# Xiezhi: An Ever-Updating Benchmark for Holistic Domain Knowledge Evaluation

**Zhouhong Gu[1]*, Xiaoxuan Zhu[1]*, Haoning Ye[1], Lin Zhang[1], Jianchen Wang[1], Yixin Zhu[1], Sihang Jiang[1], Zhuozhi Xiong[1], Zihan Li[1], Weijie Wu[1], Qianyu He[1], Rui Xu[1], Wenhao Huang[1], Jingping Liu[2], Zili Wang[3], Shusen Wang[3], Weiguo Zheng[4], Hongwei Feng[1]†, Yanghua Xiao[1,5]†,**

[1]Shanghai Key Laboratory of Data Science, School of Computer Science, Fudan University, China
[2]School of Information Science and Engineering, East China University of Science and Technology
[3]Xiaohongshu Inc.
[4]School of Data Science, Fudan University
[5]Fudan-Aishu Cognitive Intelligence Joint Research Center
{zhgu22, xxzhu22}@m.fudan.edu.cn, {hwfeng, shawyh}@fudan.edu.cn

## Abstract

New Natural Langauge Process (NLP) benchmarks are urgently needed to align with the rapid development of large language models (LLMs). We present Xiezhi, the most comprehensive evaluation suite designed to assess holistic domain knowledge. Xiezhi comprises multiple-choice questions across 516 diverse disciplines ranging from 13 different subjects with 249,587 questions and accompanied by Xiezhi-Specialty with 14,041 questions and Xiezhi-Interdiscipline with 10,746 questions. We conduct evaluation of the 47 cutting-edge LLMs on Xiezhi. Results indicate that LLMs exceed average performance of humans in science, engineering, agronomy, medicine, and art, but fall short in economics, jurisprudence, pedagogy, literature, history, and management. All the evaluation code and data are open sourced in https://github.com/MikeGu721/XiezhiBenchmark

## Introduction

Domain knowledge encompasses an in-depth comprehension of the world, necessitating the cultivation of various cognitive skills, such as memorization, abstraction, logical thinking, reasoning, and imagination. Human has exhibited unparalleled proficiency in domain knowledge, far exceeding any machine learning models in a long time. Nevertheless, recent advancements in Large Language Models (LLMs), including Bloom (Scao et al. 2022), Llama (Touvron et al. 2023), ChatGLM (Du et al. 2022), GPT4 (OpenAI 2023b; Bubeck et al. 2023) and so many other models, have shown remarkable capabilities in domain text understanding (Wei et al. 2022). It is time to propose more comprehensive and more prospective evaluations than before to explore whether LLMs have actually acquired knowledge, or just acquired a better imitation ability (Srivastava et al. 2022).

Constructing benchmarks is crucial for automatic evaluation as benchmarks facilitate efficient, systematic, and scalable comparisons among models. However, as LLMs continue to grow in size and complexity, they exhibit outstanding performance across a wide range of domain-specific tasks. This makes even the newly released benchmarks like MMLU (Hendrycks et al. 2021), BIG-bench (Srivastava et al. 2022) or HELM (Liang et al. 2022) all lag behind the capabilities of the LLMs quickily (Suzgun et al. 2022).

Considering LLMs' performance, we conclude that the benchmark used to evaluate LLMs should meet the following needs: (1) **Needs to cover more tasks (Srivastava et al. 2022)**: Cutting-edge LLMs have integrated multiple capabilities into unified Text-to-Text transformer models (Raffel et al. 2020). Therefore, the evaluation of LLMs should focus on abilities in multiple tasks. (2) **Needs to manifest the disparities among LLMs (Huang et al. 2023)**: Considering the emergent capacity of the models (Wei et al. 2022), it is likely that the SoTA LLMs by learning knowledge in different domains, now have a certain level of performance in all domains. To accurately evaluate the distinctions of LLMs with varying capacities, the benchmark should consider breaking down the evaluation dimensions into more detailed categories. This will allow for a more precise assessment of each model's capabilities and provide valuable insights into their relative strengths and weaknesses. (3) **Needs to go ahead of the training set (Bubeck et al. 2023)**: As LLMs are trained on increasingly extensive corpora, newly released benchmarks may become part of the LLMs' training data much sooner than before. A prerequisite for effective evaluation is to ensure that the benchmarks are fresher than the training data used by LLMs.

In light of the aforementioned needs, we propose a comprehensive, multi-disciplinary, auto-updating benchmark for domain knowledge evaluation. We call this benchmark Xiezhi, named after a mythical creature that symbolizes fairness and judgement. Xiezhi consists of 249587 questions with 516 disciplines, ranging from 13 different categories: philosophy, economics, law, education, literature, history, natural sciences, engineering, agriculture, medicine, military science, management, and arts. These 516 disciplines are derived from the Chinese Disciplinary Taxonomy, a comprehensive hierarchical classification system of domain knowledge pro-

*Equal Contribution

†Corresponding Authors

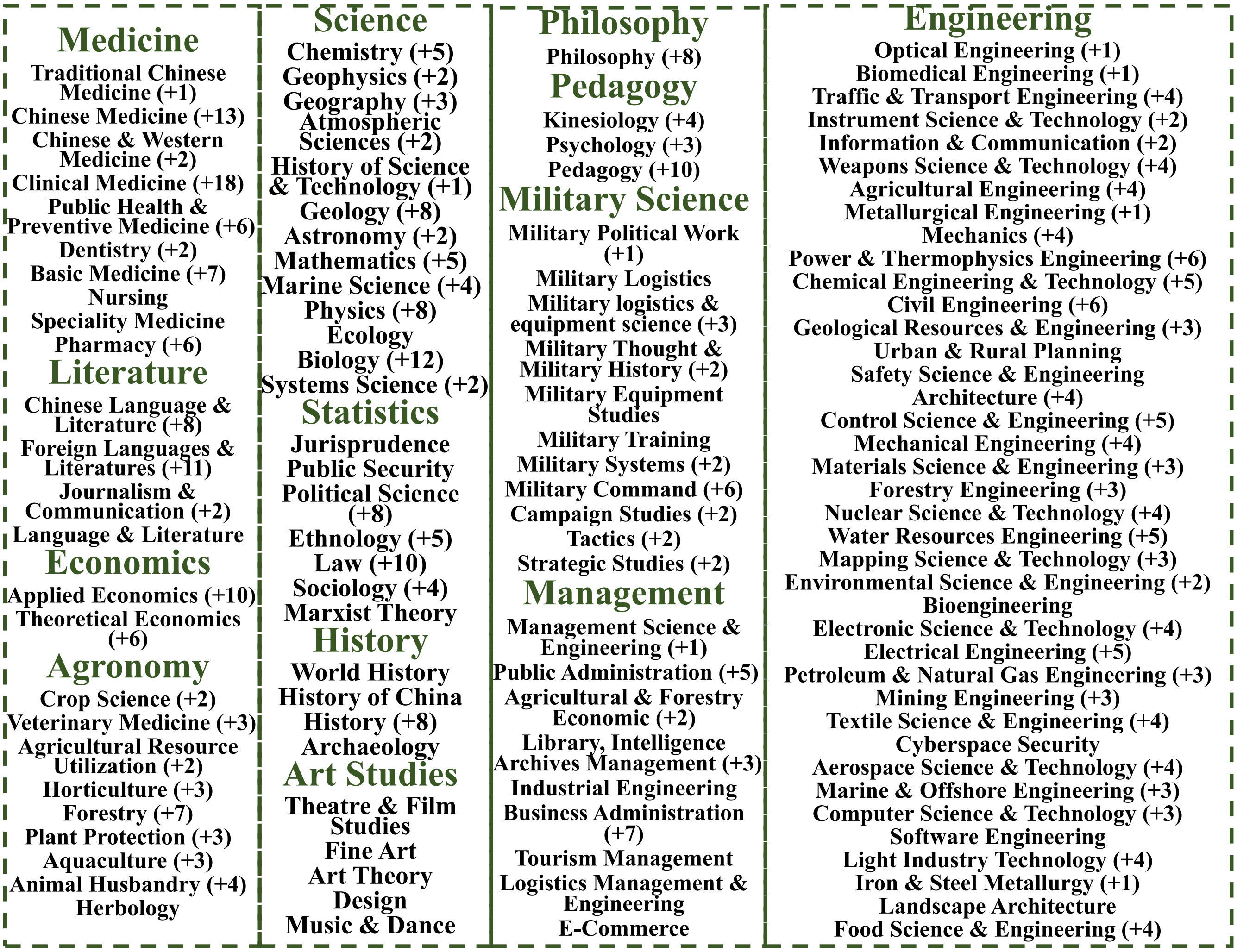
Medicine
Traditional Chinese Medicine (+1)
Chinese Medicine (+13)
Chinese & Western Medicine (+2)
Clinical Medicine (+18)
Public Health & Preventive Medicine (+6)
Dentistry (+2)
Basic Medicine (+7)
Nursing
Speciality Medicine
Pharmacy (+6)
Literature
Chinese Language & Literature (+8)
Foreign Languages & Literatures (+11)
Journalism & Communication (+2)
Language & Literature
Economics
Applied Economics (+10)
Theoretical Economics (+6)
Agronomy
Crop Science (+2)
Veterinary Medicine (+3)
Agricultural Resource Utilization (+2)
Horticulture (+3)
Forestry (+7)
Plant Protection (+3)
Aquaculture (+3)
Animal Husbandry (+4)
Herbology
Science
Chemistry (+5)
Geophysics (+2)
Geography (+3)
Atmospheric Sciences (+2)
History of Science & Technology (+1)
Geology (+8)
Astronomy (+2)
Mathematics (+5)
Marine Science (+4)
Physics (+8)
Ecology
Biology (+12)
Systems Science (+2)
Statistics
Jurisprudence
Public Security
Political Science (+8)
Ethnology (+5)
Law (+10)
Sociology (+4)
Marxist Theory
History
World History
History of China
History (+8)
Archaeology
Art Studies
Theatre & Film Studies
Fine Art
Art Theory
Design
Music & Dance
Philosophy
Philosophy (+8)
Pedagogy
Kinesiology (+4)
Psychology (+3)
Pedagogy (+10)
Military Science
Military Political Work (+1)
Military Logistics
Military logistics & equipment science (+3)
Military Thought & Military History (+2)
Military Equipment Studies
Military Training
Military Systems (+2)
Military Command (+6)
Campaign Studies (+2)
Tactics (+2)
Strategic Studies (+2)
Management
Management Science & Engineering (+1)
Public Administration (+5)
Agricultural & Forestry Economic (+2)
Library, Intelligence Archives Management (+3)
Industrial Engineering
Business Administration (+7)
Tourism Management
Logistics Management & Engineering
E-Commerce
Engineering
Optical Engineering (+1)
Biomedical Engineering (+1)
Traffic & Transport Engineering (+4)
Instrument Science & Technology (+2)
Information & Communication (+2)
Weapons Science & Technology (+4)
Agricultural Engineering (+4)
Metallurgical Engineering (+1)
Mechanics (+4)
Power & Thermophysics Engineering (+6)
Chemical Engineering & Technology (+5)
Civil Engineering (+6)
Geological Resources & Engineering (+3)
Urban & Rural Planning
Safety Science & Engineering
Architecture (+4)
Control Science & Engineering (+5)
Mechanical Engineering (+4)
Materials Science & Engineering (+3)
Forestry Engineering (+3)
Nuclear Science & Technology (+4)
Water Resources Engineering (+5)
Mapping Science & Technology (+3)
Environmental Science & Engineering (+2)
Bioengineering
Electronic Science & Technology (+4)
Electrical Engineering (+5)
Petroleum & Natural Gas Engineering (+3)
Mining Engineering (+3)
Textile Science & Engineering (+4)
Cyberspace Security
Aerospace Science & Technology (+4)
Marine & Offshore Engineering (+3)
Computer Science & Technology (+3)
Software Engineering
Light Industry Technology (+4)
Iron & Steel Metallurgy (+1)
Landscape Architecture
Food Science & Engineering (+4)

Figure 1: In Chinese mythology, the Xiezhi is a legendary creature known for its ability to discern right from wrong and uphold justice. Xiezhi Benchmark encompasses 13 distinct disciplinary categories, 118 sub-disciplines, and 385 further fine-grained disciplines, aiming to provide an extensive domain taxonomy and benchmark for fair, effective, and comprehensive domain evaluation. The number adjacent to the first-level discipline signifies the number of second-level disciplines that are further divided in Chinese discipline taxonomy.

posed by the Chinese Ministry of Education and widely acknowledged in China. We manually selected and annotated 20,124 questions from the Chinese Graduate Entrance Examination covering these 516 labels to form the Xiezhi-Meta dataset. Xiezhi-Meta is used to train an annotation model capable of estimating the relevance between questions and disciplinary labels. The annotation model subsequently tag disciplinary labels to 170k multiple-choice questions originating from diverse examinations, along with 80k multiple-choice questions auto-generated from academic surveys. To facilitate the usage of Xiezhi and align with the inclination that “consolidate increasing capabilities into single LLMs”, we also present Xiezhi-Specialty and Xiezhi-Interdiscipline in both Chinese and English verision, consisting of 14,041 and 10,746 respectively more balanced, less sensitive, and less China-centric questions. Xiezhi-Specialty encompasses questions solvable using knowledge from a single domain, while Xiezhi-Interdiscipline incorporates questions necessitating knowledge from multiple domains for resolution.

To give more precise evaluation results, we propose a new evaluation setting in this paper. We set 50 options for each multiple-choice question, as previous researchers use only 4 options, resulting in significantly reducing the accuracy of random guessing and thus better revealing the model’s real capabilities. We rank all options of each model in generation probability, as previous researchers use instructions to query the choice made by each model, to avoid inaccurate evaluations due to model’s inability in answering multiple-choice questions or errors in the generated content extraction.

To provide a detailed analysis of current development status of LLMs, as well as to demonstrate the effectiveness of the Xiezhi Benchmark and our proposed evaluation process, we conduct experiments on 47 famous LLMs across four benchmarks proposed in different works in our evaluation setting. The experiments are conducted under in 0-shot, 1-shot, 3-shot demonstration setting, which is using small number of examples to demonstrate how to solve a question, with all LLMs being evaluated on both Chinese and English versions of Xiezhi. This enables us to analyze the LLM results based on their optimal performance. Results show that the best-performing LLMs, when tested via multiple-choice questions, have surpassed the level of average practitioners in science, engineering, agronomy, and medicine in multiple-choice form of . But humans still greatly outperform all LLMs in domains of economics, jurisprudence, pedagogy, literature, history, and management. We also examined the differences in performance of various LLMs across different benchmarks. Compared to existing knowledge evaluation benchmarks, Xiezhi covers the broadest range of domains, incorporates the highest quantity of questions, and consists of the most current data. As shown in our experiments, due to the vast diversity of knowledge domains covered in Xiezhi and its fifty-to-one evaluation method, even marginal improvements in any aspect of a model can be accurately assessed. As such, it is most proficient in discerning the capability differences among various LMs, spanning from GPT-4 to LLMs with only 560M parameters. Consequently, it serves as the most appropriate benchmark for evaluating LLMs of differing competencies.

## Related Works

### Large Language Models

Recently, various companies released their LLMs, such as BARD, ERNIE Bot, Bloom (Scao et al. 2022), pythia (Biderman et al. 2023), Llama (Touvron et al. 2023), Claude, ChatGPT (OpenAI 2023a), GPT-4 (OpenAI 2023b), and ChatGLM (Du et al. 2022). Apart from their outstanding performance on trained tasks, researchers have also discovered that they emerge to have strong performance on many unseen tasks (Zhou et al. 2023; Chung et al. 2022). Consequently, the evaluation of LLMs' capabilities should focus more on a wide range of tasks over numerous diverse domains and contain samples with different difficulty levels.

The development of LLMs has spurred the growth of a series of small-scale conversational LLMs, such as Alpaca (Taori et al. 2023), Vicuna (Chiang et al. 2023), H2Ogpt (H2O.ai 2023), and Moss (Sun et al. 2023). Most of these small conversational LLMs are fine-tuned based on existing pre-trained LLMs through high-quality dialog data generated from LLMs (Ji et al. 2023; Xu et al. 2023) by parameter-efficient tuning methods (Hu et al. 2021, 2023). In order to achieve excellent performance, these models continuously acquire the latest data from the internet, and their iteration speed is much faster than LLMs. Any new benchmark will quickly become outdated as it is incorporated into the model's training data.

### Benchmarks for Knowledge Evaluation

A number of studies concentrate on assessing a model's knowledge and reasoning ability. Certain works, including HellaSwag (Zellers et al. 2019), Physical IQA (Bisk et al. 2020), and CosmosQA (Huang et al. 2019), focus on evaluating the understanding of LLMs' commonsense knowledge. Meanwhile, other research, such as MMLU (Hendrycks et al. 2021), AGI-Eval (Zhong et al. 2023), MMCU (Zeng 2023), C-Eval (Huang et al. 2023), M3KE (Liu et al. 2023), LexTreme (Niklaus et al. 2023), Big-Bench (Srivastava et al. 2022) and BIG-Bench-Hard (Suzgun et al. 2022) target at evaluating the models' proficiency in domain knowledge. However, whether these benchmarks provide effective evaluations for all language models remains debatable. This is because only LLMs with super abilities show disparities on their datasets, while small LLMs only perform at a level close to random guessing, leading to different evaluation researches having different or even contradictory results on small LLMs (Huang et al. 2023; Li et al. 2023). Furthermore, as the training corpora for models become increasingly larger, these benchmarks might lose their evaluative significance shortly after they are proposed, due to their incorporation into the training sets of LLMs.

Moreover, the rise of the generative LLMs presents its own difficulties in evaluation (Sai, Mohankumar, and Khapra 2022). Beginning with MMLU (Hendrycks et al. 2021), numerous works have proposed to use of multiple-choice questions to assess generative models. Recently, a variety of evaluation studies, such as SuperClue [1], employed an identical prompt to query all LLMs and do extraction to obtain the choice made by these LLMs. This approach requires models to have strong abilities in instruction understanding especially in multiple-choice answering, as many LLMs are unable to meet that needs, leading to unfair evaluation results.

## Xiezhi Benchmark

### Chinese Discipline Taxonomy

Chinese Discipline Taxonomy, developed by the Chinese Ministry of Education, organizes disciplines of different domains in college education. The taxonomy divides all domains into different disciplines categories and various levels of disciplines. The meanings of these levels are as follows:

**Discipline Categories**: This is the highest level of discipline taxonomy, divided according to the nature, characteristics of subjects. There are 14 subject categories in Chinese Discipline Taxonomy, including philosophy, economics, law, education, literature, history, science, engineering, agriculture, medicine, military science, management, art, and Inter-discipline.

**First-level disciplines**: A discipline category is divided into numerous first-level disciplines, each possessing relatively independent research content. For example, the "Economics" category is divided into first-level disciplines "Applied Economics" and "Theoretical Economics", and "Art Studies" consist of "Theatre & File Studies", "Fine Art" and so on.

[1] https://github.com/CLUEbenchmark/SuperCLUE

**Second-level disciplines**: These disciplines represent more subdivided areas of study or topics within the first-level discipline. For example, within the first-level discipline of "Applied Economics", further divisions include "Financial Markets", "Banking", "Insurance" and many other second-level disciplines.

As shown in Fig. 1, Xiezhi Benchmark consists of a total of 13 disciplinary categories, 118 first-level disciplines, and 385 second-level disciplines as question labels. The detailed information on the disciplines and the question amount used in Xiezhi Benchmark is listed in Tab. **??** in Appendix.

## Dataset Construction

**Data collection:** Xiezhi consists of 249,587 questions from mainly two different sources. The first category includes nearly 170k multiple-choice questions collected from six different examinations in China: *elementary school exams*, *middle school entrance exams*, *college entrance exams*, *undergraduate exams*, *graduate entrance exams*, and *adult education exams*. These questions are all open sourced and many Chinese knowledge evaluation dataset have employed these questions (Huang et al. 2023; Liu et al. 2023). The second category comprises of nearly 80k multiple choice questions generated from Chinese open-source academic surveys or reviews, which is a result come from our auto updating method.

**Auto Updating:** Our auto-updating method comprises three primary components: the construction of Xiezhi-Meta dataset, the generation of questions from open academic documents, and the automated annotation process.

**Xiezhi-Meta:** We annotated 20,124 questions collected from the Graduate Entrance Examination to form the meta version of Xiezhi through both manual efforts and chatGPT. The aim of annotation is to remove unanswerable questions and to tag each question with as many disciplines as possible.

We first used ChatGPT to tag each question with first or second-level disciplines in Chinese. In the process of tagging, we construct a prompt by concatenating the description of a question with its options, answers, and exam information with the description of each discipline to increase chatGPT's understanding of the question so that the question could be better tagged. The prompts we used is listed in Appendix Prompt, and the detail of the annotation process is described in Appendix Mannual Annotation.

**Question Generation:** Xiezhi comprises nearly 80k multiple-choice questions generated from academic surveys, as they frequently encompass well-established domain knowledge. We select Chinese academic papers across all disciplines that incorporate the terms "survey" or "review" in their titles. Subsequently, we extract several longest sentences from these surveys, which typically are the introductory sentences that contain comprehensive descriptive information pertinent to a particular field of knowledge. We identify keywords using the OpenNER method (Zhu et al. 2019) from these sentences, which are then masked to formulate the questions. To assemble the set of options for each question, the answers to all other questions in Xiezhi were sampled and combined with the standard answers for each respective question.

**Auto Annotation:** The objectives of auto annotation include the elimination of unanswerable questions and the assignment of relevant discipline labels to each question. For unanswerable questions, we extracted keywords from the Xiezhi-Meta, such as "as shown in the figure below" or "as listed in the table" and so on, and exclude questions that contain any of these keywords from collected data. We use ChatGPT and an annotation model trained by Xiezhi-Meta to do the discipline labels tagging. The annotation model, which is based on llama-7B, is used to tag coarse-grained discipline labels (The Discipline Categories in this paper) to the questions. Based on the tagged coarse-grained labels, we employ ChatGPT to assign more fine-grained labels (First and Second-level discipline labels) to the questions, in a similar manner to the labeling of Xiezhi-Meta. The detail about the training process of the annotation model and the performance of the auto annotation process is described in Appendix Auto Annotator.

**Xiezhi-Specialty & Xiezhi-Interdiscipline:** To ensure the validity of the evaluation results, we further propose two additional datasets, Xiezhi-Specialty and Xiezhi-Interdiscipline in both Chinese and English version. The trajectory of LLM development tends to consolidate multiple capabilities within individual LLMs, which may consequently yield unanticipated interdisciplinary problem-solving proficiencies. The division of Xiezhi into the Specialty and Interdiscipline datasets is designed to correspond with this evolving trend. These datasets are derived from the original Xiezhi Benchmark with the exclusion of some sensitive questions (e.g., military science) and deeply Chinese-centric questions (e.g., Literary Chinese QA, ancient Chinese poetry completion). Based on a balanced sampling strategy, Xiezhi-Specialty is constructed by selecting questions involved in 3 disciplines or less, while Xiezhi-Interdiscipline includes questions tagged by 4 disciplines or more. The down-right of Fig. 4 presents an instance of the Xiezhi-Specialty, while an instance of the Xiezhi-Interdiscipline is depicted in top-right of Fig. 4. The process of translation and annotation is delineated in Appendix Manual Annotation. Furthermore, Appendix Bias, Ethical Problems and Social Impact comprehensively discusses potential ethical challenges and our effort undertaken to mitigate them.

# Experiments

## Setup

**Models&Device:** We conducted experiments on 47 cutting-edge LLMs, the detailed descriptions of all tested LLMs are listed in Tab **??** in Appendix. Our experiments cover 45 open-source LLMs based on eight different base models: *bloom, llama, moss, pythia, gpt-neox, stablelm, chatGLM* and *falcon*. Considering the legal issues, we only show the results of two publicly recognized API-based LLMs, ChatGPT and GPT-4. Our experiment was carried out on a DGX Station with 8 80G memory Tesla A100.

**More options:** All tested LLMs need to choose the best-fit answer from 50 options for each question. Each question is set up with 3 confusing options in addition to the correct answer, and another 46 options are randomly sampled from

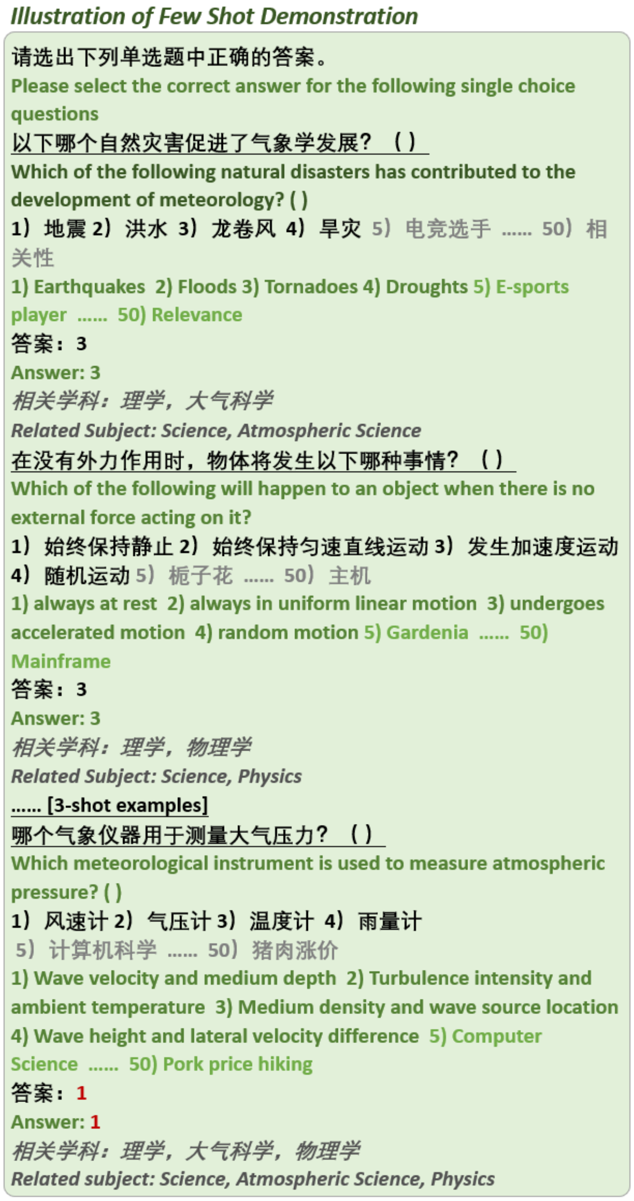

*Illustration of Few Shot Demonstration*

请选出下列单选题中正确的答案。
Please select the correct answer for the following single choice questions
以下哪个自然灾害促进了气象学发展？（）
Which of the following natural disasters has contributed to the development of meteorology? ( )
1）地震 2）洪水 3）龙卷风 4）旱灾 5）电竞选手 ...... 50）相关性
1) Earthquakes 2) Floods 3) Tornadoes 4) Droughts 5) E-sports player ...... 50) Relevance
答案：3
Answer: 3
*相关学科：理学，大气科学*
*Related Subject: Science, Atmospheric Science*
在没有外力作用时，物体将发生以下哪种事情？（）
Which of the following will happen to an object when there is no external force acting on it?
1）始终保持静止 2）始终保持匀速直线运动 3）发生加速度运动 4）随机运动 5）栀子花 ...... 50）主机
1) always at rest 2) always in uniform linear motion 3) undergoes accelerated motion 4) random motion 5) Gardenia ...... 50) Mainframe
答案：3
Answer: 3
*相关学科：理学，物理学*
*Related Subject: Science, Physics*
...... [3-shot examples]
哪个气象仪器用于测量大气压力？（）
Which meteorological instrument is used to measure atmospheric pressure? ( )
1）风速计 2）气压计 3）温度计 4）雨量计
5）计算机科学 ...... 50）猪肉涨价
1) Wave velocity and medium depth 2) Turbulence intensity and ambient temperature 3) Medium density and wave source location 4) Wave height and lateral velocity difference 5) Computer Science ...... 50) Pork price hiking
答案：1
Answer: 1
*相关学科：理学，大气科学，物理学*
*Related subject: Science, Atmospheric Science, Physics*

Figure 2: Examples of a 3-shot evaluation with Xiezhi-Interdiscipline, a question from Xiezhi-Interdiscipline and a question from Xiezhi-Specialty.

all options in all questions in Xiezhi. We obtain options from questions that have different discipline categories and select options that do not have any identical characters (for Chinese) or identical 4-gram characters (for English) to the ground truth. It is worth noting that it is possible to use WordNet, open source synonym databases, or other word construction methods to generate more confusing options. However, our experiments show that the performance of all LLMs declined dramatically when the number of options increased, even when using so many non-confusing options. This achieves our goal of exacerbating the performance gap between LLMs through new experimental settings and also shows that the traditional 4-choice setting has room for improvement.

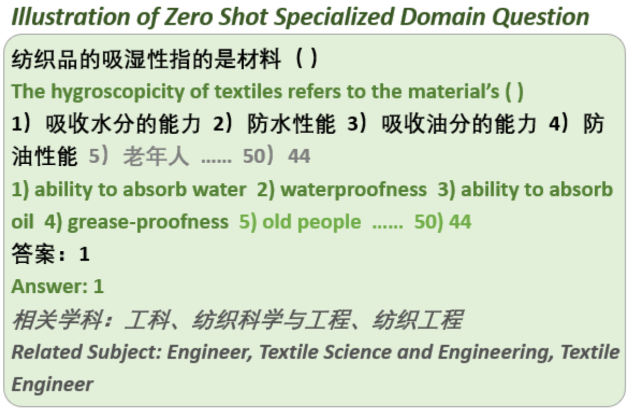

*Illustration of Zero Shot Specialized Domain Question*

纺织品的吸湿性指的是材料（）
The hygroscopicity of textiles refers to the material's ( )
1）吸收水分的能力 2）防水性能 3）吸收油分的能力 4）防油性能 5）老年人 ...... 50）44
1) ability to absorb water 2) waterproofness 3) ability to absorb oil 4) grease-proofness 5) old people ...... 50) 44
答案：1
Answer: 1
*相关学科：工科、纺织科学与工程、纺织工程*
*Related Subject: Engineer, Textile Science and Engineering, Textile Engineer*

Figure 3: Examples of a 3-shot evaluation with Xiezhi-Interdiscipline, a question from Xiezhi-Interdiscipline and a question from Xiezhi-Specialty.

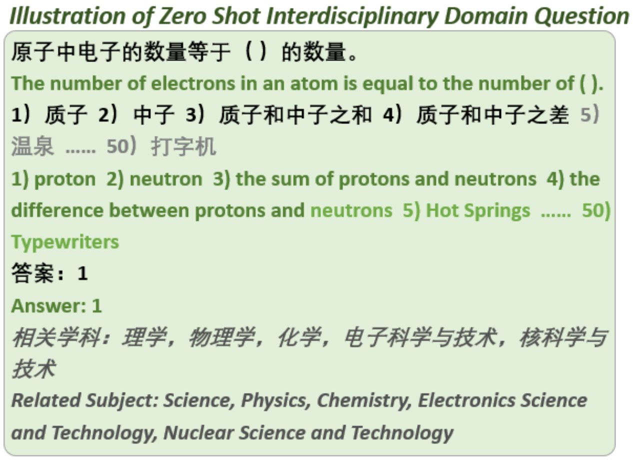

*Illustration of Zero Shot Interdisciplinary Domain Question*

原子中电子的数量等于（）的数量。
The number of electrons in an atom is equal to the number of ( ).
1）质子 2）中子 3）质子和中子之和 4）质子和中子之差 5）温泉 ...... 50）打字机
1) proton 2) neutron 3) the sum of protons and neutrons 4) the difference between protons and neutrons 5) Hot Springs ...... 50) Typewriters
答案：1
Answer: 1
*相关学科：理学，物理学，化学，电子科学与技术，核科学与技术*
*Related Subject: Science, Physics, Chemistry, Electronics Science and Technology, Nuclear Science and Technology*

Figure 4: Examples of a 3-shot evaluation with Xiezhi-Interdiscipline, a question from Xiezhi-Interdiscipline and a question from Xiezhi-Specialty.

**Few-Shot Demonstration:** Additionally, we aim to test the LLMs' understanding of demonstrations. Therefore, we evaluate the LLMs' capabilities under 0-shot, 1-shot, and 3-shot settings. Although previous researches use a 5-shot setting, our experiments have much bigger options number for each question, taking the maximum input length of each LLM into consideration, we only use at most 3 examples in our few-shot learning experiments. The examples used for demonstration were obtained from Xiezhi-Train, a dataset containing 2,555 questions absent from Xiezhi-Speciality and Xiezhi-Interdiscipline, with a minimum of two labels matching the test questions, an illustration is depicted in Fig. 4.

**Metrics:** In this section, we present mainly two experiment results: the overall performance of all LLMs across various benchmarks, and the ranking of the top eight 0-shot LLMs in

12 non-sensitive domain categories of the Xiezhi-Benchmark with the scores for top and average practitioners. For the 45 open-source models assessed in our evaluation, we calculated the probability of each model choosing every option using generative probabilities and then ranked all options accordingly based on the probabilities. Due to legal considerations, we only display the results of two publicly recognized API-based LLMs: ChatGPT and GPT-4, and we ask them to rank all given options through instructions. To represent the results of all ranking outcomes, we employed the Mean Reciprocal Rank (MRR) as the metric in this section, which calculates the reciprocal rank of the correct answer. MRR closer to 1 indicates that the model is more capable of placing the correct answer at the front of the ranking, while it suggests that the LLM tends to place the correct answer at the bottom if it is closer to 0. As a comparison, we also employ four different metrics and detailed them in Appendix Results on Other Metrics.

**Randomness:** To reduce the effect of randomness on our experiment, we set the random seed of some python libraries used in our experiment, which are `Numpy`, `Random`, and `Torch`, to 42. It is worth noting that since we used a generative probability to rank each option, this generative probability is independent of the hyperparameters to each LLMs. Nonetheless, in order to be consistent in our experiments even for details we did not notice, we still set the deterministic hyperparameters, as described in Appendix Detail Hyperparameters. Besides, Given that each question need to sample other 46 options, we constructed the set of options for each question before we started our experiment to ensure the consistency in our experiment. Also, we used string similarity during sampling to select questions that were very unlikely to be standard answers.

**Human Performance:** Since we mainly collected questions from some of the most important examinations in China, whose average scores will be released annually. Furthermore, for various academic entrance examinations, each institution will publish the average score of their recruit students. We annotate each question using the average score of the available corresponding examination and calculated the mean of all the questions within the benchmark where examination scores can be obtained. Additionally, we used the average scores publicized by several of China's top institution as a representation of a higher level of human performance. While this scoring method has its limitations, which we thoroughly analyze in Appendix Bias, Ethical Problems and Social Impact, it still provides usable human baselines for Xiezhi.

## Results of LLMs

The overall performance towards Xiezhi and baselines of all LLMs are listed in Tab. **??**. The ranking of all LLMs in each domain category is listed in Tab. 1. And here we give the most intriguing observation in the experiments.

Note: (1) The results of GPT-4 and ChatGPT are acquired through instructions, their real capabilities of them may be higher than the score listed in the tables. (2) Tab. 1 displays the optimal outcomes, which are combined performance of Xiezhi-Specialty and Xiezhi-Interdiscipline, in both Chinese and English Xiezhi. (3) At the moment of writing this paper, M3KE has solely released its training dataset. So we employed this dataset for conducting the experiments, which allowed us to execute only 0-shot experimental setups.

**Observation 1: Best Performance = Pretraining + Finetuning** Examining the overall results presented in Tab. 1, it is observed that all top-10 open-source LLMs are built upon either the llama or bloom frameworks. This suggests that obtaining the most exceptional performance is more likely through these two base models, due to their substantial potential and superior performance in domain text comprehension. Moreover, it is noted that all open-source models within the top-10 overall performance in Tab. 1 are finetuned models, which implies that only finetuned LLMs can attain the highest performance. As a result, both effective pretraining and finetuning processes are crucial components in attaining optimal performance in domain text comprehension.

**Observation 2: Most LLMs are incapable of performing stably few-shot learning from demonstrations** As shown in the "Performance-Average" in Tab. **??**, the average performance of LLMs reveals that more quantity of examples results in better model performance. However, it is not an absolute guarantee that each LLM will exhibit enhanced performance in response to an increased number of demonstrations. On the contrary, several LLMs exhibit a decline in performance as the quantity of learning examples expands. In contrast, GPT-4 and ChatGPT demonstrate a more stable improvement in their performance through few-shot learning. This can be attributed to the extensive domain knowledge possessed by GPT-4 and ChatGPT, enabling them to effectively comprehend the features embedded within the learning samples.

**Observation 3: More LLMs' parameters don't guarantee better performance** Numerous studies have posited that an increase in the number of model parameters corresponds to an enhancement in model's performance. This notion holds true when comparing LLMs that exhibit an order of magnitude difference in their parameters. For instance, Bloomz-mt with 146 billion parameters significantly outperforms Bloomz-560m with 560 million parameters. However, this argument does not consistently hold. For instance, Bloomz-7b1 surpasses Bloomz-p3 in the majority of domain tasks, and Pythia-1.4b outperforms other Pythia models with larger parameter counts across most benchmarks. A possible explanation for this phenomenon could be that LLMs with different parameter quantities are optimally suited to different amounts of pre-training and fine-tuning data (Hoffmann et al. 2022).

**Observation 4: Small LMs enhance domain capabilities at the expense of generic capabilities** In our experiments, we examined two medical LLMs: DoctorGLM and Baize-Healthcare. DoctorGLM originated from ChatGLM-6B, and Baize-Healthcare was derived from Llama-7B, with both models fine-tuned using medical domain text. Although both models have lower MRR compared to other LLMs fine-tuned based on the same base models, they each demonstrate high performance in medical domain. This suggests the augmentation of LLMs with fewer parameters in domain text comprehension, whether finetuned through exclusively domain-specific data or combining domain-specific and generic data,

| Category | Human | | Language Models | | | | |
|---|---|---|---|---|---|---|---|
| | Top | Average | | | | | |
| Phi. | 0.856✓ | 0.453✗ | gpt3.5<br>0.477 | bloomz-mt<br>0.453 | gpt4<br>0.413 | pythia-1.4b<br>0.321 | llama-7b-hf<br>0.241 |
| Eco. | 0.871✓ | 0.520✓ | gpt4<br>0.419 | bloomz-mt<br>0.310 | llama-65b-hf<br>0.290 | belle-7b-1m<br>0.255 | llama-7b-hf<br>0.234 |
| Jur. | 0.761✓ | 0.460✓ | gpt4<br>0.368 | llama-65b-hf<br>0.323 | baize-lora-7b<br>0.230 | belle-7b-0.2m<br>0.217 | gpt3.5<br>0.213 |
| Ped. | 0.854✓ | 0.510✓ | gpt4<br>0.472 | bloomz-mt<br>0.442 | gpt3.5<br>0.280 | belle-7b-0.2m<br>0.251 | baize-lora-13b<br>0.244 |
| Lit. | 0.825✓ | 0.560✓ | gpt4<br>0.417 | bloomz-mt<br>0.405 | baize-lora-7b<br>0.284 | baize-lora-13b<br>0.249 | baize-lora-7b<br>0.213 |
| His. | 0.854✓ | 0.460✓ | gpt4<br>0.437 | bloomz-mt<br>0.272 | gpt3.5<br>0.233 | belle-7b-0.2m<br>0.214 | belle-7b-1m<br>0.207 |
| Sci. | 0.926✓ | 0.394✗ | gpt4<br>0.436 | bloomz-mt<br>0.408 | gpt3.5<br>0.220 | belle-7b-1m<br>0.210 | bloomz-3b<br>0.200 |
| Eng. | 0.928✓ | 0.380✗ | gpt4<br>0.420 | gpt3.5<br>0.412 | bloomz-mt<br>0.387 | bloomz-7b1<br>0.274 | bloomz-7b1-mt<br>0.253 |
| Agr. | 0.902✓ | 0.333✗ | gpt4<br>0.515 | bloomz-mt<br>0.366 | gpt3.5<br>0.311 | bloomz-7b1-mt<br>0.224 | belle-7b-0.2m<br>0.216 |
| Med. | 0.805✓ | 0.430✗ | gpt4<br>0.469 | baize-lora-7b<br>0.279 | gpt3.5<br>0.265 | doctorglm-6b<br>0.253 | belle-7b-0.2m<br>0.223 |
| Man. | 0.857✓ | 0.513✓ | gpt4<br>0.390 | baize-lora-30b<br>0.375 | pythia-2.8b<br>0.367 | bloomz-p3<br>0.280 | belle-7b-0.2m<br>0.268 |
| Art. | 0.821✓ | 0.400✗ | gpt4<br>0.437 | baize-lora-7b<br>0.417 | bloomz-mt<br>0.377 | gpt3.5<br>0.339 | belle-7b-0.2m<br>0.238 |
| Xiezhi<br>Overall | gpt4<br>0.431 | bloomz-mt<br>0.337 | gpt3.5<br>0.267 | belle-7b-0.2m<br>0.211 | belle-7b-1m<br>0.209 | bloomz-7b1<br>0.203 | baize-lora-7b<br>0.200 |
| MMLU<br>Overall | gpt4<br>0.402 | bloomz-mt<br>0.266 | gpt3.5<br>0.240 | baize-30b (lora)<br>0.193 | bloomz-7b1-mt<br>0.189 | bloomz-7b1<br>0.167 | llama-13b<br>0.166 |
| C-Eval<br>Overall | gpt4<br>0.413 | gpt3.5<br>0.286 | bloomz-mt<br>0.204 | baize-7b (lora)<br>0.194 | baize-30b (lora)<br>0.191 | baize-13b (lora)<br>0.184 | baize-7b (lora)<br>0.178 |
| M3KE<br>Overall | gpt4<br>0.404 | gpt3.5<br>0.290 | baize-7b (lora)<br>0.231 | baize-7b (lora)<br>0.203 | bloomz-mt<br>0.161 | llama-7b<br>0.158 | baize-13b (lora)<br>0.155 |

Table 1: Ranking of all LLMs in each category in 0-shot setting. ✓ denotes human performance exceeds the state-of-the-art LLMs, whereas ✗ signifies LLMs have surpassed human performance.

will inevitably lead to a trade-off in the understanding of generic text. This observation aligns with the findings from previous research (Fu et al. 2023; Zhao et al. 2023).

### Results of Benchmarks

Based on the observations from Tab. 1, although the objective is to comprehensively evaluate the domain capabilities of LLMs, the various benchmarks still exhibit differing results, which indicates the different emphases of each benchmark. GPT-4, ChatGPT, and Bloomz-mt consistently rank within the top 10 across all four benchmarks, Baize-7b, and Bloomz-7b1 demonstrate remarkable abilities as they rank within the top 10 across three of the benchmarks. Furthermore, Xiezhi exhibits the highest variance among all LLMs in the ”Performance-Variance” of Tab. **??**, while the score of GPT-4 doesn’t always rank first like it was in other benchmark works. This indicates that the Xiezhi Benchmark excels at discerning the competence disparities among diverse LLMs and possesses the potential to appraise more potent LLMs.

## Conclusion

We introduced Xiezhi, a new benchmark that measures how well LLMs acquire and apply domain knowledge. By covering 516 subjects ranging from 13 categories with 249,587 questions, Xiezhi proposes a taxonomy of all human knowledge and assesses language understanding of the cutting-edge 47 LLMs in greatest breadth and depth among all previous benchmarks. Our research has revealed that the SOTA LLMs have outperformed practitioner experts in several domains when evaluated by multiple-choice question answering tasks. Furthermore, there is still a big gap in generic domain knowledge comprehension between larger and smaller models. Our experimental findings and the Xiezhi Benchmark we developed provide researchers with a more comprehensive understanding of their capabilities across diverse domains.

## Acknowledgements

Professor Yanghua Xiao is also a member of Research Group of Computational and AI Communication at Institute for Global Communications and Integrated Media, Fudan University. This work was supported by Science and Technology Commission of Shanghai Municipality Grant (No. 22511105902). Shanghai Municipal Science and Technology Major Project (No.2021SHZDZX0103). National Natural Science Foundation of China (No.62102095). National Natural Science Foundation of China (No. 62306112). National Natural Science Foundation of China (No. U23A20496).

## References

Biderman, S.; Schoelkopf, H.; Anthony, Q.; Bradley, H.; O'Brien, K.; Hallahan, E.; Khan, M. A.; Purohit, S.; Prashanth, U. S.; Raff, E.; Skowron, A.; Sutawika, L.; and van der Wal, O. 2023. Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling. arXiv:2304.01373.

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, 7432–7439.

Bubeck, S.; Chandrasekaran, V.; Eldan, R.; Gehrke, J.; Horvitz, E.; Kamar, E.; Lee, P.; Lee, Y. T.; Li, Y.; Lundberg, S.; et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

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.

Chung, H. W.; Hou, L.; Longpre, S.; Zoph, B.; Tay, Y.; Fedus, W.; Li, E.; Wang, X.; Dehghani, M.; Brahma, S.; et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

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.

Fu, Y.; Peng, H.; Ou, L.; Sabharwal, A.; and Khot, T. 2023. Specializing Smaller Language Models towards Multi-Step Reasoning. *arXiv preprint arXiv:2301.12726*.

H2O.ai. 2023. h2oGPT - The world's best open source GPT. https://github.com/h2oai/h2ogpt.

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

Hoffmann, J.; Borgeaud, S.; Mensch, A.; Buchatskaya, E.; Cai, T.; Rutherford, E.; de Las Casas, D.; Hendricks, L. A.; Welbl, J.; Clark, A.; et al. 2022. An empirical analysis of compute-optimal large language model training. *Advances in Neural Information Processing Systems*, 35: 30016–30030.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Hu, Z.; Lan, Y.; Wang, L.; Xu, W.; Lim, E.-P.; Lee, R. K.-W.; Bing, L.; and Poria, S. 2023. LLM-Adapters: An Adapter Family for Parameter-Efficient Fine-Tuning of Large Language Models. *arXiv preprint arXiv:2304.01933*.

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.; Gong, Y.; Deng, Y.; Peng, Y.; Niu, Q.; Ma, B.; and Li, X. 2023. Towards Better Instruction Following Language Models for Chinese: Investigating the Impact of Training Data and Evaluation. *arXiv preprint arXiv:2304.07854*.

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 preprint 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*.

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*.

Niklaus, J.; Matoshi, V.; Rani, P.; Galassi, A.; Stürmer, M.; and Chalkidis, I. 2023. Lextreme: A multi-lingual and multi-task benchmark for the legal domain. *arXiv preprint arXiv:2301.13126*.

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

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

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1): 5485–5551.

Sai, A. B.; Mohankumar, A. K.; and Khapra, M. M. 2022. A survey of evaluation metrics used for NLG systems. *ACM Computing Surveys (CSUR)*, 55(2): 1–39.

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.; Xiaotian, Z.; Zhengfu, H.; Peng, L.; Qinyuan, C.; Hang, Y.; Xiangyang, L.; Yunfan, S.; Qiong, T.; Xingjian, Z.; Ke, C.; Yining, Z.; Zhejian, Z.; Ruixiao, L.; Jun, Z.; Yunhua, Z.; Linyang, L.; Xiaogui, Y.; Lingling, W.; Zhangyue, Y.; Xuanjing, H.; and Xipeng, Q. 2023. FudanNLP Moss.

Suzgun, M.; Scales, N.; Schärli, N.; Gehrmann, S.; Tay, Y.; Chung, H. W.; Chowdhery, A.; Le, Q. V.; Chi, E. H.; Zhou, D.;

et al. 2022. Challenging BIG-Bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*.

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. https://github.com/tatsu-lab/stanford_alpaca.

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*.

Wei, J.; Tay, Y.; Bommasani, R.; Raffel, C.; Zoph, B.; Borgeaud, S.; Yogatama, D.; Bosma, M.; Zhou, D.; Metzler, D.; et al. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

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*.

Zellers, R.; Holtzman, A.; Bisk, Y.; Farhadi, A.; and Choi, Y. 2019. HellaSwag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*.

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

Zhao, W. X.; Zhou, K.; Li, J.; Tang, T.; Wang, X.; Hou, Y.; Min, Y.; Zhang, B.; Zhang, J.; Dong, Z.; et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

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*.

Zhou, C.; Li, Q.; Li, C.; Yu, J.; Liu, Y.; Wang, G.; Zhang, K.; Ji, C.; Yan, Q.; He, L.; et al. 2023. A comprehensive survey on pretrained foundation models: A history from bert to chatgpt. *arXiv preprint arXiv:2302.09419*.

Zhu, M.; Deng, Z.; Xiong, W.; Yu, M.; Zhang, M.; and Wang, W. Y. 2019. Neural Correction Model for Open-Domain Named Entity Recognition. *arXiv preprint arXiv:1909.06058*.