LHMKE: A Large-scale Holistic Multi-subject Knowledge Evaluation Benchmark for Chinese Large Language Models

Chuang Liu, Renren Jin, Yuqi Ren, Deyi Xiong*
College of Intelligence and Computing, Tianjin University
Tianjin, China
{liuc_09,rrjin,ryq20,dyxiong}@tju.edu.cn

Abstract

Chinese Large Language Models (LLMs) have recently demonstrated impressive capabilities across various NLP benchmarks and real-world applications. However, the existing benchmarks for comprehensively evaluating these LLMs are still insufficient, particularly in terms of measuring knowledge that LLMs capture. Current datasets collect questions from Chinese examinations across different subjects and educational levels to address this issue. Yet, these benchmarks primarily focus on objective questions such as multiple-choice questions, leading to a lack of diversity in question types. To tackle this problem, we propose LHMKE, a Large-scale, Holistic, and Multi-subject Knowledge Evaluation benchmark in this paper. LHMKE is designed to provide a comprehensive evaluation of the knowledge acquisition capabilities of Chinese LLMs. It encompasses 10,465 questions across 75 tasks covering 30 subjects, ranging from primary school to professional certification exams. Notably, LHMKE includes both objective and subjective questions, offering a more holistic evaluation of the knowledge level of LLMs. We have assessed 11 Chinese LLMs under the zero-shot setting, which aligns with real examinations, and compared their performance across different subjects. We also conduct an in-depth analysis to check whether GPT-4 can automatically score subjective predictions. Our findings suggest that LHMKE is a challenging and advanced testbed for Chinese LLMs.

Keywords: Chinese LLMs, Evaluation Benchmark, Knowledge Evaluation

1. Introduction

We have recently witnessed an influx of large language models (LLMs) , which are either proprietary or open-source. Among them, the proliferation of Chinese LLMs (Du et al., 2022; Zeng et al., 2023a, 2021; Sun et al., 2021; Wu et al., 2021) has been remarkable, with over 60 models unveiled this year alone.[1] The evaluation of these models has consequently emerged as a critical concern.

Traditional benchmarks may no longer suffice as they are typically designed to assess specific tasks like machine translation or question answering. However, LLMs, having been trained on a variety of instructions (Shen et al., 2023a), possess the capability to respond to and perform a diverse array of questions and tasks. This indicates a need for more comprehensive benchmarks for their evaluation. A direct approach would be to amalgamate various independent tasks into a unified benchmark for a holistic evaluation of LLMs. Examples of such integrated benchmarks include SuperGLUE (Wang et al., 2019) and BIG-bench (Srivastava et al., 2022).

To measure the knowledge acquistion and application of LLMs, simply matching superficial semantic clues in text is not enough. A more effective way to evaluate LLMs is using questions from human exams, which cover various subjects and educational levels. For example, MMLU (Hendrycks et al., 2021) contains 57 subjects from college and high school courses. Some Chinese benchmarks for LLMs, such as M3KE (Liu et al., 2023a), C-Eval (Huang et al., 2023), and CMMLU (Li et al., 2023a), follow the same design philosophy as MMLU. However, these benchmarks only focus on one type of questions: multiple-choice questions. This form of questions, despite facilitating the automatic evaluation of LLMs in knowledge application, is not adequate to assess the capabilities of LLMs comprehensively and deeply as LLMs only need to make simple judgments (shortcuts might be exploited during this decision process).

In contrast, human exams include different types of subjective questions, e.g., writing, conditional analysis, conceptual explanations, in addition to multiple-choice questions (MCQs). Different from MCQs, subjective questions are normally equipped with standard or reference answers that are used to compare with answers provided by testers. Because of this, assessing tester answers often requires a broad range of knowledge, rather than word matching. In real human exams, this is usually done by experienced teachers as reviewers. However, this manual assessment is not desirable for testing LLMs as it is usually time-consuming and expensive.

Fortunately, advanced LLMs, such as GPT-4 (OpenAI, 2023), seem to be a promising automatic assessor in comparing answers with reference answers. Recent studies show that advanced LLMs,

*Corresponding author.

[1] https://github.com/wgwang/LLMs-In-China

LREC-COLING 2024, pages 10476–10487

Benchmark	Language	# Tasks	# Objective Q	# Subjective Q	# Numbers ToQ	Standardized S
MMCU (Zeng, 2023)	Zh	51	11,900	0	1	X
C-Eval (Huang et al., 2023)	Zh	52	13,948	0	1	X
CMMLU (Li et al., 2023a)	Zh	67	11,528	0	1	X
M3KE (Liu et al., 2023a)	Zh	71	20,477	0	1	X
Xiezhi (Gu et al., 2023)	Zh	516	249,587	0	1	X
GAOKAO (Zhang et al., 2023b)	Zh	9	1,781	1,030	3	X
CG-Eval (Zeng et al., 2023b)	Zh	55	0	11,000	3	X
LHMKE (Ours)	Zh	75	7,884	2,581	32	✓

Table 1: The comparison between LHMKE and other related benchmarks. Q: Question. ToQ: Type of Question. S: Scoring.

if equipped with well-designed prompts or personalized roles (Chan et al., 2023), are able to compare different pairs of answers and provide scores that are consistent with human evaluators. This inspires and encourages us to build new datasets with multiple types of questions (including subjective questions) for comprehensively and automatically evaluating LLMs.

We hence propose LHMKE, a **L**arge-scale, **H**olistic and **M**ulti-subject **K**nowledge **E**valuation benchmark for Chinese LLMs. LHMKE covers 30 subjects with 75 tasks, and each question in LHMKE is sourced from the realistic standard exams with a specific score. This allows us to standardize each subject to a uniform scoring system. We compare LHMKE with other related benchmarks in Table 1.

We have evaluated 11 Chinese LLMs on the proposed LHMKE, focusing only on LLMs instruction-tuned by Supervised Fine-tuning (SFT) (Ouyang et al., 2022) and Reinforcement Learning from Human Feedback (RLHF) (Christiano et al., 2017; Stiennon et al., 2020; Ouyang et al., 2022) because of their remarkable capabilities (Chung et al., 2022; Wei et al., 2022; Sanh et al., 2022). Generally, experimental results show that current Chinese LLMs have difficulty in achieving high scores, with noticeable performance gaps across different subjects. Most LLMs perform better in the elementary and secondary school exams than exams of other education levels. Unsurprisingly, the newest versions of LLMs, such as ChatGLM-6B[2] and ChatGLM2-6B[3], surpass their earlier versions. Moreover, the tested LLMs exhibit expertise in the Teacher Certification subject within the career development group and the Education subject within the college group. However, the newest versions of LLMs are not always better than old versions. This implies that while current LLMs have improved in subjects related to basic education, they still face challenges in other domains.

Additionally, we have compared various methods for scoring subjective questions in LHMKE. These methods include traditional metrics, using one LLM as a reviewer, and using two LLMs as reviewers. We have also used GPT-3.5[4] and GPT-4 (OpenAI, 2023) as our initial evaluators. We define different prompts to instruct them to mimic human reviewers when grading LLM-autored answers. Our results suggest that GPT-4 with appropromiate prompts matches most closely with human scorers.

Our main contributions in the paper:

- We introduce LHMKE, a comprehensive, multi-subject knowledge evaluation benchmark for Chinese LLMs, which to date encompasses the largest number of question types in alignment with the major Chinese education system.
- We have conducted tests on a broad of latest open-source SFT/RLHF Chinese LLMs under a zero-shot setting.
- We have evaluated the performance of each LLM across different subjects. In addition, various evaluation methods for automatically scoring LLM-generated answers of subjective questions have also been explored on LHMKE. We release LHMKE (data and evaluation scripts) at `https://github.com/tjunlp-lab/LHMKE`.

2. Related work

A variety of benchmarks (Guo et al., 2023) have been developed to evaluate Chinese LLMs capacity for knowledge acquisition and application. Unlike other datasets designed for assessing language comprehension (Xu et al., 2023; Yu et al., 2023), reasoning (Zhang et al., 2023a; Wang et al., 2023; Shi et al., 2023), role-play (Shen et al., 2023b), bias (Huang and Xiong, 2023) and interaction with environments (Li et al., 2023b; Zhuang et al., 2023; Liu et al., 2023b; Zhou et al., 2023), these benchmarks focus on measuring the knowledge acquired during training, which is a fundamental aspect of understanding the capabilities of LLMs. Current Chinese knowledge evaluation benchmarks consist

[2]https://github.com/THUDM/ChatGLM-6B

[3]https://github.com/THUDM/ChatGLM2-6B

[4]https://platform.openai.com/docs/guides/gpt

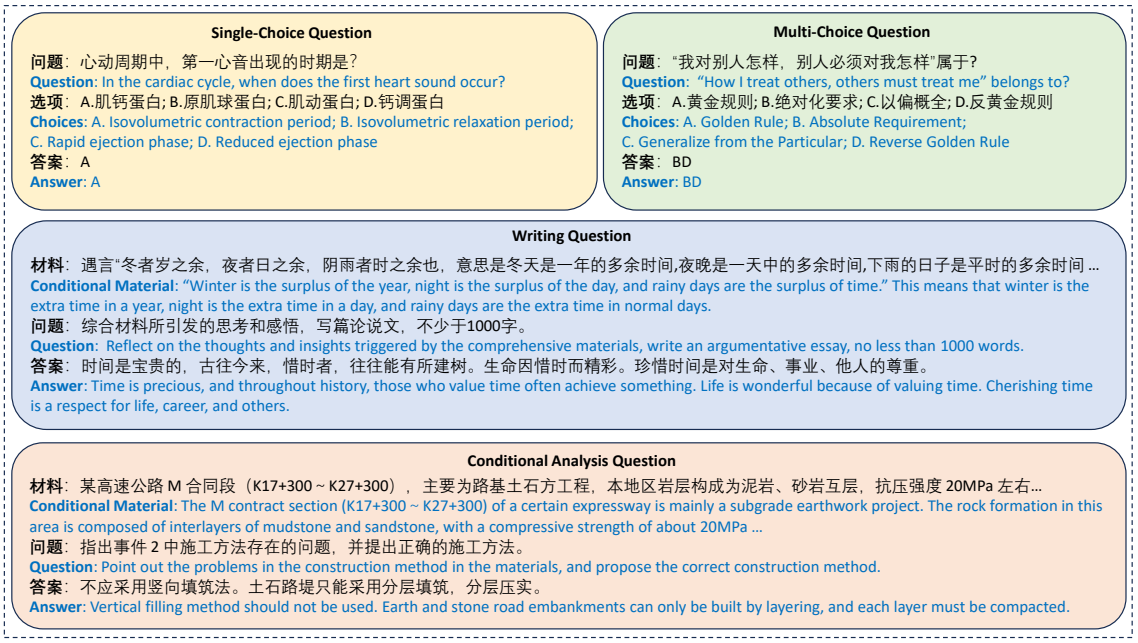

Figure 1: Examples in LHMKE. The yellow example of a objective question with single-choice from Western Medicine subject. The green example of a objective question with multi-choice from Psychological Counselor subject. The blue example of subjective question with writing from Teacher Certification . The orange example of subjective question with conditional analysis from Construction Practical Examination.

of multiple-choice questions collected from various examinations, with LLM performance evaluated in terms of accuracy.

C-Eval, proposed by Huang et al. (2023), comprises 13,948 multiple-choice questions across 52 tasks. Concurrent to C-Eval, M3KE (Liu et al., 2023a) collects 20,477 multiple-choice questions on 71 tasks. MMCU (Zeng, 2023) and CMMLU (Li et al., 2023a) are similar benchmarks consisting of multiple-choice questions, containing 11,900 and 11,528 questions respectively. Xiezhi (Gu et al., 2023) stands out due to its size, encompassing 249,587 questions across 516 subjects.

Despite the rapid expansion of Chinese benchmarks, there is a noticeable lack of subjective questions. To enhance the diversity of question types, GAOKAO-Bench (Zhang et al., 2023b) and CG-Eval (Zeng et al., 2023b) are proposed accordingly. GAOKAO-Bench (Zhang et al., 2023b) includes real Chinese college entrance examination questions, comprising 1,781 objective questions and 1,030 subjective questions. Meanwhile, CG-Eval (Zeng et al., 2023b) assesses Chinese text generation capabilities with 11,000 subjective questions across three question types.

In comparison to these works, LHMKE is a comprehensive Chinese benchmark that not only spans the entire Chinese educational spectrum from primary school to career development but also includes both objective and subjective questions from standard Chinese examinations. Figure 1 shows examples in LHMKE.

3. LHMKE

LHMKE encompasses a broad spectrum of Chinese education levels, ranging from primary school tests to professional exams, with a total of 75 tasks across 30 subjects. The data in LHMKE, which are all standard exam questions, have been meticulously collected online by college students to ensure their quality.

LHMKE is divided into three distinct groups: Elementary and Secondary School, College, and Career Development. The elementary and secondary school group includes educational levels of primary school, middle school, and high school. The college group comprises questions from major fields of study in line with the Chinese postgraduate examination. The career development group features questions from popular professional qualification examinations.

We have instructed collectors to expand the scale of each subject based on its standard score. This means that the total score for each subject should be a multiple of its standard score, facilitating the quantification of LLMs' performances. Furthermore, we have capped the number of questions in each subject at 300 for efficiency. For example, the score in the primary school math exam is 100 with 20 questions. This indicates that we require 15 sets of primary school math questions to accumulate

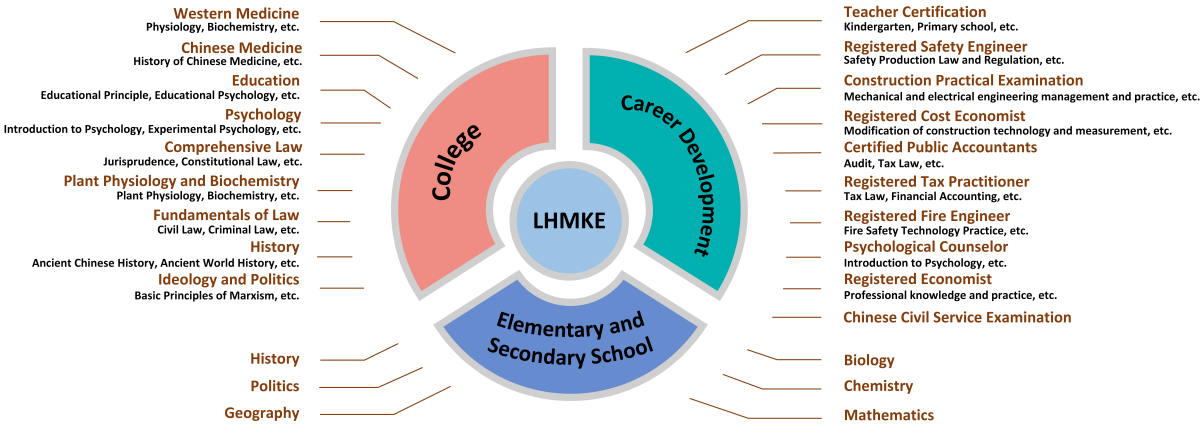

Figure 2: Main subjects in LHMKE.

a total of 300 questions, with the overall scores amounting to 1500. To maintain completeness, we have included several complete examination papers for every subject as far as possible. We only replace undesired questions with those from the same examination papers from other years.

In total, we have amassed 10,465 questions across 30 subjects, which correspond to 75 tasks. All subjects and tasks in LHMKE are presented in Figure 2.

3.1. the Group of Elementary and Secondary School

This group is composed of nine subjects. We have selected mathematics at the primary school level, as the mathematics subjects at higher educational levels often involve a large number of equations, which is not the primary objective of LHMKE. In addition to mathematics, we also select history, politics, and biology, which are taught in both middle school and high school. Additionally, the subjects of chemistry and geography are included from middle school and high school, respectively.

3.2. the Group of College

This group consists of 11 subjects with 37 tasks, spanning a variety of fields. Specifically, we include Psychology, Education, History, Ideology and Politics, Western Medicine, Chinese Medicine, Comprehensive Law, Fundamentals of Law, and Plant Physiology and Biochemistry in this group. Each Law subject is divided into sections for law students and non-law students to meet the requirements of Chinese educational departments. Moreover, each subject is a comprehensive examination, in other words, they are all composed of several tasks. For instance, the Psychology subject covers Developmental and Educational Psychology, Experimental Psychology, Introduction to Psychology and Psychology of Teaching.

	G.ESS	G.C	G.CD
No.S	9	11	10
No.T	9	37	29
No.Q	3,555	3,554	3,356
No.OQ	2,031	2,819	3,034
No.SQ	1,524	735	322
No.Avg.Q	395	323.1	335.6
No.Max.Q	609	344	421
No.Min.Q	314	295	270
Avg.OQ Tokens	127.0	92.4	116.2
Avg.OA Tokens	1.0	1.4	1.4
Avg.SQ Tokens	217.7	189.0	165.8
Avg.SA Tokens	75.0	196.7	74.7

Table 2: Overall statistics of LHMKE. G.ESS: the Group of Elementary and Secondary School. G.C: the Group of College. G.CD: the Group of Career Development. S: Subject. T: Task. Q: Question. OQ: Objective Question. SQ: Subjective Question.

3.3. the Group of Career Development

In this group, we have collected 10 Chinese professional qualification examinations as subjects. These include the Chinese Civil Service Examination, Teacher Certification (General Qualifications), Registered Safety Engineer, Certified Public Accountants, Psychological Counselor, Construction Practical Examination, Registered Fire Engineer, Registered Tax Practitioner, Registered Economist and Registered Cost Economist. Similar to the college group, each subject also comprises multiple tasks. For example, the Certified Public Accountants includes four tasks: audit, tax law, corporate strategy and risk management.

Components	Prompt
Description of the given Role	We would like to get your feedback on Answer 2's performance in answering the above user question with reference to the question and Answer 1's answer, as Answer 1's answer is completely correct. Please rate the accuracy of Answer 2's answer to the questions. Both Answer 2 will receive an overall score on a scale of 1 to 10, with higher scores indicating better overall performance. Please first provide a full explanation of your evaluation, avoiding any possible bias, and make sure that Answer 2's score is obtained by referring strictly to Answer 1's response. Then, output two lines representing the scores of Answers 1 and 2. Note that Answer 1's score is always 10.
Scoring Standards	Score Answer 2 strictly according to the following scale: (a) If Answer 2's response answers the question correctly or matches the main points of Answer 1's response exactly and does not contain any errors of detail, a score of 10 is given; (b) If Answer 2's response partially answers the question or partially matches Answer 1's response, a score in the range of 0 to 5 is given; (c) If Answer 2's answer lacks important details or knowledge points compared to Answer 1's answer a score in the range of 0 to 3 is given; (d) If Answer 2's response is irrelevant to the question or inconsistent with Answer 1's response, a score of 0 is given; (e) If Answer 2's response has a clear knowledge error, a score of 0 is given; (f) If Answer 2's response excerpts large portion of content from the question, a score of 0 is given.

Table 3: The prompt given to the evaluator in our experiments.

LLMs	P.M (100)	M.P (100)	M.H (100)	M.B (100)	M.C (100)	H.P (100)	H.H (100)	H.B (90)	H.G (100)	Totals (890)
ChatGLM-6B	14.4	57.6	56.9	43	42.6	37.4	51.9	28.9	45.1	377.8
ChatGLM2-6B	44.5	80.4	72.7	73.7	73.4	69.3	62.2	**57.8**	59.3	593.3
Baichuan2-7B-Chat	19.8	86.7	79.8	64.2	46.0	74.3	**71.8**	47.1	**72.7**	562.4
Baichuan2-13B-Chat	**46.9**	**89.8**	83.7	65.6	71.1	**78.0**	71.6	53.6	70.3	630.6
BELLE-7B	19.2	50.7	48.2	38.1	39.0	37.1	40.5	26.7	38.9	338.4
Chinese-Alpaca-2-7B	16.3	50.8	54.0	42.1	34.1	30.7	40.0	28.1	33.5	329.6
Chinese-Alpaca-2-13B	12.1	62.3	57.5	47.8	38.0	48.2	44.7	31.5	42.0	384.1
MOSS-SFT-16B	7.3	54.1	42.6	29.9	16.0	47.1	37.4	18.8	55.4	308.6
Qwen-7B-Chat	41.6	86.3	**85.9**	75.0	**76.2**	73.5	66.7	55.7	70.8	**631.7**
InternLM-Chat-7B	33.6	83.0	77.3	**75.2**	71.1	64.6	60.8	57.7	52.4	575.7
InternLM-Chat-7B-v1.1	42.7	76.0	68.8	69.8	63.8	65.9	61.4	54.4	51.9	554.7

Table 4: Overall results in the Elementary and Secondary School group, and the numbers in the parentheses represent the total score of the subject in the official exam. P.M: Primary School Math. M.P: Middle School Politics. M.H: Middle School History. M.B: Middle School Biology. M.C: Middle School Chemistry. H.P: High School Politics. H.H: High School History. H.B: High School Biology. H.G: High School Geography.

4. Dataset Statistic

Table 2 presents the overall statistics of LHMKE. The numbers of subjects in three groups are 9, 11 and 10, respectively. And subjects in the elementary and secondary school, college, and career development group involve of different numbers of tasks, individually cover 9, 37 and 29 tasks, respectively. There are 3,555, 3,554 and 3,356 questions in the three group, which can be classified into objective or subjective questions. Specifically, the numbers of objective questions in each group are 2,031, 2,819 and 3,034, and the numbers of subjective question are 1,524, 735 and 322, respectively. Besides, the maximum numbers of questions in the three groups are 609, 344 and 421, and the minimum numbers are 314, 295 and 270, respectively. Additionally, the average lengths of objective questions in each group are 127.0, 92.4, and 116.2 respectively, with corresponding answer lengths of 1.0, 1.4, and 1.4. The average lengths of subjective questions in each group are 217.7, 189.0, and 165.8, with their respective answer lengths being 75.0, 196.7, and 74.7.

5. Experiments

We evaluated a series of Chinese LLMs on LHMKE to understand their capabilities on human exams with a wide variety of question types.

5.1. Assessed LLMs

We accessed a wide range of Chinese LLMs. These LLMs are instruction-tuned by SFT/RLHF including ChatGLM-6B[5], ChatGLM2-6B[6], Baichuan2-7B-Chat/13B[7] (Baichuan, 2023), Qwen-Chat-7B[8] (Bai et al., 2023), MOSS-SFT-16B[9], BELLE-7B (Ji

[5]https://github.com/THUDM/ChatGLM-6B
[6]https://github.com/THUDM/ChatGLM2-6B
[7]https://github.com/baichuan-inc/Baichuan2
[8]https://github.com/QwenLM/Qwen
[9]https://huggingface.co/fnlp/moss-moon-003-sft

et al., 2023), InternLM-Chat-7B/13B (Team, 2023) and Chinese-Alpaca-2-7B/13B (Cui et al., 2023; Taori et al., 2023).

5.2. Evaluation Metrics

Due to multiple question types collected in LHMKE, we evaluated the results of LLMs using different metrics. For objective questions, we used accuracy as the evaluation metric.

For subjective questions, drawing inspiration from Chateval (Chan et al., 2023), we employed GPT-4 as an evaluator, giving it the role of a reviewer. As depicted in Table 3, the prompt is designed with a detailed description of roles and scoring standards.

5.3. Results

We compare the overall scores of each LLM across all subject groups, with each group having its own standard score. This makes it convenient to identify the strengths and weaknesses of different LLMs.

Table 4 provides the results of assessed LLMs over the Elementary and Secondary School group. It has been observed that no single LLM outperforms others across all subjects. Qwen-7B-Chat, narrowly leading Baichuan2-Chat-13B, achieves the highest total scores on this group. Baichuan2-Chat-13B achieves the highest scores in 3 subjects, excelling in Primary School Math, Middle School Politics, and High School Politics, which is closely followed by Baichuan2-Chat-7B and Qwen-7B, leading in High School History and High School Geography, Middle School History and Middle School Chemistry, respectively. ChatGLM2-6B and InternLM-Chat-7B are each leading on High School Biology and Middle School Biology, separately. However, other LLMs lag significantly behind these models. When comparing LLMs of different sizes, such as Chinese-Alpaca-2-13B and Chinese-Alpaca-2-7B, the larger model demonstrates superior performance. Similarly, between different versions of the same LLM like ChatGLM-6B and ChatGLM2-6B, the latest version consistently outperforms its predecessor. Yet InternLM-Chat-7B-v1.1 is not as good as InternLM-Chat-7B, which may be caused by the different instruction data used by these two versions. Interestingly, despite MOSS-SFT-16B being the largest model in our experiments, it does not achieve the highest score in any subjects, indicating that model size is not necessarily indicative of performance.

Unlike the findings in the Elementary and Secondary School group, we observe different trends in the College group (Table 5). Overall, LLMs tend to perform better in social science subjects than natural science subjects. Most models achieve higher scores in Psychology, Education, and History but struggled with Plant Physiology and Biochemistry and Medicine. On this group, Baichuan2-Chat-13B obtains the highest score in terms of both individual and total scores. Apart from Baichuan2-Chat-13B, Baichuan2-Chat-7B outperforms other models in eight subjects while the remaining highest scores are achieved by ChatGLM2-6B, InternLM-Chat-7B and InternLM-Chat-7B-v1.1, separately. Although Qwen-7B-Chat does not achieve the highest score in any subject, its results are competitive. Interestingly, Chinese-Alpaca-2-13B shows a completely different performance pattern in this group; it outperforms ChatGLM2-6B in History, Comprehensive Law, Fundamentals of Law, Fundamentals of Law for non-law students and Plant Physiology and Biochemistry but scores lowest in Chinese Medicine. This could be attributed to the imbalance in the training data for Chinese-Alpaca-2-13B. Additionally, ChatGLM-6B outperforms ChatGLM2-6B in Education and Western Medicine while Chinese-Alpaca-2-7B follows similar trends to Chinese-Alpaca-2-13B in Psychology and Chinese Medicine.

Finally, we evaluated these LLMs on the Career Development group, as shown in Table 6. Despite the fact that the top LLM remains Baichuan2-13B-Chat, the overall performance of LLMs in this group is markedly subpar. It is clear that if the subject is closely associated with the educational level in the Elementary and Secondary School group, LLMs are likely to achieve a high score such as Teacher Certification. However, LLMs encounter difficulties with certain professional domain knowledge, such as various types of certification examinations. Furthermore, Baichuan2-13B-Chat has obtained the lowest score in Psychological Counselor compared with its overall performances, which is in stark contrast to its performance of Psychology in the College group. This suggests that even though current LLMs have acquired extensive knowledge from various data sources, a significant gap still exists between them and domain experts.

6. Analysis

We provide in-depth analyses of LHMKE, which includes a comparison of each LLM's overall performance on objective questions versus subjective questions, and an explanation as to why GPT-4 with careful prompting is the most suitable evaluator.

6.1. Comparing LLM Performance between Objective and Subjective Questions

Figure 3 presents a comparative analysis of the performance of various LLMs on objective and subjective questions. It is clear that the majority of the evaluated LLMs exhibit superior performance on sub-

LLMs	P (300)	Edu (300)	H (300)	IP (100)	WM (300)	CM (300)	CL (150)	FL (150)	CL* (150)	FL* (150)	PPaB (150)	Totals (2350)
ChatGLM-6B	89.4	123.1	67.3	56.0	98.8	77.0	61.1	34.1	60.4	41.4	26.6	735.2
ChatGLM2-6B	102.6	119.4	86.9	60.5	78.8	94.3	67.4	45.6	71.1	53.2	39.7	819.5
Baichuan2-7B-Chat	136.4	173.5	173.1	59.2	113.0	88.5	96.3	80.9	95.4	70.4	74.7	1161.4
Baichuan2-13B-Chat	**158.9**	**196.4**	**181.6**	**73**	**135.0**	**103.3**	**107.9**	**83.0**	**104.2**	**74.3**	**89.8**	**1307.4**
BELLE-7B	78.8	83.7	58.8	32.0	83.3	55.5	40.0	29.6	44.0	30.2	31.1	567.0
Chinese-Alpaca-2-7B	65.0	91.6	95.3	30.7	28.5	21.5	50.0	40.4	49.6	40.3	44.5	557.4
Chinese-Alpaca-2-13B	63.4	113.2	95.8	38.9	25.25	3.0	74.0	52.4	67.5	53.8	50.8	638.0
MOSS-SFT-16B	63.5	91.3	77.2	36.8	39.8	35.3	50.8	38.7	51.9	39.6	31.3	556.2
Qwen-7B-Chat	107.9	144.7	116.1	54.1	114.0	70.8	78.7	70.3	80.6	65.3	56.0	958.5
InternLM-Chat-7B	95.8	110.3	83.3	59.7	112.0	89.5	38.9	65.4	74.3	60.7	71.3	861.2
InternLM-Chat-7B-v1.1	122.3	130.5	102.3	53.1	114.3	84.3	65.7	51.2	65.4	54.9	45.1	889.1

Table 5: Overall results in the College group, and the numbers in the parentheses represent the total score of the subject in the official exam. P: Psychology. Edu: Education. H: History. IP: Ideology and Politics. WM: Western Medicine. CM: Chinese Medicine. CL: Comprehensive Law. FL: Fundamentals of Law. CL*: Comprehensive Law for non-law students. FL*: Fundamentals of Law for non-law students. PPaB: Plant Physiology and Biochemistry.

LLMs	CCSE (100)	TC (150)	RSE (100)	CPA (100)	PC (100)	CPE (120)	RFE (120)	RTP (140)	RE (140)	RCE (100)	Totals (1170)
ChatGLM-6B	38.5	74.7	25.0	23.3	23.9	38.8	22.7	19.7	21.0	21.5	309.1
ChatGLM2-6B	28.3	102.4	30.3	21.0	44.9	34.0	21.3	22.4	26.3	21.5	352.4
Baichuan2-7B-Chat	40.9	108.1	31.8	27.8	16.3	9.1	27.7	20.5	32.0	23.0	337.2
Baichuan2-13B-Chat	39.2	**110.8**	**38.0**	29.3	7.4	**60.0**	**31.7**	28.9	39.7	27.3	**412.3**
BELLE-7B	33.8	55.0	21.5	14.5	24.2	26.1	23.0	15.6	27.0	17.5	258.2
Chinese-Alpaca-2-7B	12.0	55.7	13.0	11.2	0.6	26.1	14.7	7.04	5.7	5.0	151.0
Chinese-Alpaca-2-13B	9.2	63.7	11.8	12.1	1.8	35.1	10.0	4.8	4.7	5.5	158.7
MOSS-SFT-16B	26.2	57.8	20.0	12.7	10.5	4.6	20.0	11.1	12.7	12.3	187.9
Qwen-7B-Chat	41.3	86.1	32.3	27.7	30.0	54.7	28.0	20.4	34.7	25.5	380.7
InternLM-Chat-7B	42.9	53.7	35.5	20.5	**47.5**	37.7	31.3	36.6	**44.7**	**28.8**	379.2
InternLM-Chat-7B-v1.1	**43.0**	68.5	31.0	**42.8**	43.4	47.2	26.0	**38.9**	41.7	25.8	408.3

Table 6: Overall results in the Career Development group, and the numbers in the parentheses represent the total score of the subject in the official examination. CCSE: Chinese Civil Service Examination. TC: Teacher Certification (General Qualifications). RSE: Registered Safety Engineer. CPA: Certified Public Accountants. PC: Psychological Counselor. CPE: Construction Practical Examination. RFE: Registered Fire Engineer. RTP: Registered Tax Practitioner. RE: Registered Economist. RCE: Registered Cost Economist.

jective questions as compared to objective questions, with InternLM-Chat-7B and InternLM-Chat-7B-v1.1 being notable exceptions. This, however, should not be misconstrued to suggest that subjective questions are less challenging. In contrast to objective questions, subjective questions offer the possibility of partial scoring even when the answers are not entirely accurate. The most commendable performance on subjective questions is demonstrated by Baichuan2-7B and 13B, while InternLM-Chat-7B and InternLM-Chat-7B-v1.1 emerge as the top-performing LLMs for objective questions. These observations underscore the potential of a balanced mix of objective and subjective questions for a more nuanced evaluation of LLMs.

6.2. Analysis for Evaluating Subjective Question

In this section, we conducted a comparative analysis on the evaluation methods for subjective questions, demonstrating that our evaluator outperforms others. Initially, we randomly selected 100 subjective questions from the outputs of evaluated LLMs as examples. Each selection comprised a predicted answer and a corresponding reference answer. Subsequently, we enlisted three postgraduate students to score these predictions, establishing a human benchmark for comparison with our evaluator. This implies that an optimal evaluator would align more closely with human scoring.

To identify the most effective evaluator, we employed GPT-3.5 and GPT-4 as base models. Despite these models having demonstrated their evaluative capabilities, careful design is still required to better align them with human scoring. Broadly speaking, we explored two settings: careful prompting and multi-agent debates. For the careful prompting setting, we designed several prompts, as illustrated in Table 3, directing the evaluator to adhere to them, thereby enabling precise scoring. For multi-agent debates, we utilized two LLMs as reviewers; after the first reviewer assigned a score,

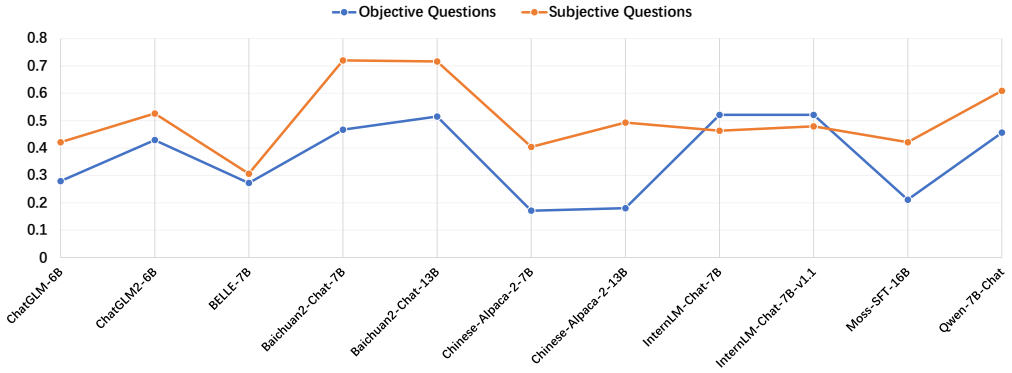

Figure 3: Comparing each LLM's performance in objective questions vs. subjective questions.

Metrics	Evaluator	TS	AS
Traditional Metrics	BLEU	5.0	0.05
	ROUGE-1	274	2.74
	ROUGE-2	113	1.13
	ROUGE-L	164	1.64
Careful Prompting	GPT-3.5	592	5.92
	GPT-4	500	5.0
Multi-agent Debates	GPT-3.5 & GPT-3.5	585	5.85
	GPT-3.5 & GPT-4	528	5.28
	GPT-4 & GPT-4	503	5.03
Manual Assessment	Human Scorers	276	2.76

Table 7: Comparing different evaluators with human markers. TS: Total Score. AS: Average Score.

Metrics	Evaluators	S	P
Traditional Metrics	BLEU	0.332	0.212
	ROUGE-1	0.398	0.347
	ROUGE-2	0.358	0.281
	ROUGE-L	0.390	0.331
Careful Prompting	GPT-3.5	0.353	0.366
	GPT-4	0.704	0.683
Multi-agent Debates	GPT-3.5 & GPT-3.5	0.358	0.365
	GPT-3.5 & GPT-4	0.633	0.605
	GPT-4 & GPT-4	0.712	0.694

Table 8: Comparing Spearman and Pearson correlation coefficients between different evaluators and human. S: Spearman. P: Pearson.

Indicator	GPT-3.5		GPT-4	
	GP	CP	GP	CP
Average Score	7.76	5.92	6.83	**5.0**
Spearman.C	0.381	0.353	**0.734**	0.704
Pearson.C	0.378	0.366	**0.692**	0.683

Table 9: Comparing careful prompting with general prompting. GP: General Prompting. CP: Careful Prompting. C: Coefficients.

the subsequent reviewer was tasked with verifying the score and granted the authority to modify it if deemed necessary. In addition, we examined major traditional metrics such as BLEU and ROUGE-n.

Table 7 presents the outcomes of all experimental evaluators. We observe that all traditional metrics tend to yield a low score, with BLEU being particularly low at 0.05. Although the scores provided by ROUGE are akin to the average human score, they still exhibit less correlation with human scores, as depicted in Table 8. When evaluators are given a detailed prompt to guide their scoring, the score assigned by GPT-4 is closer to human evaluators than that of GPT-3.5, with average scores of 5.0 and 5.92 respectively. This suggests that GPT-4 may be a superior evaluator, potentially due to its inherent capability to better understand the question and reference answer compared to GPT-3.5. Results in Table 8 suggest that the multi-agent debate method, which uses two GPT-4 reviewers achieves the highest correlation to human scorers. Nevertheless, the improvements over a single GPT-4 are marginal and the debate approach would incur double cost. Moreover, while combination of GPT-3.5 and GPT-4 significantly outperforms two GPT-3.5s, it is still inferior to two GPT-4s. Table 8 also suggests that traditional metrics (e.g., BLEU) are not adequate to evaluate LLMs on subjective questions given their low correlations to human scorers. Consequently, we opted to employ GPT-4 with careful prompting as our experimental evaluator for subjective questions.

6.3. Careful Prompting vs. General Prompting

Table 9 compares the results between careful prompting and general prompting. General prompting refers to the prompts provided to LLMs that do not include the scoring standards shown in Table 3. Despite general prompting appearing to achieve a higher correlation with human scorers, particularly with GPT-4, it suffers in terms of average score. This suggests that while GPT-4 with general prompting can mimic the process of human scoring, it tends to assign high scores to inaccurate predictions. Therefore, the inclusion of scoring standards

in the prompts is necessary.

6.4. Inter-annotator Agreement Analysis

We computed the standard deviation of each annotator's scores on 100 randomly selected subjective questions with respect to a reference. Specifically, we observe 23 instances where the standard deviation is 0, 10 instances where the standard deviation falls between 0 and 1, 29 instances where it falls between 1 and 2, 17 instances between 2 and 3, 11 instances between 3 and 4, 6 instances between 4 and 5, and 4 instances where the standard deviation exceeds 5.

We further find that the standard deviation of all annotators on 79 questions is less than 3, indicating that annotators are able to provide similar scores for most questions. Additionally, when a model produces an obviously incorrect response, all annotators provide the same answer. For instance, all annotators assign 0 points to 22 questions simultaneously. However, they collectively award a perfect score of 10 to only one question. This suggests that different annotators maintain distinct scoring criteria to some extent, in line with the nature of subjective evaluation.

7. Conclusion

We have constructed a new benchmark, LHMKE, to evaluate Chinese LLMs across a diverse range of question types and subjects, spanning from elementary school to professional certifications. LHMKE includes 30 subjects, 75 tasks, and 10,456 questions. We observe that all evaluated state-of-the-art Chinese LLMs struggle on LHMKE. We will publicly release the benchmark to serve as a new testbed for Chinese LLMs.

8. Ethics Statement

This work presents LHMKE, a large-scale, holistic, and multi-subject knowledge evaluation benchmark for Chinese large language models. All data in LHMKE are collected from public sources. All testing instances in LHMKE are carefully scrutinized to exclude any examples with ethical concern.

Acknowledgements

The present research was supported by Zhejiang Lab (No. 2022KH0AB01). We would like to thank the anonymous reviewers for their insightful comments.

9. Bibliographical References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Baichuan. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *CoRR*, abs/2308.07201.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4299–4307.

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. 2022. Scaling instruction-finetuned language models. *CoRR*, abs/2210.11416.

Yiming Cui, Ziqing Yang, and Xin Yao. 2023. Efficient and effective text encoding for chinese llama and alpaca. *arXiv preprint arXiv:2304.08177*.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 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), pages 320–335, Dublin, Ireland. Association for Computational Linguistics.

Zhouhong Gu, Xiaoxuan Zhu, Haoning Ye, Lin Zhang, Jianchen Wang, Sihang Jiang, Zhuozhi Xiong, Zihan Li, Qianyu He, Rui Xu, Wenhao Huang, Zili Wang, Shusen Wang, Weiguo Zheng, Hongwei Feng, and Yanghua Xiao. 2023. Xiezhi: An ever-updating benchmark for holistic domain knowledge evaluation. *CoRR*, abs/2306.05783.

Zishan Guo, Renren Jin, Chuang Liu, Yufei Huang, Dan Shi, Supryadi, Linhao Yu, Yan Liu, Jiaxuan Li, Bojian Xiong, and Deyi Xiong. 2023. Evaluating large language models: A comprehensive survey. *CoRR*, abs/2310.19736.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Yufei Huang and Deyi Xiong. 2023. CBBQ: A chinese bias benchmark dataset curated with human-ai collaboration for large language models. *CoRR*, abs/2306.16244.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. 2023. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Yunjie Ji, Yong Deng, Yan Gong, Yiping Peng, Qiang Niu, Lei Zhang, Baochang Ma, and Xiangang Li. 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*.

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

Minghao Li, Yingxiu Zhao, Bowen Yu, Feifan Song, Hangyu Li, Haiyang Yu, Zhoujun Li, Fei Huang, and Yongbin Li. 2023b. Api-bank: A comprehensive benchmark for tool-augmented llms. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 3102–3116. Association for Computational Linguistics.

Chuang Liu, Renren Jin, Yuqi Ren, Linhao Yu, Tianyu Dong, Xiaohan Peng, Shuting Zhang, Jianxiang Peng, Peiyi Zhang, Qingqing Lyu, Xiaowen Su, Qun Liu, and Deyi Xiong. 2023a. M3KE: A massive multi-level multi-subject knowledge evaluation benchmark for chinese large language models. *CoRR*, abs/2305.10263.

Ruibo Liu, Ruixin Yang, Chenyan Jia, Ge Zhang, Denny Zhou, Andrew M Dai, Diyi Yang, and Soroush Vosoughi. 2023b. Training socially aligned language models in simulated human society. *arXiv preprint arXiv:2305.16960*.

OpenAI. 2023. Gpt-4 technical report. *OpenAI*.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *CoRR*, abs/2203.02155.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal V. Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Févry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. 2022. Multitask prompted training enables zero-shot task generalization. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Tianhao Shen, Renren Jin, Yufei Huang, Chuang Liu, Weilong Dong, Zishan Guo, Xinwei Wu, Yan Liu, and Deyi Xiong. 2023a. Large language model alignment: A survey. *CoRR*, abs/2309.15025.

Tianhao Shen, Sun Li, and Deyi Xiong. 2023b. Roleeval: A bilingual role evaluation benchmark for large language models. *CoRR*, abs/2312.16132.

Dan Shi, Chaobin You, Jiantao Huang, Taihao Li, and Deyi Xiong. 2023. CORECODE: A common

sense annotated dialogue dataset with benchmark tasks for chinese large language models. *CoRR*, abs/2312.12853.

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

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F. Christiano. 2020. Learning to summarize from human feedback. *CoRR*, abs/2009.01325.

Yu Sun, Shuohuan Wang, Shikun Feng, Siyu Ding, Chao Pang, Junyuan Shang, Jiaxiang Liu, Xuyi Chen, Yanbin Zhao, Yuxiang Lu, Weixin Liu, Zhihua Wu, Weibao Gong, Jianzhong Liang, Zhizhou Shang, Peng Sun, Wei Liu, Xuan Ouyang, Dianhai Yu, Hao Tian, Hua Wu, and Haifeng Wang. 2021. ERNIE 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. *CoRR*, abs/2107.02137.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Stanford alpaca: an instruction-following llama model (2023). *URL https://github. com/tatsu-lab/stanford_alpaca*.

InternLM Team. 2023. Internlm: A multilingual language model with progressively enhanced capabilities. https://github.com/InternLM/InternLM.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 3261–3275.

Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. 2023. Scibench: Evaluating college-level scientific problem-solving abilities of large language models. *arXiv preprint arXiv:2307.10635*.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Shaohua Wu, Xudong Zhao, Tong Yu, Rongguo Zhang, Chong Shen, Hongli Liu, Feng Li, Hong Zhu, Jiangang Luo, Liang Xu, et al. 2021. Yuan 1.0: Large-scale pre-trained language model in zero-shot and few-shot learning. *arXiv preprint arXiv:2110.04725*.

Liang Xu, Anqi Li, Lei Zhu, Hang Xue, Changtai Zhu, Kangkang Zhao, Haonan He, Xuanwei Zhang, Qiyue Kang, and Zhenzhong Lan. 2023. Superclue: A comprehensive chinese large language model benchmark. *arXiv preprint arXiv:2307.15020*.

Jifan Yu, Xiaozhi Wang, Shangqing Tu, Shulin Cao, Daniel Zhang-li, Xin Lv, Hao Peng, Zijun Yao, Xiaohan Zhang, Hanming Li, Chunyang Li, Zheyuan Zhang, Yushi Bai, Yantao Liu, Amy Xin, Nianyi Lin, Kaifeng Yun, Linlu Gong, Jianhui Chen, Zhili Wu, Yunjia Qi, Weikai Li, Yong Guan, Kaisheng Zeng, Ji Qi, Hailong Jin, Jinxin Liu, Yu Gu, Yuan Yao, Ning Ding, Lei Hou, Zhiyuan Liu, Bin Xu, Jie Tang, and Juanzi Li. 2023. Kola: Carefully benchmarking world knowledge of large language models. *CoRR*, abs/2306.09296.

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. 2023a. 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.

Hui Zeng. 2023. Measuring massive multitask chinese understanding. *CoRR*, abs/2304.12986.

Hui Zeng, Jingyuan Xue, Meng Hao, Chen Sun, Bin Ning, and Na Zhang. 2023b. Evaluating the generation capabilities of large chinese language models. *arXiv preprint arXiv:2308.04823*.

Wei Zeng, Xiaozhe Ren, Teng Su, Hui Wang, Yi Liao, Zhiwei Wang, Xin Jiang, ZhenZhang Yang, Kaisheng Wang, Xiaoda Zhang, Chen Li, Ziyan Gong, Yifan Yao, Xinjing Huang, Jun Wang, Jianfeng Yu, Qi Guo, Yue Yu, Yan Zhang, Jin Wang, Hengtao Tao, Dasen Yan, Zexuan Yi, Fang Peng, Fangqing Jiang, Han Zhang, Lingfeng Deng, Yehong Zhang, Zhe Lin, Chao Zhang, Shaojie Zhang, Mingyue Guo, Shanzhi Gu, Gaojun Fan, Yaowei Wang, Xuefeng Jin, Qun Liu, and Yonghong Tian. 2021. Pangu-α: Large-scale autoregressive pretrained chinese language models with auto-parallel computation. *CoRR*, abs/2104.12369.

Sarah J Zhang, Samuel Florin, Ariel N Lee, Eamon Niknafs, Andrei Marginean, Annie Wang, Keith Tyser, Zad Chin, Yann Hicke, Nikhil Singh, et al. 2023a. Exploring the mit mathematics and eecs curriculum using large language models. *arXiv preprint arXiv:2306.08997*.

Xiaotian Zhang, Chunyang Li, Yi Zong, Zhengyu Ying, Liang He, and Xipeng Qiu. 2023b. Evaluating the performance of large language models on GAOKAO benchmark. *CoRR*, abs/2305.12474.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. 2023. Webarena: A realistic web environment for building autonomous agents. *CoRR*, abs/2307.13854.

Yuchen Zhuang, Yue Yu, Kuan Wang, Haotian Sun, and Chao Zhang. 2023. Toolqa: A dataset for LLM question answering with external tools. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.