

# ZhuJiu: A Multi-dimensional, Multi-faceted Chinese Benchmark for Large Language Models

Baoli Zhang<sup>1,\*</sup>, Haining Xie<sup>1,2,\*</sup>, Pengfan Du<sup>1,2</sup>, Junhao Chen<sup>3</sup>, Pengfei Cao<sup>1</sup>,  
Yubo Chen<sup>1,2</sup>, Shengping Liu<sup>4</sup>, Kang Liu<sup>1,2</sup> and Jun Zhao<sup>1,2</sup>

<sup>1</sup>Institute of Automation, Chinese Academy of Sciences

<sup>2</sup>School of Artificial Intelligence, University of Chinese Academy of Sciences

<sup>3</sup>Harbin Engineering University, <sup>4</sup>Beijing Unisound Information Technology Co., Ltd  
{baoli.zhang, pengfei.cao, yubo.chen, kliu, jzhao}@nlpr.ia.ac.cn, yisuanwang@hrbeu.edu.cn  
{xiehaining21, dupengfan22}@mailsucas.ac.cn, liushengping@unisound.com

## Abstract

The unprecedented performance of large language models (LLMs) requires comprehensive and accurate evaluation. We argue that for LLMs evaluation, benchmarks need to be comprehensive and systematic. To this end, we propose the ZhuJiu benchmark, which has the following strengths: (1) **Multi-dimensional ability coverage**: We comprehensively evaluate LLMs across 7 ability dimensions covering 51 tasks. Especially, we also propose a new benchmark that focuses on knowledge ability of LLMs. (2) **Multi-faceted evaluation methods collaboration**: We use 3 different yet complementary evaluation methods to comprehensively evaluate LLMs, which can ensure the authority and accuracy of the evaluation results. (3) **Comprehensive Chinese benchmark**: ZhuJiu is the pioneering benchmark that fully assesses LLMs in Chinese, while also providing equally robust evaluation abilities in English. (4) **Avoiding potential data leakage**: To avoid data leakage, we construct evaluation data specifically for 37 tasks. We evaluate 10 current mainstream LLMs and conduct an in-depth discussion and analysis of their results. The ZhuJiu benchmark and open-participation leaderboard are publicly released at <http://www.zhujiu-benchmark.com/> and we also provide a demo video at <https://youtu.be/qypkJ89L1Ic>.

## 1 Introduction

With the continuous development of large language models (LLMs), the emergence of GPT4 (OpenAI, 2023) is enough to trigger a new wave of technology. Various types of LLMs have recently been rapidly developing, such as Llama2 (Touvron et al., 2023) and ChatGLM2 (Du et al., 2022), demonstrating impressive generalization abilities and broad applicability. Therefore, it is crucial to conduct comprehensive and objective evaluations

of LLMs to fully understand their strengths and limitations.

Specifically, on the one hand, for **applicators**, they need to understand the overall performance of LLMs or the advantages of LLMs in a specific aspect. Constructing comprehensive and authoritative benchmarks can help applicators significantly improve the efficiency of using LLMs. On the other hand, for **developers**, the improvement direction of LLMs requires accurate evaluation results as guidance. An objective and fair benchmark can help them carry out relevant research work on LLMs more targetedly.

To this end, scholars conduct extensive research on evaluations for LLMs and construct some superior benchmarks. Normally, the evaluation for LLMs includes two aspects: ability evaluation and evaluation method. Although **traditional benchmarks** such as GLUE (Wang et al., 2018), SuperGLUE (Wang et al., 2019) and CUGE (Yao et al., 2021) still have a role to play in evaluating LLMs, their limitations are becoming increasingly apparent due to the growing diversity of evaluation dimensions and methods for LLMs. For the **ability evaluation** of LLMs, recent work proposes excellent benchmarks for LLMs in one or several aspects, such as knowledge, reasoning, language, safety and hallucination (Liang et al., 2022; Jifan Yu, 2023; Sun et al., 2023a; Amayuelas et al., 2023; Li et al., 2023). However, a comprehensive evaluation of LLMs remains insufficient. For the **evaluation method** of LLMs, there are currently 3 main categories: (1) *Metrics Evaluation*: Evaluating LLMs using existing datasets and corresponding metrics (Liang et al., 2022); (2) *ChatGPT Evaluation*: Using GPT-like LLMs to generate evaluation data and compare the response results of different LLMs (Wang et al., 2023c); (3) *Model Arena*: constructing one-on-one model arenas where humans compare the evaluation results of models based on their own judgment (Zheng et al., 2023).

<sup>1,\*</sup>Co-first authors, they contributed equally to this work.

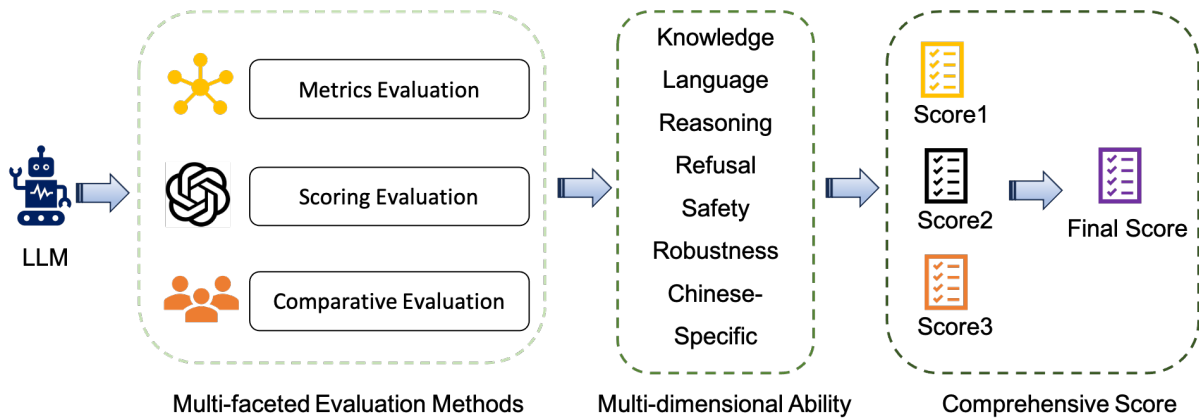


Figure 1: The evaluation process of LLM using ZhuJiu.

Despite these successful efforts for LLMs’ evaluations, existing studies still suffer from several limitations: (1) Current benchmarks tend to focus on evaluating LLMs on a single dimension of their abilities, which can not provide a comprehensive evaluation of LLMs. (2) Most benchmarks only use a single evaluation method, which may not provide an accurate evaluation of all the abilities of LLMs. For example, while HELM(Liang et al., 2022) uses metrics to evaluate LLMs, it may not measure all abilities such as long-text generation or machine translation, etc. (3) The cross-lingual abilities of LLMs, especially for Chinese, have garnered growing attention. However, the lack of a comprehensive Chinese benchmark for LLMs remains a critical issue. (4) Many current benchmarks only use public datasets for evaluation, risking potential data leakage. The results of evaluations based on this data lack credibility.

In this paper, we propose the ZhuJiu Benchmark to solve above mentioned problems, which can fill the gap in the development of a comprehensive benchmark for evaluating LLMs in Chinese. The advantages of the ZhuJiu are as follows: (1) **Multi-dimensional ability coverage:** we evaluate LLMs from 7 ability dimensions, including *knowledge, Chinese-specific, language, reasoning, refusal, safety and robustness abilities*, covering 51 datasets to provide a comprehensive performance assessment. In addition, we also proposed a new paradigm for evaluating the knowledge ability. (2) **Multi-faceted evaluation methods coordination:** we use *Metrics Evaluation, Scoring Evaluation, and Comparative Evaluation* for comprehensively evaluating LLMs to ensure authoritative and accurate evaluation results. (3) **Comprehensive Chinese benchmark:** ZhuJiu is the pioneering Chi-

nese benchmark that can comprehensively evaluate LLMs, while allowing equivalent assessment in English. (4) **Avoiding potential data leakage:** in addition to collecting 14 commonly used datasets, we construct 37 datasets for the evaluation of LLMs, ensuring maximum avoidance of data leakage and evaluation fairness. The overall evaluation process is shown in Figure 1.

We also release an online evaluation platform that supports multiple functions including visualizations of evaluation results, participating in model arena and submission of evaluation model, etc. Moreover, we evaluate 10 publicly available LLMs, including ChatGLM (Du et al., 2022), BELLE (Yunjie Ji and Li, 2023), ChatGPT (OpenAI, 2022), and so on. Based on the experimental results, we observe some interesting phenomena and summarize them in 4.2.

In summary, the contributions of this paper are as follows:

- We propose ZhuJiu, the first Chinese benchmark that covers multi-dimensions of ability and employs multi-faceted evaluation methods in collaboration. Meanwhile in the ZhuJiu we construct a novel benchmark for evaluating knowledge ability and 37 evaluation datasets to prevent data leakage issues.
- We release an online evaluation platform that enables users to evaluate LLMs. We will continue to improve the platform, and update the evaluation leaderboard.
- Using the ZhuJiu benchmark, we evaluate 10 current LLMs, to comprehensively and deeply explore their abilities, providing valuable insights to inform future LLM development.

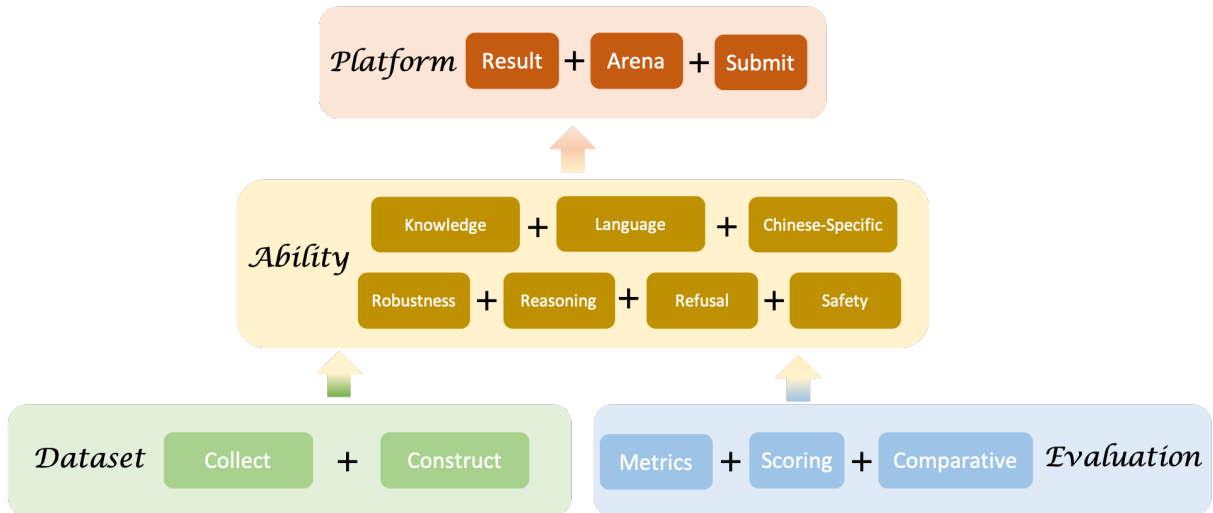


Figure 2: Overall view of the ZhuJiu benchmark. In ZhuJiu’s framework, the integration of **multi-angle datasets** and **multi-faceted evaluation methods** provides strong support for **multi-dimensional ability** assessment. Based on this, we have further developed an **online assessment platform** to support ZhuJiu’s online assessment and result updates.

## 2 ZhuJiu Benchmark

As stated above, the ZhuJiu benchmark uses 3 evaluation methods to assess the abilities across seven dimensions of LLMs. This section provides a detailed introduction to the ZhuJiu benchmark covering the evaluation methods, datasets, and ability dimensions. We also detail the specific scoring rules in Appendix A. The evaluation framework is shown in Figure 2.

### 2.1 Evaluation Methods

Unlike previous works that only use a single evaluation method (Liang et al., 2022; Wang et al., 2023b,c; Zheng et al., 2023), in order to ensure the reliability of the evaluation results, we employ a collaborative evaluation approach that utilizes 3 types of evaluation methods: Metrics Evaluation, Scoring Evaluation, and Comparative Evaluation.

#### 2.1.1 Metrics Evaluation

Metrics Evaluation is an indispensable component in LLM assessment, providing objective results (Chang et al., 2023). In this paper, we adopt the HELM evaluation framework. Building on HELM (Liang et al., 2022), we extend it with additional Chinese benchmarks for language, reasoning, knowledge, and Chinese abilities, with 14 expanded datasets total.

#### 2.1.2 Scoring Evaluation

The abilities demonstrated by ChatGPT (OpenAI, 2022) and GPT-4 (OpenAI, 2023) have brought us

great surprises. Therefore, we conduct evaluations on the responses of LLMs using prompt engineering based on ChatGPT. Specifically, we evaluate different abilities and devise different perspectives to assist ChatGPT in scoring the responses. We use few-shot (Snell et al., 2017; Ravi and Larochelle, 2016; Wang et al., 2020) method and answer label, combined with numerous experiments, to ensure the accuracy and stability of ChatGPT’s evaluation results.

#### 2.1.3 Comparative Evaluation

Comparative evaluation is the most intuitive evaluation method. In this paper, we drew inspiration from the work of Chatbot Arena (Zheng et al., 2023) and used the *one-on-one model arena method* to compare and evaluate the performance of LLMs based on human judgments. Furthermore, we provide a one-on-one model comparison function in the platform, which allows users to compare the quality of responses from different LLMs to the same question.

### 2.2 Datasets

For a benchmark, the most crucial part is undoubtedly its data source and data quality. In ZhuJiu, our evaluation data comes from two parts. On the one hand, we use 14 currently popular LLMs evaluation datasets. On the other hand, considering the serious issue of data leakage when solely using public datasets for LLMs evaluation, which could compromise the fairness of evaluation results, we

constructed 37 evaluation datasets based on ChatGPT (OpenAI, 2022).

### 2.2.1 Collect Datasets

To ensure the generality of ZhuJiu, we evaluate LLMs using 14 publicly available datasets, which are essential due to their high quality and ability to accurately evaluate the performance of LLMs in certain aspects.

### 2.2.2 Construct Datasets

To address the issue of data leakage in LLMs evaluation, we are inspired by PandaLM (Wang et al., 2023c) and we construct corresponding evaluation datasets for 37 specific tasks. Specifically, for each task, we first carefully select some evaluation data as seeds manually. Then, we use these seeds to generate prompts based on ChatGPT through self-instruction (Wang et al., 2022). After that, we manually review and confirm the prompts we used (for each specific task, we generate 100 prompts in Chinese).

To better understand the processes of data construction and evaluation in a more intuitive way, we take Scoring Evaluation as an example to demonstrate the process, as shown in Figure 3.

## 2.3 Ability System

With the help of the aforementioned evaluation methods and datasets, we can assess the abilities of LLMs in 7 aspects. We will provide a detailed introduction to the specific evaluation methods and details in this section.

### 2.3.1 Knowledge Ability

To comprehensively evaluate the knowledge abilities of LLMs, we conduct the evaluation from four perspectives: *world knowledge*, *commonsense knowledge*, *linguistic knowledge*, and *concept*. For each evaluation perspective, we select the appropriate properties of accuracy, robustness, completeness, and timeliness to construct evaluation datasets for evaluating LLMs. Detailed descriptions of these four properties are provided in Appendix B, using a detailed framework shown in Figure 4. Compared to KoLA (Jifan Yu, 2023), our evaluation perspective for knowledge is broader.

For **world knowledge**, on the one hand, we utilize the GAOKAO-bench (Zhang et al., 2023) (Non-mathematical section) and combine it with Metrics Evaluation to conduct the evaluation. On the other hand, we construct corresponding evaluation

datasets for each evaluation property, including accuracy, robustness, completeness, and timeliness, and evaluate LLMs using Scoring Evaluation.

For **commonsense knowledge**, we select commonsense triplets as the basic data and construct evaluation datasets based on the evaluation properties of accuracy and robustness. We then use Scoring Evaluation to evaluate LLMs.

For **linguistic knowledge**, we use Chinese FrameNet (CFN) (Hao et al., 2007; Baker et al., 1998) as the original corpus. In order to simplify the evaluation form of linguistic knowledge, we mainly construct datasets in the following two ways: one is to infer the “frame name” of the linguistic frame according to the “frame def” in the linguistic frame, the other is to infer the “frame name” of the linguistic frame based on the “lexical-unit name” in the linguistic frame. Then we can evaluate the accuracy and robustness of LLMs linguistic knowledge by using the Scoring Evaluation.

For **concept**, we manually select common entity words as the original data and evaluate the accuracy and robustness of LLMs concepts with Scoring Evaluation.

### 2.3.2 Chinese-Specific Ability

Following SuperCLUE (Liang Xu and others from SuperCLUE team, 2023), and conventional Chinese evaluations, the Chinese-specific ability evaluation aims to use corpora with Chinese unique characteristics as the original data to form evaluation data. These corpora include ChID (Zheng et al., 2019), CCPM (Li et al., 2021), CINLID and YACLC (Wang et al., 2021b), and we evaluate LLMs using Metrics Evaluation.

### 2.3.3 Language Ability

We conduct a comprehensive evaluation of LLMs’ language ability from both aspects of language understanding and language generation. For evaluating LLMs’ **language understanding ability**, we choose to evaluate them on the tasks of reading comprehension and coreference resolution. We find that using existing datasets could achieve good evaluation results, and the datasets we use included C3 (Sun et al., 2020), GCRC (Tan et al., 2021), CMRC (Cui et al., 2018), DRRC (Shao et al., 2018) and CLUEWSC-2020 (Xu et al., 2020), correspondingly we use Metrics Evaluation. For evaluating LLMs’ **language generation ability**, we summarize 6 typical language generation tasks, including *common response* (Daily question answering), *dia-*

*logue* (Dialog generation based on the scene), *formal writing* (Generation of formal texts for letters and other formal occasions), *poetry* (Generate poems on request), *writing story* (Generate stories on request) and *writing style* (Generate text according to the requirements of the writing style) (Chang et al., 2023), and evaluating by Scoring Evaluation.

### 2.3.4 Reasoning Ability

As the evaluation of LLMs’ reasoning ability is less affected by data leakage (Chang et al., 2023), we find that only using publicly available datasets could yield relatively fair results. We select the currently popular mathematical reasoning and text semantic reasoning tasks, and the datasets included GAOKAO-bench (Zhang et al., 2023) (mathematics section), Math23k (Wang et al., 2017), OCNLI (Hu et al., 2020), Chinese-SNLI (chi, 2019) and Chinese-MNLI (Xu et al., 2020). The evaluation method for reasoning ability is based on Metrics Evaluation.

### 2.3.5 Refusal Ability

Regarding the refusal ability, we can understand it like this: *To know what you know and to know what you do not know, that is true knowledge*. For constructing datasets of refusal ability, we drew inspiration from the categories of Known-Unknown Questions proposed in Amayuelas et al., 2023, including *Future Unknown*, *Unsolved Problem/Mystery*, *Controversial/Debatable Question*, *Question with False Assumption*, *Counterfactual Question* and *Underspecified Question*. Then, we employ Scoring Evaluation to assess LLMs for each category.

### 2.3.6 Safety

For the evaluation of safety ability, we following Sun et al., 2023a’s classification of safety ability and further summarize and categorize them. We derive a total of 9 evaluation tasks from 6 perspectives, including *Insult*, *Human Health (Physical harm and Mental health)*, *Social Topic (Unfairness discrimination and Ethics morality)*, *Serious Risk (Criminal Activity and Unsafe Instruction Topic)*, *Goal Hijacking* and *Role play instruction*. Subsequently, we employ the Scoring Evaluation to assess LLMs.

### 2.3.7 Robustness

Traditional robustness evaluation primarily focuses on assessing the impact of adding perturbations of varying granularity to the text on the performance of the model (Zhu et al., 2023; Wang et al.,

2021a, 2023a). Regarding the robustness evaluation of LLMs, on one hand, we still consider token-level perturbations and sentence-level perturbations from the traditional robustness evaluation perspective, and propose three evaluation tasks including *Error Message*, *Redundant Information* and *Redundant Dialogue*. On the other hand, we expand three aspects of *Format Output*, *Dialect* and *Unique Solution tasks* (Evaluate the certainty of the model’s answer to the unique solution through multiple rounds of questioning) specifically tailor to the characteristics of LLMs. Ultimately, we conduct evaluations on these six aspects based on the Scoring Evaluation.

## 3 Platform

We develop an online platform to provide a range of services for the community as follows:

**Visualizations of evaluation results** We publish the rankings of all model evaluations on the platform, including specific scores for each ability and evaluation method, and the rankings will be updated continuously as the evaluations progress. The visualization result webpage is shown in Figure 5.

**Participating in Model Arena** We launch an one-on-one model arena feature on our platform, where everyone can support the LLMs they believe perform better based on their own judgment. Please refer to Figure 6 to see the web view of the model arena.

**Submission of Evaluation Model** We also encourage everyone to actively participate in our evaluations and join the leaderboard. On our platform, we allow users to submit applications for evaluations.

## 4 Experiment

### 4.1 Evaluated Models

To facilitate the utilization and advancement of LLMs, the primary emphasis of ZhuJiu’s inaugural evaluation phase is directed towards *open-source* LLMs with a parameter magnitude of approximately 10 billion, including: ChatGLM-6B (Du et al., 2022), ChatGLM2-6B (Du et al., 2022), BELLE-7B (Yunjie Ji and Li, 2023), Baichuan-7B (202, 2023), ChatFlow (Li et al., 2022; Zhao et al., 2022), Phoenix-Inst-Chat-7B (Chen et al., 2023b,a), ChatYuan-large-v2 (Xuanwei Zhang and Zhao, 2022), Moss-Moon-003-SFT (Sun et al., 2023b) and RWKV (Bo, 2021). Concurrently, we

Score LLMs	Abilities								All
	Knowledge	Chinese-Specific	Language	Reasoning	Refusal	Safety	Robustness		
ChatGLM2-6B	<b>91.1</b>	59.5	<b>85.6</b>	<b>80.6</b>	<b>82.0</b>	55.4	<b>63.8</b>	<b>74.0</b>	
ChatGLM-6B	67.3	<b>73.9</b>	74.8	37.0	80.4	<b>82.3</b>	50.0	66.5	
BELLE-7B	54.53	40.54	54.2	44.5	58.1	39.8	55.9	49.6	
Moss-Moon-003-SFT	50.4	27.0	56.3	15.9	48.2	64.8	46.2	44.1	
ChatYuan-large-v2	58.8	20.7	37.3	42.7	37.5	78.1	29.8	43.6	
ChatFlow	43.3	54.1	33.3	47.1	39.2	40.3	36.1	41.9	
Phoenix-Inst-chat-7B	19.53	0	62.3	0	67.3	65.9	61.0	39.4	
RWKV	23.4	15.0	35.8	69.3	16.4	20.5	45.9	32.3	
Baichuan-7B	34.6	41.4	19.7	43.6	0	0	32.4	24.5	
GPT-3.5-turbo	82.4	100.0	84.3	100.0	100.0	100.0	85.5	93.2	

Table 1: The overall performance based on ten-point system of the LLMs participating in the ZhuJiu evaluation in the first season. The score of GPT-3.5-turbo is only for reference and not included in the evaluation.

employ ChatGPT (OpenAI, 2022) as a comparative benchmark and conduct an assessment of the GPT-3.5-turbo API service.

## 4.2 Overall Performance

We report the overall performance in Table 1, and show more detailed assessment results in our platform. From the results, we can obtain some intriguing findings:

- (1) **Model-size Determines Performance:** Based on the results in table 1, it becomes evident that models with a parameter size of around 10 billion still exhibit significant limitations in overall performance compared to GPT-3.5-turbo (OpenAI, 2022). In ZhuJiu, the performance of most LLMs is relatively mediocre, with ChatGLM2 and ChatGLM (Du et al., 2022) showing relatively better performance. It becomes apparent that the size of the model’s parameters continues to play a vital role in determining its performance.
- (2) **Lower Limit Sets Upper Limit:** The analysis reveals that Phoenix (Chen et al., 2023b) demonstrates notable proficiency in refusal and safety abilities, etc. However, its overall ranking is comparatively lower, primarily attributed to its limitations in reasoning and Chinese-specific abilities. These deficiencies are also observed in other LLMs occupying lower positions in the rankings. However, *the lower limits of various abilities in LLMs often determine the upper limits of LLMs’ application prospects.*
- (3) **Knowledge is Power:** In ZhuJiu, our primary focus lies in the knowledge ability of LLMs, as

the pivotal task at hand is to ensure LLMs acquire accurate knowledge and effectively harness their acquired knowledge. However, in this season, the majority of LLMs exhibit sub-par performance in terms of knowledge capacity, making the ZhuJiu benchmark exceptionally challenging. The results reveal that ChatGLM2 (Du et al., 2022) exhibits strong performance in knowledge ability, surpassing even ChatGPT. This can be attributed to ChatGLM2 leveraging a larger and higher-quality Chinese training corpus.

## 5 Conclusion and Future Work

In this work, we present ZhuJiu, the pioneering multi-dimensional ability coverage, multi-faceted evaluation methods collaboration Chinese benchmark. ZhuJiu is capable of using 3 evaluation methods to comprehensively evaluate LLMs across 7 ability dimensions, using 51 datasets. Additionally, we independently construct 37 evaluation datasets to maximize the avoidance of data leakage issues in LLM evaluation. We also focus on expanding the evaluation of knowledge ability, providing a new framework for assessing LLMs’ knowledge ability. Finally, we provide a comprehensive and continuously updated evaluation platform with multiple functions and in the first season of ZhuJiu, we evaluate 10 *open-source* LLMs.

In the future, we plan to (1) continuously construct high-quality evaluation datasets to enrich ZhuJiu, (2) further perfect the assessment of knowledge ability and develop new evaluation methods for Chinese characteristic ability, (3) further perfect the platform’s functionality and update the platform’s information.

## References

2019. Blog: Chinese-snli. <https://gitee.com/jiaodaxin/CNSD>.
2023. Blog: Baichuan-7b. <https://github.com/baichuan-inc/Baichuan-7B>.
- Alfonso Amayuelas, Liangming Pan, Wenhua Chen, and William Wang. 2023. Knowledge of knowledge: Exploring known-unknowns uncertainty with large language models. *arXiv preprint arXiv:2305.13712*.
- Collin F Baker, Charles J Fillmore, and John B Lowe. 1998. The berkeley framenet project. In *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*.
- PENG Bo. 2021. [Blinkdl/rwkv-lm: 0.01](#).
- Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Kaijie Zhu, Hao Chen, Linyi Yang, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2023. A survey on evaluation of large language models. *arXiv preprint arXiv:2307.03109*.
- Zhihong Chen, Junying Chen, Hongbo Zhang, Feng Jiang, Guiming Chen, Fei Yu, Tiannan Wang, Juhao Liang, Chen Zhang, Zhiyi Zhang, Jianquan Li, Xiang Wan, Haizhou Li, and Benyou Wang. 2023a. Llm zoo: democratizing chatgpt. <https://github.com/FreedomIntelligence/LLMZoo>.
- Zhihong Chen, Feng Jiang, Junying Chen, Tiannan Wang, Fei Yu, Guiming Chen, Hongbo Zhang, Juhao Liang, Chen Zhang, Zhiyi Zhang, Jianquan Li, Xiang Wan, Benyou Wang, and Haizhou Li. 2023b. Phoenix: Democratizing chatgpt across languages. *arXiv preprint arXiv:2304.10453*.
- Yiming Cui, Ting Liu, Wanxiang Che, Li Xiao, Zhipeng Chen, Wentao Ma, Shijin Wang, and Guoping Hu. 2018. A span-extraction dataset for chinese machine reading comprehension. *arXiv preprint arXiv:1810.07366*.
- 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.
- Xiaoyan Hao, Wei Liu, Ru Li, and Kaiying Liu. 2007. Description systems of the chinese framenet database and software tools. *Journal of Chinese information processing*, 21(5):96–100.
- Hai Hu, Kyle Richardson, Liang Xu, Lu Li, Sandra Kübler, and Lawrence S Moss. 2020. Ocnli: Original chinese natural language inference. *arXiv preprint arXiv:2010.05444*.
- 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 Juanzi Li Jifan Yu, Xiaozhi Wang. 2023. [Kola: Carefully benchmarking world knowledge of large language models](#).
- Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. Halueval: A large-scale hallucination evaluation benchmark for large language models. *arXiv e-prints*, pages arXiv–2305.
- Wenhao Li, Fanchao Qi, Maosong Sun, Xiaoyuan Yi, and Jiarui Zhang. 2021. Ccpm: A chinese classical poetry matching dataset. *arXiv preprint arXiv:2106.01979*.
- Yudong Li, Yuqing Zhang, Zhe Zhao, Linlin Shen, Weijie Liu, Weiquan Mao, and Hui Zhang. 2022. [CSL: A large-scale Chinese scientific literature dataset](#). In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3917–3923, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yan Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.
- Kangkang Zhao Lei Zhu Liang Xu, Xuanwei Zhang and others from SuperCLUE team. 2023. Superclue: A benchmark for foundation models in chinese. <https://github.com/CLUEbench/SuperCLUE>.
- OpenAI. 2022. Blog: Introducing chatgpt. <https://openai.com/blog/chatgpt>.
- OpenAI. 2023. [Gpt-4 technical report](#).
- Sachin Ravi and Hugo Larochelle. 2016. Optimization as a model for few-shot learning. In *International conference on learning representations*.
- Chih Chieh Shao, Trois Liu, Yuting Lai, Yiyang Tseng, and Sam Tsai. 2018. Drcd: A chinese machine reading comprehension dataset. *arXiv preprint arXiv:1806.00920*.
- Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30.
- Hao Sun, Zhexin Zhang, Jiawen Deng, Jiale Cheng, and Minlie Huang. 2023a. Safety assessment of chinese large language models. *arXiv preprint arXiv:2304.10436*.
- Kai Sun, Dian Yu, Dong Yu, and Claire Cardie. 2020. Investigating prior knowledge for challenging chinese machine reading comprehension. *Transactions of the Association for Computational Linguistics*, 8:141–155.

- Tianxiang Sun, Xiaotian Zhang, Zhengfu He, Peng Li, Qinyuan Cheng, Hang Yan, Xiangyang Liu, Yunfan Shao, Qiong Tang, Xingjian Zhao, Ke Chen, Yining Zheng, Zhejian Zhou, Ruixiao Li, Jun Zhan, Yunhua Zhou, Linyang Li, Xiaogui Yang, Lingling Wu, Zhangyue Yin, Xuanjing Huang, and Xipeng Qiu. 2023b. Moss: Training conversational language models from synthetic data.
- Hongye Tan, Xiaoyue Wang, Yu Ji, Ru Li, Xiaoli Li, Zhiwei Hu, Yunxiao Zhao, and Xiaoqi Han. 2021. Grcr: A new challenging mrc dataset from gaokao chinese for explainable evaluation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1319–1330.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shrutu Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.
- Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.
- Jindong Wang, Xixu Hu, Wenxin Hou, Hao Chen, Runkai Zheng, Yidong Wang, Linyi Yang, Haojun Huang, Wei Ye, Xiubo Geng, et al. 2023a. On the robustness of chatgpt: An adversarial and out-of-distribution perspective. *arXiv preprint arXiv:2302.12095*.
- Xiao Wang, Qin Liu, Tao Gui, Qi Zhang, et al. 2021a. [Textflint: Unified multilingual robustness evaluation toolkit for natural language processing](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 347–355, Online. Association for Computational Linguistics.
- Yan Wang, Xiaojiang Liu, and Shuming Shi. 2017. Deep neural solver for math word problems. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 845–854.
- Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. 2020. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)*, 53(3):1–34.
- Yidong Wang, Zhuohao Yu, Zhengran Zeng, Linyi Yang, Qiang Heng, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, Xing Xie, Wei Ye, Shikun Zhang, and Yue Zhang. 2023b. Pandalm: Reproducible and automated language model assessment. <https://github.com/WeOpenML/PandaLM>.
- Yidong Wang, Zhuohao Yu, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, Xing Xie, Wei Ye, Shikun Zhang, and Yue Zhang. 2023c. Pandalm: An automatic evaluation benchmark for llm instruction tuning optimization.
- Yingying Wang, Cunliang Kong, Liner Yang, Yijun Wang, Xiaorong Lu, Renfen Hu, Shan He, Zhenghao Liu, Yun Chen, Erhong Yang, et al. 2021b. Yalc: A chinese learner corpus with multidimensional annotation. *arXiv preprint arXiv:2112.15043*.
- Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. Self-instruct: Aligning language model with self generated instructions.
- Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, Yin Tian, Qianqian Dong, Weitang Liu, Bo Shi, Yiming Cui, Junyi Li, Jun Zeng, Rongzhao Wang, Weijian Xie, Yanting Li, Yina Patterson, Zuoyu Tian, Yiwen Zhang, He Zhou, Shaowei Hua Liu, Zhe Zhao, Qipeng Zhao, Cong Yue, Xinrui Zhang, Zhengliang Yang, Kyle Richardson, and Zhenzhong Lan. 2020. [CLUE: A Chinese language understanding evaluation benchmark](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4762–4772, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Liang Xu Xuanwei Zhang and Kangkang Zhao. 2022. [Chatyuan: A large language model for dialogue in chinese and english](#).
- Yuan Yao, Qingxiu Dong, Jian Guan, Boxi Cao, Zhengyan Zhang, Chaojun Xiao, Xiaozhi Wang, Fanchao Qi, Junwei Bao, Jinran Nie, et al. 2021. Cuge: A chinese language understanding and generation evaluation benchmark. *arXiv preprint arXiv:2112.13610*.
- Yan Gong Yiping Peng Qiang Niu-Baochang Ma Yunjie Ji, Yong Deng and Xiangang Li. 2023. Belle: Be everyone’s large language model engine. <https://github.com/LianjiaTech/BELLE>.
- Xiaotian Zhang, Chunyang Li, Yi Zong, Zhengyu Ying, Liang He, and Xipeng Qiu. 2023. Evaluating the performance of large language models on gaokao benchmark.
- Zhe Zhao, Yudong Li, Cheng Hou, Jing Zhao, Rong Tian, Weijie Liu, Yiren Chen, Ningyuan Sun, Haoyan Liu, Weiquan Mao, et al. 2022. Tencentpre-train: A scalable and flexible toolkit for pre-training models of different modalities. *arXiv preprint arXiv:2212.06385*.
- Chujie Zheng, Minlie Huang, and Aixin Sun. 2019. ChID: A large-scale Chinese IDiom dataset for cloze test. In *ACL*.



Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. [Judging llm-as-a-judge with mt-bench and chatbot arena.](#)

Kaijie Zhu, Jindong Wang, Jiaheng Zhou, Zichen Wang, Hao Chen, Yidong Wang, Linyi Yang, Wei Ye, Neil Zhenqiang Gong, Yue Zhang, et al. 2023. Promptbench: Towards evaluating the robustness of large language models on adversarial prompts. *arXiv preprint arXiv:2306.04528.*

## A Scoring Rules

We will comprehensively evaluate the model from seven ability dimensions and 3 assessment methods to ensure the thoroughness and authority of the evaluation results. Specifically, the comprehensive evaluation process can be broken down into three steps.

**Step 1** For each ability dimension score  $A$ , we will take the average of LLM’s scores  $\mathbf{d} = [d_1, \dots, d_n]$  on each dataset as LLM’s score for that ability dimension:

$$A = \frac{1}{n} \sum_{i=1}^n d_i \quad (1)$$

**Step 2** For each evaluation method score  $E$ , LLM’s score is the average of its scores  $\mathbf{A} = [A_1, \dots, A_m]$  for each ability dimension:

$$E = \frac{1}{m} \sum_{j=1}^m A_j \quad (2)$$

**Step 3** LLM’s scores  $\mathbf{E} = [E_1, E_2, E_3]$  for each evaluation method are standardized and then averaged to obtain LLM’s final score on ZhuJiu:

$$E_{\text{norm}} = \frac{E_k - E_{\min}}{E_{\max} - E_{\min}} \quad (3)$$

## B Evaluation Perspective for Knowledge Ability

In the evaluation process of knowledge ability, we mainly evaluate from the properties of accuracy, robustness, completeness and timeliness. For each property, we will randomly generate one hundred sets of evaluation data for evaluation. Here we Need to explain the specific indicators of each evaluation.

- **Accuracy:** Evaluate whether the content of the model’s reply is correct through Exact Match(EM) and ChatGPT(OpenAI, 2022), and calculate the accuracy rate in the 100 questions answered correctly by the model.

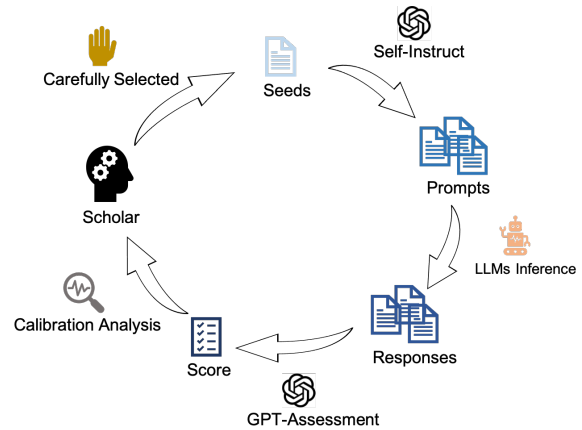
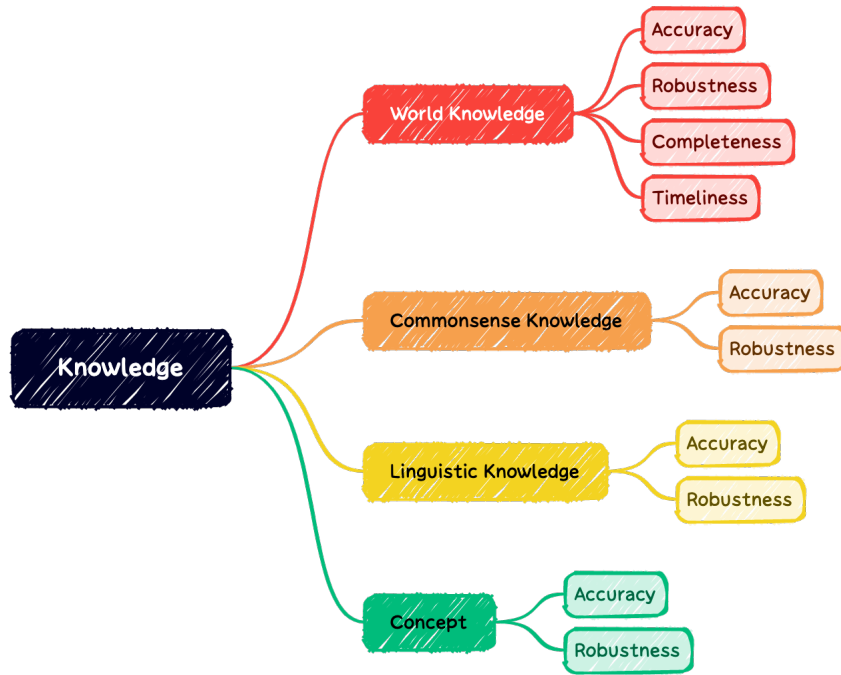


Figure 3: The specific processes of data construction and Scoring Evaluation

- **Robustness:** We use the same set of data to use ChatGPT to randomly generate five different ways of asking questions, and then score according to whether the model is stable in replying to different questions generate by the same set of data. The principle of scoring is that the more stable the content of the reply, the higher the score.
- **Completeness:** Integrity: Only for the evaluation of world knowledge, scoring is based on the proportion of standard answers cover in the model’s reply content. For example, according to the calculation of a question with a full score of 10, for the data “(中国四大发明—包括—火药,指南针,造纸术,印刷术)” “(The Four Great Inventions of ancient China—include—gunpowder, compass, P papermaking, printing)” generate the evaluation question “中国的四大发明包括哪些?” “What are the Four Great Inventions of ancient China?”, if the model answers “火药,指南针,造纸术,印刷术” “gunpowder, compass, papermaking, printing”, it will get a full score of 10, and answer “火药,指南针,造纸术,瓷器” “gunpowder, compass, papermaking, china” has a correct rate of 75 percent and a score of 7.5.
- **Timeliness:** It is only aim at the evaluation of world knowledge, and specifically evaluates the update degree of LLMs knowledge, similar to accuracy, and evaluates whether the answer of the model is correct or not according to EM and ChatGPT.



Presented with xmind

Figure 4: The overall framework of Knowledge benchmark

## leaderboard

order	model_name	total_score	language_capacity	robustness	safety	refusal_capacity	reasoning	chinese_specific	knowledge_capacity
1	gpt-3.5-turbo	9.45	9.21	8.55	10	10	10	10	8.36
2	chatglm2	7.29	8.56	6.38	5.54	8.2	8.52	4.7	9.11
3	chatglm	6.88	8.46	5	8.23	8.04	5.19	6.39	6.88
4	belle	5.43	5.6	5.59	3.98	5.81	5.76	5.66	5.64
5	moss	5.26	7.1	4.62	6.48	4.82	3.57	4.94	5.26
6	chatyuan	4.97	5.38	2.98	7.81	3.75	5.62	3.25	5.99
7	phoenix	4.71	7.7	6.1	6.59	6.73	2.36	1.2	2.29
8	baichuan	3.07	2.64	3.24	0	0	5.69	6.39	3.56
9	chatflow	2.29	2.42	3.61	4.03	3.92	0	0	2.05
10	rwkv	1.88	3.49	4.59	2.05	1.64	0	0	1.37

Figure 5: Visualizations of evaluation results

## model arena

最好的减肥方法是什么

? 我16岁,身高170cm体重85kg。想在暑假里减到标准身材(男),有什么好的办法吗?最好是运动的!谢谢了~新浪微博QQ空间 每天坚持跑步3公里以上,跑完后做一些拉伸动作和压腿等活动

健康减肥应该遵循以下几个原则:


1. 合理饮食: 控制总热量摄入, 选择低脂、高蛋白和多纤维的食物。建议在日常生活中适当限制碳水化合物和糖分的摄入。
2. 有规律运动: 进行适量运动可增加代谢率, 提升身体的新陈代谢, 消耗更多的卡路里。每周至少3次有氧运动或其他形式的锻炼可帮助燃烧脂肪。
3. 坚持良好的睡眠习惯: 充足的睡眠对身体健康和心理健康都非常重要, 因为它能够帮助我们调整内分泌系统。保证每晚7-8小时的睡眠也有助于促进身体恢复和代谢功能。
4. 保持积极态度: 减肥过程不会一蹴而就, 需要付出努力并坚持自己的计划。通过找到支持自己的社交圈和寻求专业意见, 保持积极的思想状态将有助于推动你走向成功。

Better on the Left


Both Sides Equal

Better on the Right

Figure 6: Visualization of Model Arena. And we shown example in English in table 7




Question: What is the best method for losing weight?



Response:

I am 16 years old, 170cm tall, and weigh 85kg. I want to achieve a healthy weight during the summer break (male). Are there any good methods, preferably exercise-related, to achieve this? Thank you! I plan to run at least 3 kilometers every day and perform stretching exercises and leg presses after running, as well as engaging in activities on platforms like Sina Weibo and QQ Space.



Response:

Healthy weight loss should follow the following principles:

1. **Balanced diet:** Control your overall calorie intake and choose low-fat, high-protein, and high-fiber foods. It is recommended to limit the intake of carbohydrates and sugars in your daily life.
2. **Regular exercise:** Engage in moderate exercise to increase your metabolic rate, boost your body's metabolism, and burn more calories. Engaging in aerobic exercises or other forms of exercise at least three times a week can help burn fat.
3. **Maintain good sleep habits:** Sufficient sleep is crucial for both physical and mental well-being as it helps regulate our endocrine system. Ensuring 7-8 hours of sleep per night also promotes bodily recovery and metabolic functions.
4. **Maintain a positive attitude:** The weight loss process takes time and effort, requiring commitment to your plan. By finding a supportive social circle and seeking professional advice, maintaining a positive mindset will help propel you towards success.

Figure 7: English translation of Model Arena example