retention, abstraction, logical reasoning, analytical thinking and imagination. For LLMs, such educational knowledge serves as a highly suitable tool for assessing comprehensive capabilities.

Some benchmarks like GLUE (Wang et al. 2018), SuperGLUE (Wang et al. 2019), MMLU (Hendrycks et al. 2021a), RACE (Lai et al. 2017), GSM8K (Cobbe et al. 2021), LAMBADA (Paperno et al. 2016), TriviaQA (Joshi et al. 2017), Truthfulqa (Lin, Hilton, and Evans 2021), HELM (Liang et al. 2022) and BIG-bench (Srivastava et al. 2022), effectively evaluate the performance of LLMs at a given moment and measuring whether the model can be further improved. However, these benchmarks predominantly cater to the English language. Consequently, when evaluating Chinese LLMs, disparities arising from inherent linguistic dissimilarities can lead to inequitable comparisons.

Therefore, several benchmarks for the Chinese language have been proposed, such as CLUE (Xu et al. 2020), FewCLUE (Xu et al. 2021), SuperCLUE, APE210K (Zhao et al. 2020), GAOKAO (Zhang et al. 2023), L-Eval (An et al. 2023), AGI-Eval (Zhong et al. 2023), MMCU (Zeng 2023), CMMLU (Li et al. 2023), C-Eval (Huang et al. 2023), M3KE (Liu et al. 2023) and Xiezhi (Gu et al. 2023). It is important to underscore that while these sets span a wide range of subjects, they may display restricted coverage of knowledge points and an incomplete knowledge system.

Taking into account the dynamic progression of LLMs, we have identified certain limitations prevalent in previous benchmarks. Firstly, The accuracy on some benchmarks GLUE (Yang et al. 2022), SuperGLUE (Wang et al. 2019) have reached over 90, surpassing human performance, and it is impossible to accurately measure whether the model ability has improved. Secondly, M3KE (Liu et al. 2023), AGI-EVAL (Zhong et al. 2023), C-EVAL (Huang et al. 2023) and Xiezhi (Gu et al. 2023) rely on questions sourced from examination papers. These examination papers, constrained by their respective paper lengths, may not comprehensively cover the full range of knowledge points, such benchmarks fall short of covering the entire knowledge structure of a given subject. There is no comprehensive Chinese benchmark that currently targets K12 (6 years of elementary school, 3 years of middle school, and 3 years of high school) educational knowledge points.

Using CoT testing mode on some models can generate the inference steps of the model to answer the question, which

*Authors contribute equally.

†Corresponding author

greatly improves the interpretability and verifiability of the model (Nye et al. 2021), (Wei et al. 2023). It is important to have a benchmark with high-quality analytical reasoning steps and metrics that allow for a fair assessment of the reasoning process. At present, there are few evaluation sets with inference steps. GSM8K provides a relatively comprehensive set of logical reasoning steps, but its content coverage is limited. On some evaluation sets similar to C-EVAL, their explanations are generated by GPT-4, the quality of which is relatively hard to ensure. Lastly, given the rapid pace at which LLMs undergo updates, there is a possibility that numerous benchmarks may have inadvertently incorporated training data, resulting in data contamination.

In order to accurately assess the K12 knowledge capability demonstrated by LLMs, we propose a new benchmark based on our self-built multi-level K12 knowledge graph question bank. The knowledge graphs and questions involved all come from our question bank, which has accumulated over decades, and has been carefully reviewed by professional teachers to avoid issues of discrimination and bias, ensure the quality of the questions and the coverage of knowledge. In particular, the benchmark encompasses all K12-related knowledge-based subjects, categorized into primary school, middle school and high school stages. Furthermore, we have incorporated two more complex evaluation methods for single-choice and multiple-answer questions, which further reduce the possibility of data contamination and explore the ability of LLMs to apply knowledge and reasoning: shuffling options and adding distractors, to achieve a more comprehensive ability assessment.

Previous work on evaluating the performance of LLMs has tended to focus on the results, without evaluating the inference steps, which makes it difficult to objectively study their correctness (independent of the final answer). We extract 7,334 mathematical problems with more than two parsing steps from our benchmark as the reasoning evaluation subset. Reference (Golovneva et al. 2022), we use several interpretability metrics to evaluate CoT results for the reasoning subset.

Our main contributions are summarized as follows.

- We propose Knowledge Graph Based Benchmark, an evaluation set for Chinese LLMs, constructed based on our self-developed multi-level knowledge graph. Up to now, it covers the most comprehensive knowledge points in Chinese K12 field.
- We have tested a wide range of open-source Chinese LLMs and some representative English LLMs. And we present a detailed analysis of the model results.
- We derive a subset that evaluates the ability of inference steps to evaluate the higher-order reasoning ability of LLMs by the interpretability metrics.

Related Works

Large Language Models Recently, several prominent models have been released by various institutions, including BLOOM (Scao et al. 2022), Llama (Touvron et al. 2023), Alpaca (Taori et al. 2023), Vicuna (Chiang et al. 2023), ChatGPT (OpenAI 2023a), GPT-4 (OpenAI 2023b). These models have demonstrated remarkable performance across a wide range of tasks, even surpassing human capabilities in certain aspects. Furthermore, there has been a rapid development in Chinese LLMs, including ChatGLM (Du et al. 2022), Baize (Xu et al. 2023), Belle (Ji et al. 2023), InternLM, Qwen, Baichuan and Moss (Sun et al. 2023) which have exhibited a remarkable comprehension of the Chinese language. To assess the LLM's ability to achieve human-level or superhuman performance, it is crucial to evaluate its fundamental capabilities. The evaluation set should be as comprehensive and detailed as possible. A more challenging and newer Chinese evaluation set is crucial for evaluating the capabilities of LLMs effectively.

Benchmarks There are several studies that focus on evaluating a model's level of knowledge and capacity for reasoning. Physical IQA (Bisk et al. 2020) and CosmosQA (Huang et al. 2019) primarily assess the model's comprehension of common knowledge. GLUE (Wang et al. 2018), MMLU (Hendrycks et al. 2021b), AGI-EVAL (Zhong et al. 2023) and C-EVAL (Huang et al. 2023) are designed to evaluate the model's overall capability. Additionally, M3KE (Liu et al. 2023) introduces a multi-level and multi-disciplinary domain knowledge test set, while Xiezhi (Gu et al. 2023) presents an extensive domain knowledge evaluation set with the largest number of questions. However, these benchmarks are mainly collected from standardized test papers and often focus on university-level subjects. As a result, due to the influence of the overall length of the test papers, it cannot be guaranteed to cover all the necessary knowledge points, nor can it fully reflect the entire knowledge structure of the subject. And if considering the human learning process, there is a lack of evaluation sets that comprehensively cover knowledge points in the K12 field.

Some prior works (Golovneva et al. 2022) (Deng et al. 2021) (Yuan, Neubig, and Liu 2021) on evaluating reasoning chains have proposed reasonable metrics. In this work, we extract some mathematical problems as the evaluation subset to evaluate the reasoning abilities of LLMs.

Therefore, we propose a K12 domain knowledge evaluation set based on self-developed knowledge graphs, that ensures comprehensive coverage of all K12 knowledge points.

Knowledge Graph Based Benchmark

Design Principle Our benchmark encompasses a diverse range of topics in K12 education, covering primary, middle and high school levels. It includes all major subjects mentioned above, covering textbook materials, teaching materials, experimental knowledge and more. It has undergone multiple levels of knowledge graph classification and annotation by professional teachers.

Discipline Categories We referred to the Chinese Discipline Taxonomy for primary, middle and high school subjects and divided the first level labels into the following subjects: Chinese, Mathematics, English, Politics, History, Geography, Physics, Chemistry and Biology. For the convenience of statistical analysis, we divide it into two main categories, Arts and Sciences. The Arts include Chinese, En-

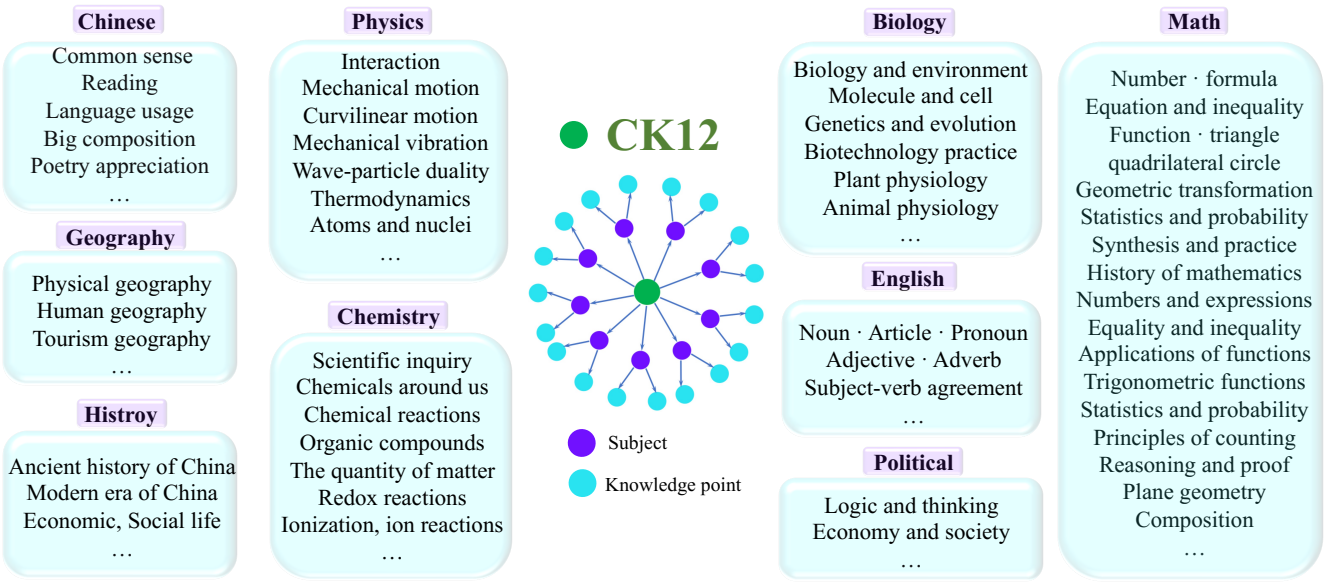

Figure 1: Our proposed benchmark is built around knowledge graphs, and encompasses 9 k12 education subjects. 584 knowledge points, and 1,989 further fine-grained knowledge points, aiming to provide a knowledge benchmark for the fair, effective, and comprehensive evaluation.

glish, History, Politics and Geography; the Sciences include Mathematics, Chemistry, Biology and Physics.

The first level indicates the subject. The second level called level-1 knowledge points further refine each respective subject covering textbook materials, teaching materials, and experimental knowledge of K12. Lastly, the third level labels which are called level-2 knowledge points continue to be divided based on specific level-1 knowledge point categories. Notably, a third level encompasses no fewer than 10 questions, rendering it a comprehensive unit of evaluation. As shown in Figure 1, our proposed benchmark includes 9 K12 disciplines and 584 level-1 knowledge points. There are 1,989 level-2 knowledge points in total, so we only show the level-2 knowledge points of mathematics subject in Figure 2.

The knowledge graph serves as the cornerstone of our benchmark, meticulously crafted by experienced teachers to encompass every single knowledge point across all K12-related subjects, and encapsulate a complete architecture of the knowledge system. By leveraging this knowledge graph, our benchmark assesses the holistic foundational aptitude and knowledge expansion capacity of LLMs in the K12 domain, providing a comprehensive overview of the abilities of various LLMs in these knowledge domains.

According to bloom's taxonomy of educational objectives, the developmental trajectory of LLM's cognitive abilities should progress from basic memory and understanding to analysis, application, ultimately evaluation and creation. These abilities that LLM needs to possess align with the intentions of our human education. Analogous to human learning, we start from elementary school, through middle school and high school, continue into university and even more advanced domains like doctoral studies, our cognitive abilities largely follow this pattern. There is a lack of comprehensive assessments specifically designed for K12 knowledge. The benchmark we propose is built on a knowledge graph constructed by professional teachers, ensuring both the comprehensive knowledge structure and coverage in K12.

To minimize the risk of data contamination, we implemented a shuffling mechanism and introduced interference options for the choices in our choice questions to further assess LLM's ability to apply knowledge and reasoning ability. These interference options were extracted from the choices of other choice questions within the same subject. We randomly selected 10 options and ranked them accordingly. In the section of Comparison on Data, the comparison results of the random data, the data with added interference options and the original data are provided.

It is crucial to highlight that 74% percent of the questions in our proposed benchmark incorporate a meticulous analysis. This analysis have undergone rigorous scrutiny by professional educators and exhibit exceptionally high quality. A mathematics sample with analysis is shown in few-shot-CoT in Figure 4. We will also make these high-quality question explanations available to the public, hoping that can help the open community optimize and develop even higher level large-scale models. Furthermore, we extract 7,334 mathematical problems with more than two parsing steps from our benchmark as the reasoning evaluation subset and use several interpretability metrics to evaluate reasoning step results for the reasoning subset.

Data Collection Our benchmark consists of 39,452 questions sourced exclusively from a private test bank. Moreover, each question has been carefully selected and analyzed to avoid issues of discrimination and bias, ensure the quality and fairness of the data. These questions encompass a

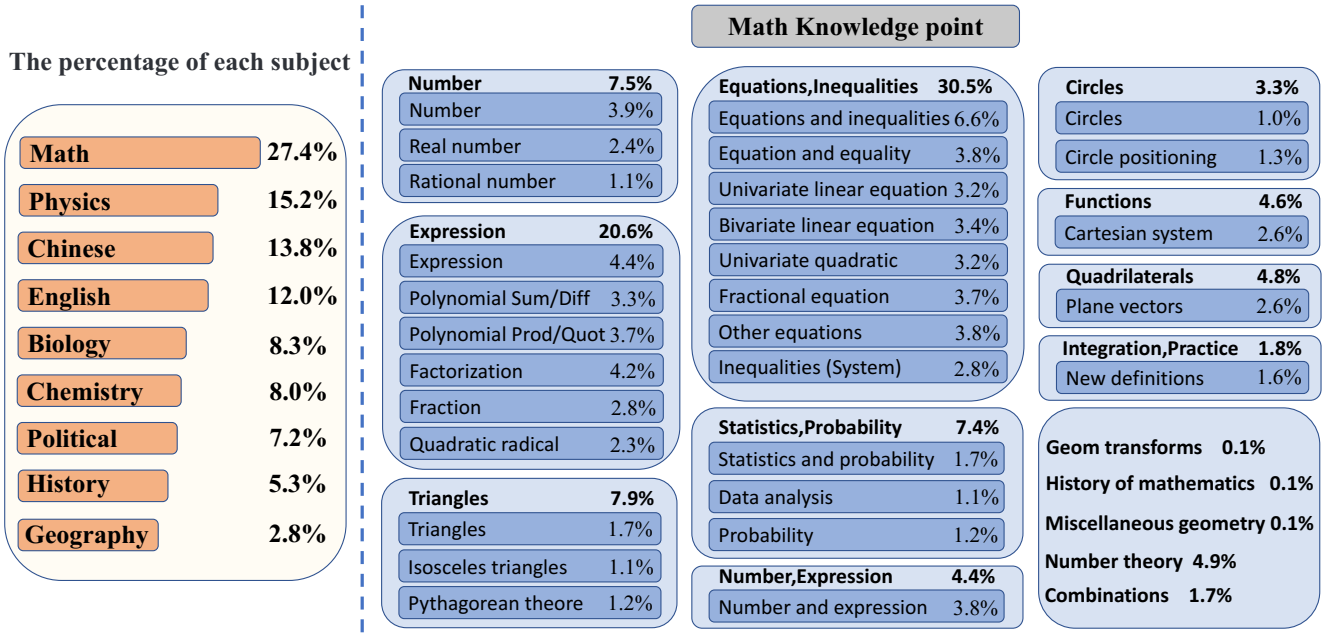

Figure 2: The figure on the left is the statistics of subjects in our benchmark.The figure on the right shows statistics for fine-grained Math knowledge points, including all level-1 and level-2 points above a 1% share.

diverse range of origins, including internal confidential papers, external schools examination papers and independent enrollment plan papers. To ensure comprehensive coverage, on each third level label of our knowledge graph, we have selected 10 questions for each type of question, which includes single choice, multiple choice, blank filling, judgment and sorting. However, some of the third level labels only contain one or several types of questions. Therefore, according to the proportion of each type of question, they constitute 48.9%, 12.9%, 7.7%, 25.2%, and 0.5% respectively.

We classify and analyze the original questions and do some processing. Firstly, some transcription questions and dictation questions are removed, which are not significant for model ability evaluation. Then, we found that some blank filling questions mixed choice and blank filling, and we removed these questions. The main reason is that the question stem does not directly indicate what the question should do, which can lead to confusion.

Scoring Step-By-Step Reasoning In order to make a reasonable evaluation of the reasoning ability of the LLMs, we selected 7,334 math problems with more than two analysis steps as the reasoning test set. Subsequently, we proceeded to evaluate the few-shot-CoT results of all test models.

According to (Golovneva et al. 2022), We selected five metrics to evaluate the accuracy of the reasoning step. Faithfulness-Step measures if the model misinterpreted the problem statement. Source-Consistency evaluates whether the reasoning steps are consistent with the content of the question. Reasoning Alignment is to evaluate whether the reasoning steps match the parsing. Hallucination was performed to assess whether the responses contained irrelevant reasoning steps. Redundancy is to evaluate whether the inference step contains information that is not required to solve the problem. While Source-Consistency was assessed using an NLI model that was trained to classify hypothesis-context pairs into entailment, neutral and contradiction classes (Laurer et al. 2022). The Reasoning Alignment was assessed using the all-mpnet-base-v2 model and the other metrics are assessed by using roscoe-512-roberta-base model which is finetuning on SimCSE (Gao, Yao, and Chen 2021) to extract alignment vector (Deng et al. 2021) and then calculating them through predefined rules. A detailed description of the calculation steps can be found in the Appendix.

Experiments

In this section, we will introduce our evaluation setups, including identical prompts, few-shot and CoT samples, error analysis and comparison of results. And we use accuracy as the metric to evaluate different LLMs.

Prompt In order to accommodate different models, we standardized the Prompt, but some models require additional specific instructions, such as Moss, which should be modified to "$< |Human| >: question < eoh > \n < |Moss| >:$". The provided Figure 3 and Figure 4 showcase the prompt used for the evaluation of single-choice questions, encompassing four prevalent test modes: zero-shot, few-shot, zero-shot-CoT, and few-shot-CoT. The test prompt for other question types can be found in the Appendix.

Few-shot and CoT Setup We extract a set of 5 questions from the evaluation set as few-shots examples. However, if the number of tokens surpass 1024 during the testing phase, we adjust the input to 3 or 1 shots. For CoT testing (Kojima

Zero-shot

下面是一道选择题，题干之后是四个选项，请从ABCD中选择出正确的选项，直接给出正确选项对应的结果即可。题目:
Here is a single-choice question. Following the question stem are four options, please choose the correct option from ABCD and provide the corresponding result. Question:

题目内容... Question content …
特别注意，不需要给解析，并且回答的格式应为“正确答案是：
Special attention, no need for interpretation, and the format of the answer should be "The correct answer is:

Few-shot

请选出下列单选题的正确答案，直接给出正确选项对应的结果即可
Please select the answer for the following single-choice questions. Simply provide the corresponding result for the correct option.

例1: 题目内容... Example 1: Question content …
答案：... Answer: …
…[5-shot examples]…

题目内容... Question content …

Figure 3: Examples of zero-shot and few-shot evaluation with the single choice question.

Zero-shot-CoT

下面是一道选择题，题干之后是四个选项，请从ABCD中选择出正确的选项，直接给出正确选项对应的结果即可。题目:
Here is a single-choice question. Following the question stem are four options, please choose the correct option from ABCD and provide the corresponding result. Question:

题目内容... Question content …
特别注意，不需要给解析，并且回答的格式应为“正确答案是：
Special attention, no need for interpretation, and the format of the answer should be "The correct answer is:

让我们一步一步思考。
Let's think step by step.

Few-shot-CoT

请选出下列单选题的正确答案，直接给出正确选项对应的结果即可
Please select the answer for the following single-choice questions. Simply provide the corresponding result for the correct option.

例1: 题目内容... Example 1: Question content …
答案: 让我们一步一步思考。答案内容...
Answer: Let's think step by step. Answer content…
…[5-shot examples]…

题目内容... Question content …
正确答案是：让我们一步一步思考
True Answer: Let's think step by step:

Figure 4: Examples of zero-shot-CoT and few-shot-CoT evaluation with the single choice question. The green text is the analysis process.

et al. 2022), (Wei et al. 2022), adding “Let's think it step by step” at the end of question. The detailed setup is shown in Figure 4 .

Assessed Models In order to conduct a comprehensive evaluation of the current LLMs on our benchmark, we selected the current 8 state-of-the-art models for testing, including pretrained model Llama, as well as models that have undergone finetuning, such as ChatGPT, InternLM and ChatGLM2. These models span a range of sizes, including 6B, 7B, 13B, 16B, 65B and so on. For evaluation, ChatGPT and GPT-4 models are tested through the utilization of interfaces, ChatGPT uses 2023.03.15 version and GPT-4 uses 2023.06.13. During testing, the temperature parameter for the pretrained model is set to 0.8 and the top-p parameter is set to 0.95, as referenced from (Huang et al. 2023). For other models, the temperature parameter is set to 0 and the top-p parameter is set to 1. A detailed description of these LLMs can be found in the Appendix.

Error Analysis We categorized the common errors of LLMs into three types, with examples provided in Figure 5. The first type is errors caused by training corpus. The Llama mainly relies on English pre-training corpus, tends to produce illogical answers. However, CN-Alpaca which utilizes more Chinese training corpus performs notably better. The second type is problem-solving errors, it's a consensus that LLMs lack sufficient inferential ability. The third type is gibberish outputs common to many models, the extent of these problems varies as seen in Figure 5 for Baichuan. We will provide more cases in the Appendix.

Comparison of Models While all of it are capable of processing Chinese input, their proficiency levels differ. The Llama model, due to its relatively small Chinese corpus, exhibits relatively weak proficiency in Chinese, as the first error type mentioned above. On the other hand, ChatGPT and GPT-4, benefiting from a certain amount of Chinese corpus in their training data, showcase relatively better Chinese proficiency. The remaining models are trained using substantial amounts of Chinese corpus. Among the models we tested, only Llama65B are pretrained models. As shown in Table 1, it consistently fall behind the pretrained+finetuning model across all evaluated criteria, even though they have a significantly larger parameter count than the latter. These findings effectively demonstrate the significance of the finetuning process. Both effective pretrained and finetuning are crucial components in achieving optimal performance in domain-specific text comprehension. Additionally, we noticed that for the finetuning model, the number of parameters does not necessarily reflect the quality of the results. For example, the accuracy of the InternLM-6B is higher than that of CN-Alpaca and Baichuan with 13B parameters on many questions. A possible explanation is high-quality data focus on Chinese may lead to better results in Chinese Benchmarks.

Comparison on Evaluation Setup The pretrained model exhibits enhanced accuracy with the implementation of few-shot testing, whereas most finetuning models are incapable of performing stably few-shot learning from demonstrations. Demonstrate the few-shot abilities of the finetuned large models have all deteriorated. But encouragingly, we still observe ChatGPT benefit from few-shot setting in Single, Filling and Sorting. This can be ascribed to the extensive domain knowledge integrated within ChatGPT, equipping it with the capacity to discern and encapsulate the distinctive features inherent in the learning samples (Gu et al. 2023).

Compared with few-shot and few-shot CoT, most models

	Model	GPT-4	ChatGPT	InternLM	ChatGLM2	CN-Alpaca	Baichuan	Moss	Llama
	Params	⊘	⊘	7B	6B	13B	13B	16B	65B
Single	ZS	**70.1** / 61.3	52.3 / 41.6	**68.4 / 48.2**	**57.5 / 44.7**	**44.1 / 32.7**	**41.5 / 29.4**	27.2 / **25.6**	30.0 / 26.1
	FS	⋆ / ⋆	**55.5 / 44.0**	67.1 / 45.3	51.6 / 44.6	30.7 / 26.8	25.1 / 17.0	**28.6** / 25.3	**37.9 / 29.1**
	ZS-CoT	69.1 / **62.6**	51.9 / 41.2	60.7 / 41.6	43.1 / 37.3	40.8 / 27.7	34.7 / 15.9	23.5 / 19.3	22.0 / 19.7
	FS-CoT	⋆ / ⋆	47.4 / 42.3	53.8 / 35.8	44.0 / 32.7	28.6 / 21.9	15.7 / 8.6	24.0 / 16.4	24.8 / 20.5
Single*	ZS	**69.7** / 59.8	**54.4 / 42.7**	**66.3 / 45.8**	**56.2** / 43.1	**41.6 / 31.7**	**40.7 / 29.6**	27.9 / 25.4	29.3 / 25.0
	FS	⋆ / ⋆	**54.4 / 42.7**	64.4 / 42.9	50.9 / **44.6**	32.9 / 27.8	28.4 / 18.6	**29.7 / 26.6**	**36.4 / 28.1**
	ZS-CoT	67.8 / **61.0**	50.0 / 41.8	59.6 / 40.3	42.3 / 37.2	38.8 / 28.3	33.5 / 15.8	26.4 / 21.9	23.2 / 20.8
	FS-CoT	⋆ / ⋆	47.1 / 41.7	52.1 / 35.7	42.8 / 32.7	28.8 / 22.5	17.1 / 7.7	25.1 / 16.6	27.0 / 20.5
Single**	ZS	**62.2** / 52.8	44.5 / **36.5**	**54.5 / 37.2**	**45.9 / 35.8**	21.8 / 21.8	**32.4 / 21.2**	24.7 / **23.3**	**15.2** / 15.7
	FS	⋆ / ⋆	**45.3** / 31.7	54.2 / 33.4	33.8 / 27.6	6.7 / 10.8	14.3 / 12.4	18.5 / 15.5	11.9 / 11.2
	ZS-CoT	59.6 / **53.0**	44.1 / 36.1	49.1 / 30.2	35.3 / 30.9	**26.3 / 24.9**	30.4 / 17.3	**25.0 / 23.3**	14.9 / **16.3**
	FS-CoT	⋆ / ⋆	32.6 / 25.0	41.5 / 26.9	35.9 / 27.0	16.3 / 13.9	15.8 / 8.8	18.3 / 15.6	10.7 / 11.0
Multi	ZS	**47.8** / 30.7	**33.2 / 14.7**	**34.5 / 17.7**	**29.6 / 18.6**	5.2 / **3.7**	10.3 / **7.2**	3.6 / 1.4	4.9 / 3.4
	FS	⋆ / ⋆	30.0 / 13.5	29.5 / 12.5	24.3 / 15.1	**5.7** / 3.6	**13.5** / 5.4	3.6 / 4.3	**12.9 / 8.2**
	ZS-CoT	46.7 / **31.3**	32.0 / 14.0	16.0 / 4.5	23.2 / 17.5	4.3 / 2.1	10.9 / 2.9	2.3 / 0.9	5.2 / 3.3
	FS-CoT	⋆ / ⋆	13.3 / 4.1	13.8 / 5.4	8.8 / 2.7	6.4 / 2.4	5.6 / 0.7	**4.9 / 5.9**	2.7 / 3.5
Multi*	ZS	**48.5** / 29.6	27.4 / 12.3	**31.9 / 17.0**	**31.5 / 17.8**	5.8 / **3.2**	11.6 / **6.9**	**4.1 / 1.7**	2.5 / 2.0
	FS	⋆ / ⋆	27.2 / 12.3	27.5 / 12.1	22.0 / 13.8	5.2 / 4.4	**11.7** / 5.9	3.1 / 2.9	**4.5 / 4.1**
	ZS-CoT	46.6 / **30.6**	**35.3 / 13.7**	17.9 / 4.3	21.5 / 16.3	4.2 / 2.1	9.7 / 3.5	3.1 / 1.5	1.8 / 1.7
	FS-CoT	⋆ / ⋆	10.9 / 3.3	10.2 / 5.1	7.4 / 2.4	**5.9** / 3.0	5.4 / 0.6	3.8 / **5.7**	0.4 / 0.9
Multi**	ZS	**33.7** / 11.4	15.3 / 5.2	17.6 / 7.3	**17.0 / 6.6**	2.6 / **2.0**	**11.9 / 6.6**	**2.7** / 1.0	**0.2** / 0.2
	FS	⋆ / ⋆	10.2 / 3.1	**20.8 / 9.4**	8.5 / 6.0	**1.7** / 1.6	5.3 / 3.2	2.1 / 0.4	0.0 / **0.3**
	ZS-CoT	33.4 / **14.0**	**15.8 / 5.3**	9.9 / 2.4	14.7 / 4.4	3.1 / 1.5	9.5 / 4.0	1.8 / **1.4**	0.1 / 0.1
	FS-CoT	⋆ / ⋆	8.7 / 2.0	8.0 / 2.5	6.7 / 1.7	2.0 / 1.0	6.6 / 0.6	2.1 / 0.4	0.0 / 0.0
Filling	ZS	**41.8 / 10.9**	32.8 / 14.1	30.8 / 10.9	18.5 / 17.1	17.8 / 9.0	**34.8** / 10.5	12.2 / 7.5	14.9 / 9.0
	FS	⋆ / ⋆	**35.9** / 28.5	**31.8** / 12.5	**18.8** / 24.7	**20.4** / 13.4	29.1 / 5.3	11.2 / 7.3	**19.9 / 10.6**
	ZS-CoT	41.3 / **10.9**	32.9 / 18.4	22.9 / **25.8**	16.7 / 21.4	9.3 / 6.5	33.1 / **13.2**	9.7 / 9.0	4.7 / 4.9
	FS-CoT	⋆ / ⋆	25.9 / **31.6**	20.9 / 18.4	17.4 / **34.3**	16.4 / **14.5**	0.4 / 11.3	**13.4 / 24.2**	1.0 / 2.1
Judging	ZS	70.5 / 73.0	**64.7** / 60.7	**65.0 / 63.8**	60.7 / **64.4**	**57.5** / 52.8	**64.0 / 58.2**	51.5 / 50.0	38.6 / 32.3
	FS	⋆ / ⋆	60.2 / 59.7	62.5 / 60.4	**62.1** / 58.5	54.2 / **55.4**	54.5 / 52.4	**56.6 / 57.8**	49.6 / **51.9**
	ZS-CoT	**71.2 / 73.3**	62.8 / **60.8**	62.6 / 62.2	61.0 / 64.3	54.6 / 52.3	60.6 / 55.4	49.9 / 55.5	**50.1** / 47.0
	FS-CoT	⋆ / ⋆	61.3 / 59.9	58.6 / 60.1	61.5 / 63.3	55.1 / 49.9	55.1 / 48.1	52.6 / 57.2	47.3 / 40.7
Sor.	ZS	**51.5** / -	10.6 / -	1.7 / -	**2.2** / -	0.4 / -	**1.7** / -	**1.3** / -	0.4 / -
	ZS-CoT	51.1 / -	**16.0** / -	1.7 / -	1.3 / -	0.4 / -	**1.7** / -	0.0 / -	0.4 / -

Table 1: Models tested in four modes. Arts and science results are separated by /. Bold indicates the best result among the four mode on the same model, subject and question type. * stands for shuffle options, ** stands for adds distractor, - means no questions, ⋆ indicates pending updates(GPT4 token limits). The abbreviations description: Chinese-Alpaca-plus(CN-Alpaca), zero-shot(ZS), few-shot(FS), zero-shot-CoT(ZS-CoT), few-shot-CoT(FS-CoT) and Sorting(Sor.).

Model	Acc	FS	SC ↓	RA	Hal	Red
GPT-4	**29.8**	85.5	**11.3**	**83.7**	97.7	84.6
ChatGPT	24.5	84.9	13.8	82.9	97.4	84.1
InternLM	22.0	**89.7**	12.3	74.0	**98.1**	83.1
Baichuan	17.8	85.5	34.9	77.2	96.5	80.2
ChatGLM2	22.0	85.3	13.5	82.8	97.8	**85.5**

Table 2: Results on CoT evaluation in Mathematic. Faithfulness-Step (FS), Source-Consistency(SC), Reasoning Alignment (RA), Hallucination (Hal), Redundancy (Red).

Knowledge point	Acc	Knowledge point	Acc
Science of Heat	**1.00**	Mechanics	0.93
Mechanical Motion	0.30	Mass and Density	0.17
Knowledge of Force	0.40	Elasticity	0.33
Energy	0.37	Acoustics	0.95
Ultimate Questions	0.00	Falsification Trial	0.00
Science Practice	0.00	Atomic Physics	0.00

Table 3: Accuracy of some Level-1 knowledge points in physics subject, results from ChatGLM2 with zero-shot.

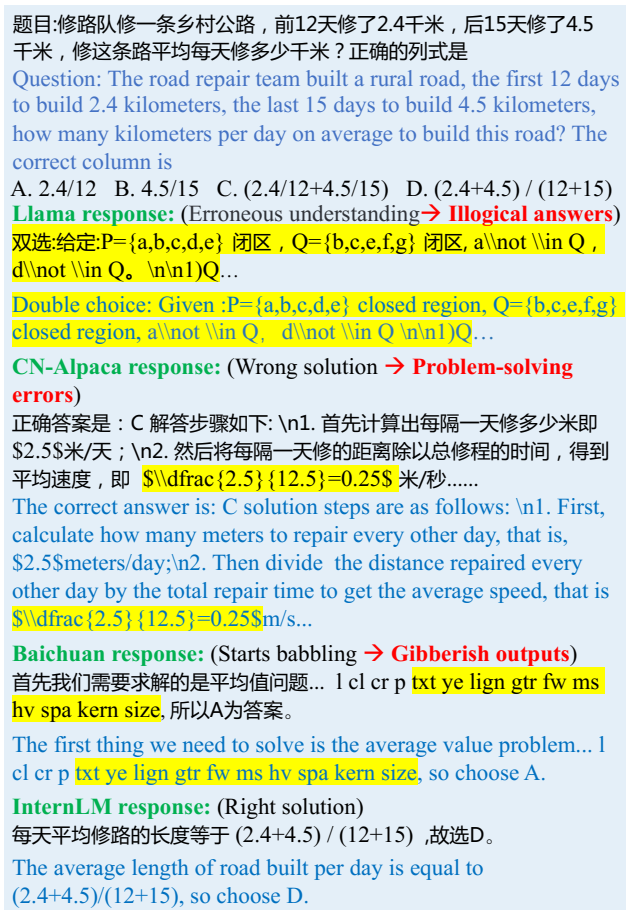

Figure 5: Examples of error question from different LLMs. Due to the length, please zoom in the image for a better view. Yellow indicates the marked errors, red signifies the type of errors.

demonstrate diminished performance, this suggests that CoT prompting does not necessarily improve results for many subjects. The hypothesis for this behavior is that these two instruction-tuned models did not incorporate CoT data during their instruction tuning phase, which may have led to an overfitting of the provided instruction data at the expense of their CoT reasoning abilities (Huang et al. 2023).

Comparison on Data Regarding choice questions, it is evident that shuffling the options has minimal impact on the model's performance, almost within a margin of 1 percentage point. This demonstrates that the models can comprehend the meaning of the questions and possess the ability to reason to find the correct answer from a limited set of simple options. The differences between models may primarily due to variations in their knowledge and reasoning capabilities. However, the introduction of distractors significantly impacts the outcomes of all models, resulting in a significant decline, particularly noticeable in multiple-choice questions. This observation attests to the heightened difficulty posed by this approach, manifesting as reduced model ability when encountering such questions. Through a comparative analysis of Art and Science, we observe that the accuracy of Art questions consistently surpasses that of Science questions by a margin exceeding 10 percentage points. This finding provides evidence that current LLMs excel in memory retention and comprehension, while their logical reasoning abilities remain comparatively deficient.

Comparison on Step-By-Step Reasoning We have selected some models from the few-shot-CoT results to evaluate the inference steps. Note that the evaluation metric for the inference step is only a relative reference score, and its absolute value holds little significance. Based on the observations from Table 2, GPT-4 demonstrated the highest accuracy and achieved the best results on Source-Consistency and Reasoning Alignment, indicating its superior reasoning ability. Reasoning Alignment is the most intuitive metric for evaluating the reasoning steps. It is noteworthy that ChatGLM2 achieved a high alignment score despite its low accuracy. Conversely, InternLM showed high accuracy but low alignment scores. In response to this anomaly, we randomly sampled the results of GPT-4, ChatGLM2 and InternLM, conducted manual subjectivity evaluations for the models' output. It mainly gives a score for the similarity between the analyse and the model inference step. The subjective evaluation scores were 1813, 2235 and 815, respectively, which are consistent with the trend of the Reasoning Alignment score and support the credibility of the evaluation metrics.

Results on Physics Level-1 Knowledge Point We conducted an assessment of the accuracy of some level-1 knowledge points in the junior high school Physics subject, based on the results obtained from ChatGLM2-6B. In Table 3, it is evident that ChatGLM2-6B exhibits a relatively limited proficiency in certain knowledge points, including Ultimate Questions, Falsification Trial, Science Practice and Atomic Physics, all of which have an accuracy score of 0. The accuracy analysis of all knowledge points effectively guides the model toward enhancing its own capabilities. Moreover, the deficient segments can be offset by incorporating pertinent data from knowledge points that exhibit lower accuracy.

Conclusion

We introduced the Knowledge Graph Based Benchmark, a new comprehensive benchmark based on a self-developed knowledge graph that measures how well the overall capacity of LLMs acquire and apply the K12 domain knowledge. Our Benchmark covers 9 subjects covering 584 level-1 knowledge points and 1,989 further fine-grained level-2 knowledge points, with 39,452 different types of questions. We explore 8 different cutting edge LLMs on our benchmark, these models span a range of sizes,including 6B, 7B, 13B, 16B, 65B and so on, containing pretrained models and finetuning models. Among several evaluation setups, different types of questions and scoring step-by-step reasoning, we have extensively explored the issue of incomplete coverage of K12 education topics in LLMs and wish to identify optimization pathways to address certain knowledge gaps.

Acknowledgements

This work was supported by National Key R&D Program of China, under Grant No. 2020AAA0104500.

References

An, C.; Gong, S.; Zhong, M.; Li, M.; Zhang, J.; Kong, L.; and Qiu, X. 2023. L-Eval: Instituting Standardized Evaluation for Long Context Language Models. arXiv:2307.11088.

Bisk, Y.; Zellers, R.; Gao, J.; Choi, Y.; et al. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34:05, 7432–7439.

Chiang, W.-L.; Li, Z.; Lin, Z.; Sheng, Y.; Wu, Z.; Zhang, H.; Zheng, L.; Zhuang, S.; Zhuang, Y.; Gonzalez, J. E.; Stoica, I.; and Xing, E. P. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality.

Cobbe, K.; Kosaraju, V.; Bavarian, M.; Chen, M.; Jun, H.; Kaiser, L.; Plappert, M.; Tworek, J.; Hilton, J.; Nakano, R.; Hesse, C.; and Schulman, J. 2021. Training Verifiers to Solve Math Word Problems. arXiv:2110.14168.

Deng, M.; Tan, B.; Liu, Z.; Xing, E. P.; and Hu, Z. 2021. Compression, transduction, and creation: A unified framework for evaluating natural language generation. *arXiv preprint arXiv:2109.06379*.

Du, Z.; Qian, Y.; Liu, X.; Ding, M.; Qiu, J.; Yang, Z.; and Tang, J. 2022. GLM: General Language Model Pretraining with Autoregressive Blank Infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 320–335.

Gao, T.; Yao, X.; and Chen, D. 2021. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*.

Golovneva, O.; Chen, M.; Poff, S.; Corredor, M.; Zettlemoyer, L.; Fazel-Zarandi, M.; and Celikyilmaz, A. 2022. Roscoe: A suite of metrics for scoring step-by-step reasoning. *arXiv preprint arXiv:2212.07919*.

Gu, Z.; Zhu, X.; Ye, H.; Zhang, L.; Wang, J.; Jiang, S.; Xiong, Z.; Li, Z.; He, Q.; Xu, R.; et al. 2023. Xiezhi: An Ever-Updating Benchmark for Holistic Domain Knowledge Evaluation. *arXiv preprint arXiv:2306.05783*.

Hendrycks, D.; Basart, S.; Kadavath, S.; Mazeika, M.; Arora, A.; Guo, E.; Burns, C.; Puranik, S.; He, H.; Song, D.; et al. 2021a. Measuring coding challenge competence with apps. *arXiv preprint arXiv:2105.09938*.

Hendrycks, D.; Basart, S.; Kadavath, S.; Mazeika, M.; Arora, A.; Guo, E.; Burns, C.; Puranik, S.; He, H.; Song, D.; et al. 2021b. Measuring coding challenge competence with apps. *arXiv preprint arXiv:2105.09938*.

Huang, L.; Bras, R. L.; Bhagavatula, C.; and Choi, Y. 2019. Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. *arXiv preprint arXiv:1909.00277*.

Huang, Y.; Bai, Y.; Zhu, Z.; Zhang, J.; Zhang, J.; Su, T.; Liu, J.; Lv, C.; Zhang, Y.; Lei, J.; et al. 2023. C-Eval: A Multi-Level Multi-Discipline Chinese Evaluation Suite for Foundation Models. *arXiv preprint arXiv:2305.08322*.

Ji, Y.; Deng, Y.; Gong, Y.; Peng, Y.; Niu, Q.; Zhang, L.; Ma, B.; and Li, X. 2023. Exploring the Impact of Instruction Data Scaling on Large Language Models: An Empirical Study on Real-World Use Cases. *arXiv preprint arXiv:2303.14742*.

Joshi, M.; Choi, E.; Weld, D. S.; and Zettlemoyer, L. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*.

Kojima, T.; Gu, S. S.; Reid, M.; Matsuo, Y.; and Iwasawa, Y. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35: 22199–22213.

Lai, G.; Xie, Q.; Liu, H.; Yang, Y.; and Hovy, E. 2017. RACE: Large-scale ReAding Comprehension Dataset From Examinations. arXiv:1704.04683.

Laurer, M.; van Atteveldt, W.; Casas, A.; and Welbers, K. 2022. Less Annotating, More Classifying: Addressing the Data Scarcity Issue of Supervised Machine Learning with Deep Transfer Learning and BERT-NLI. *Political Analysis*, 1–33.

Li, H.; Zhang, Y.; Koto, F.; Yang, Y.; Zhao, H.; Gong, Y.; Duan, N.; and Baldwin, T. 2023. CMMLU: Measuring massive multitask language understanding in Chinese. arXiv:2306.09212.

Liang, P.; Bommasani, R.; Lee, T.; Tsipras, D.; Soylu, D.; Yasunaga, M.; Zhang, Y.; Narayanan, D.; Wu, Y.; Kumar, A.; et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Lin, S.; Hilton, J.; and Evans, O. 2021. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*.

Liu, C.; Jin, R.; Ren, Y.; Yu, L.; Dong, T.; Peng, X.; Zhang, S.; Peng, J.; Zhang, P.; Lyu, Q.; et al. 2023. M3KE: A Massive Multi-Level Multi-Subject Knowledge Evaluation Benchmark for Chinese Large Language Models. *arXiv preprint arXiv:2305.10263*.

Nye, M.; Andreassen, A. J.; Gur-Ari, G.; Michalewski, H.; Austin, J.; Bieber, D.; Dohan, D.; Lewkowycz, A.; Bosma, M.; Luan, D.; Sutton, C.; and Odena, A. 2021. Show Your Work: Scratchpads for Intermediate Computation with Language Models. arXiv:2112.00114.

OpenAI. 2023a. ChatGPT: Optimizing Language Models for Dialogue. https://openai.com/blog/chatgpt.

OpenAI. 2023b. GPT-4 Technical Report. arXiv:2303.08774.

Paperno, D.; Kruszewski, G.; Lazaridou, A.; Pham, Q. N.; Bernardi, R.; Pezzelle, S.; Baroni, M.; Boleda, G.; and Fernández, R. 2016. The LAMBADA dataset: Word prediction requiring a broad discourse context. *arXiv preprint arXiv:1606.06031*.

Scao, T. L.; Fan, A.; Akiki, C.; Pavlick, E.; Ilić, S.; Hesslow, D.; Castagné, R.; Luccioni, A. S.; Yvon, F.; Gallé, M.; et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Srivastava, A.; Rastogi, A.; Rao, A.; Shoeb, A. A. M.; Abid, A.; Fisch, A.; Brown, A. R.; Santoro, A.; Gupta, A.; Garriga-Alonso, A.; et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Sun, T.; Zhang, X.; He, Z.; Li, P.; Cheng, Q.; Yan, H.; Liu, X.; Shao, Y.; Tang, Q.; Zhao, X.; Chen, K.; Zheng, Y.; Zhou, Z.; Li, R.; Zhan, J.; Zhou, Y.; Li, L.; Yang, X.; Wu, L.; Yin, Z.; Huang, X.; and Qiu, X. 2023. MOSS: Training Conversational Language Models from Synthetic Data.

Taori, R.; Gulrajani, I.; Zhang, T.; Dubois, Y.; Li, X.; Guestrin, C.; Liang, P.; and Hashimoto, T. B. 2023. Stanford alpaca: An instruction-following llama model.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Wang, A.; Pruksachatkun, Y.; Nangia, N.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; and Bowman, S. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

Wang, A.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; and Bowman, S. R. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Chi, E.; Le, Q.; and Zhou, D. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.; Le, Q.; and Zhou, D. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. arXiv:2201.11903.

Xu, C.; Guo, D.; Duan, N.; and McAuley, J. 2023. Baize: An Open-Source Chat Model with Parameter-Efficient Tuning on Self-Chat Data. *arXiv preprint arXiv:2304.01196*.

Xu, L.; Hu, H.; Zhang, X.; Li, L.; Cao, C.; Li, Y.; Xu, Y.; Sun, K.; Yu, D.; Yu, C.; et al. 2020. CLUE: A Chinese language understanding evaluation benchmark. *arXiv preprint arXiv:2004.05986*.

Xu, L.; Lu, X.; Yuan, C.; Zhang, X.; Xu, H.; Yuan, H.; Wei, G.; Pan, X.; Tian, X.; Qin, L.; et al. 2021. Fewclue: A chinese few-shot learning evaluation benchmark. *arXiv preprint arXiv:2107.07498*.

Yang, L.; Zhang, S.; Qin, L.; Li, Y.; Wang, Y.; Liu, H.; Wang, J.; Xie, X.; and Zhang, Y. 2022. GLUE-X: Evaluating Natural Language Understanding Models from an Out-of-distribution Generalization Perspective. *arXiv preprint arXiv:2211.08073*.

Yuan, W.; Neubig, G.; and Liu, P. 2021. Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34: 27263–27277.

Zeng, H. 2023. Measuring Massive Multitask Chinese Understanding. *arXiv preprint arXiv:2304.12986*.

Zhang, X.; Li, C.; Zong, Y.; Ying, Z.; He, L.; and Qiu, X. 2023. Evaluating the Performance of Large Language Models on GAOKAO Benchmark. arXiv:2305.12474.

Zhao, W.; Shang, M.; Liu, Y.; Wang, L.; and Liu, J. 2020. Ape210K: A Large-Scale and Template-Rich Dataset of Math Word Problems. arXiv:2009.11506.

Zhong, W.; Cui, R.; Guo, Y.; Liang, Y.; Lu, S.; Wang, Y.; Saied, A.; Chen, W.; and Duan, N. 2023. AGIEval: A Human-Centric Benchmark for Evaluating Foundation Models. *arXiv preprint arXiv:2304.06364*.