ZhuJiu-Knowledge: A Fairer Platform for Evaluating Multiple Knowledge Types in Large Language Models

Pengfan Du[1,2*], **Sirui Liang**[1*], **Baoli Zhang**[1*], **Pengfei Cao**[1],
Yubo Chen[1,2†], **Kang Liu**[1,2,3†] and **Jun Zhao**[1,2]
[1]Institute of Automation, Chinese Academy of Sciences
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences
[3]Shanghai Artificial Intelligence Laboratory
{baoli.zhang,pengfei.cao,yubo.chen,kliu,jzhao}@nlpr.ia.ac.cn
{dupengfan22}@mails.ucas.ac.cn, liangsirui2024@ia.ac.cn

Abstract

The swift advancement in large language models (LLMs) has heightened the importance of model evaluations. LLMs have acquired a substantial amount of knowledge, and evaluating the knowledge of these LLMs is crucial. To address this, we introduce the ZhuJiu-Knowledge benchmark which carefully considers the following factors: (1) **For knowledge scope**, we concentrate on three domains: commonsense knowledge, world knowledge, language knowledge, which comes from ATOMIC, Conceptnet, Wikidata, and Wordnet. (2) **For data construction**, to prevent data contamination, we utilize knowledge derived from corpora and knowledge graphs to formulate novel questions that are ensured not to appear in the training corpus. A multitude of prompts is purposefully devised to mitigate the impact of prompt design on evaluation and to further analyze the LLMs' sensitivity to various prompts. (3) **For evaluation criteria**, we propose a novel voting methodology for assessing generative text, aligning the model's evaluation with human preferences to reduce biases inherent in individual model assessments. We evaluate 14 current mainstream LLMs and conduct a comprehensive discussion and analysis of their results. The ZhuJiu-Knowledge benchmark and open-participation leaderboard are publicly released at http://zhujiu-knowledge.top/ and we also provide a demo video at https://youtu.be/QJp4qlEHVH8.

1 Introduction

The unprecedented performance of LLMs, such as GPT4 (Achiam et al., 2023) and Llama2 (Touvron et al., 2023), has garnered significant attention and made their evaluation a focal point as the field progresses(Guo et al., 2023; Chang et al., 2023). LLMs have acquired a substantial amount of knowledge, and evaluating the knowledge of these LLMs is crucial. Existing efforts have focused on evaluating the knowledge stored within the model (e.g., Petroni et al., 2019; Yu et al., 2023). However, these works still face several challenges.

Constructing a reasonable benchmark for evaluating knowledge involves careful consideration of several key factors: (1) **Knowledge Scope**. Most evaluations are limited to world knowledge related to entities and relations, lacking assessments of commonsense knowledge and language knowledge (Liang et al., 2022; Zhang et al., 2023). This limitation arises because these two types of knowledge are often expressed in the form of events or sentences. Currently, making unbiased evaluations of the generated sentences is still a difficult problem for LLMs. (2) **Data Constructinon**. Most evaluation platforms have limitations on assessing data. Firstly, the evaluated data usually are leaked and embedded in the target LLMs in the process of pre-training or SFT, after such data are released publicly. As a result, making an evaluation on such data would be biased (Brown et al., 2020; Zhou et al., 2023a). Secondly, existing knowledge probing strategies usually relied on the given prompts heavily. Existing methods only used just one prompt for each piece of knowledge. If the target LLM does not understand the given prompt well, it will not obtain better results (Webson and Pavlick, 2021; Abdou et al., 2022). (3) **Evaluation Criteria**. Evaluating knowledge using multiple-choice questions and true or false questions may introduce certain biases obviously. Assessing generated text with QA questions requires a reasonable evaluation metric for the generated content. Traditional evaluation criteria such as GLUE (Wang et al., 2018), ROUGE (Lin, 2004), and RECALL have inherent limitations, often leading to a gap between evaluation results and users' subjective experiences. While manual evaluation is highly

[1]*Co-first authors, they contributed equally to this work.

[2]†Corresponding author

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: System Demonstrations), pages 194–206
June 16-21, 2024

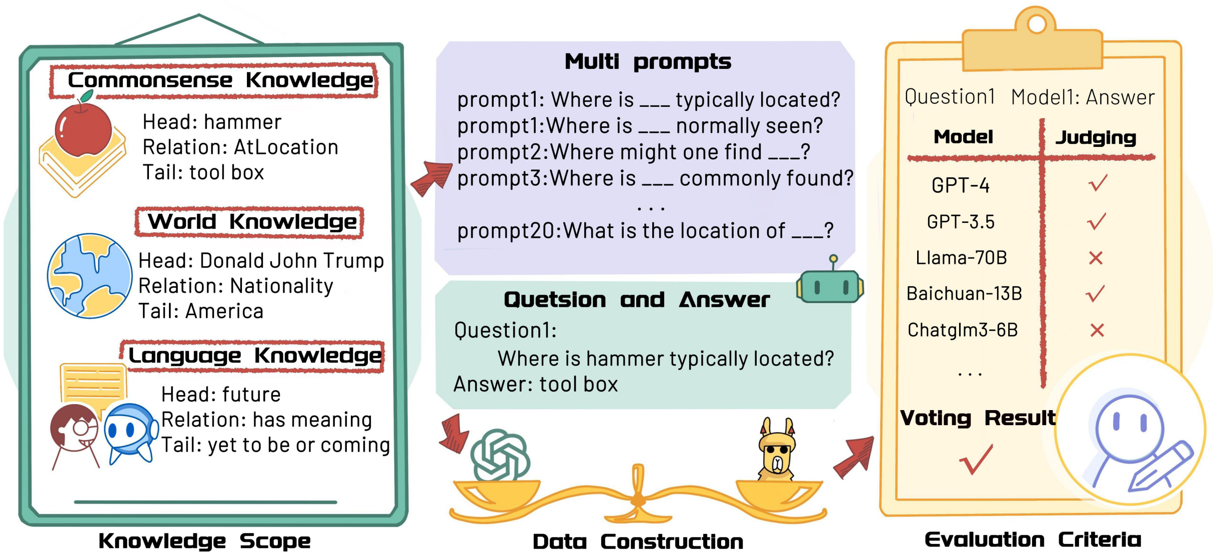

Figure 1: The evaluation process of LLM using ZhuJiu-Knowledge.

reliable, it is time-consuming and labor-intensive (Karpinska et al., 2021). Therefore, designing a rational evaluation metric is also a critically important issue.

This paper constructs ZhuJiu-Knowledge, a fairer platform for evaluating multiple Knowledge types in LLMs, which is designed to assess LLMs' capabilities in commonsense knowledge, world knowledge, and language knowledge. During the construction process, we fully consider the aforementioned three factors.

For Knowledge Scope, we choose to evaluate commonsense knowledge, world knowledge, and language knowledge for the following reasons: (i) Commonsense knowledge is often implicitly embedded in the texts, which presents a challenge for LLMs to acquire substantially from textual corpora. A proficient understanding of commonsense is fundamental for LLMs to generate reasonable responses (Davis, 2023). (ii) World knowledge plays a critical role in LLMs' performance. Based on memory and manipulation for world knowledge, LLMs can provide more accurate and relevant responses to user questions (Hendrycks et al., 2020; Zhong et al., 2023). (iii) A deep understanding of grammar, semantics, and pragmatics enables LLMs to grasp language nuances and process text effectively (Chuang et al., 2020; Dentella et al., 2023).

For data construction, we simultaneously address data contamination(Zhou et al., 2023b; Sainz et al., 2023) and prompt sensitivity issues. (i) To address data contamination, some studies adopt machine-unreadable (Huang et al., 2023) data or constantly evolving data. Considering the pressing demand for high-quality data in the training process of LLMs, it can be foreseen that these data will also be trained by LLMs in the near future. To overcome this limitation, the paper proposes to leverage knowledge from existing knowledge graphs and corpora to construct evaluation questions. The advantage of this method is to allow for the automatic generation of questions tailored to the targeted knowledge domains. (ii) To better assess LLMs' prompt sensitivity, we propose designing multiple prompts for each piece of knowledge. For each knowledge, we design multiple prompts to investigate whether an LLM can consistently generate correct answers when given varying prompts

For evaluation criteria, conventional methods typically relied on traditional evaluation metrics or more advanced models. However, these methods often fail to fully capture LLMs' comprehensive abilities and users' subjective experiences (Li et al., 2023; Chiang and Lee, 2023), while purely manual evaluations are labor-intensive. To bridge the gap between objective metrics and subjective preferences, we propose an innovative approach integrating human-computer collaboration, which employs a vote-like evaluation strategy. Multiple LLMs are employed to evaluate the given answers, where some pivot LLMs are selected as the evaluation models because they have higher correlations with human evaluations. The final result is obtained through voting.

We also release an online evaluation platform that supports multiple functions including visualizations of evaluation results and submission of evaluation models, etc. Moreover, we evaluate 14 publicly available LLMs. Based on the experimental results, we obtain some intriguing observations and summarize them in Section 4.

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

- We construct ZhuJiu-Knowledge, a fairer platform for evaluating commonsense knowledge, language knowledge, and world knowledge in LLMs.
- We present a novel benchmark construction technique to evaluate LLMs in commonsense knowledge, world knowledge, and language knowledge. This benchmark is designed to reduce data contamination and prompt sensitivity, accompanied by a human-aligned evaluation strategy to yield more reliable results.
- Using the ZhuJiu-Knowledge benchmark, we evaluate 14 current LLMs across three types of knowledge, providing insights for the improvement and enhancement of LLM.

2 ZhuJiu-Knowledge Benchmark

As stated above, the ZhuJiu-Knowledge benchmark conducts precise evaluations of models across three knowledge types. This section provides a detailed introduction to the ZhuJiu-Knowledge benchmark covering the knowledge selection, task design, data construction, and evaluation method. We also detailed the process of problem construction in Appendix A. The Evaluation framework is shown in Figure 1.

2.1 Knowledge Scope

The remarkable proficiency exhibited by the LLM has propelled evaluations towards more challenging and diverse tasks. The excessive pursuit of broadening task coverage has resulted in the neglect of knowledge itself within the evaluation process (Suzgun et al., 2022). To address these challenges, we advocate for classifying knowledge, designing tasks tailored to different knowledge types, and evaluating the model's ability in this knowledge with finer granularity. We have chosen the following three kinds of knowledge as our focus of assessment.

2.1.1 Commonsense Knowledge

We start with two commonsense knowledge graphs, Atomic (Hwang et al., 2021) and ConceptNet (Speer et al., 2017). (i) Atomic captures diverse relations on every day and inferences about others' mental states in symbolic form, it is represented by triples, e.g., *(PersonX adopts a dog, xNeed, to go to the pet store.)* (ii) ConceptNet encompasses the general attributes of common entities, also represented as triplets, e.g., *(single rose, HasProperty, beautiful)*. By contextualizing knowledge from these knowledge graphs, we transform abstract knowledge into specific questions, thereby assessing the model's ability to comprehend and apply this abstract knowledge in a contextualized manner.

2.1.2 World Knowledge

This paper employs Wikidata[1] as the knowledge source for generating evaluation queries. This knowledge base is renowned for its high quality and comprises billions of data triples. Additionally, we have introduced two refined evaluation metrics to gauge the model's knowledge proficiency. (i) To evaluate the model's understanding of knowledge at different frequencies, we have classified knowledge into three categories: (1) High-frequency Knowledge, characterized by interlinking occurrences exceeding 1000; (2) Common-frequency Knowledge, with interlinking occurrences between 100 and 1000; and (3) Low-frequency Knowledge, where interlinking occurrences are fewer than 100. (ii) To access LLMs' sensitivity to timeliness, temporally relevant knowledge triples are selected.

2.1.3 Language Knowledge

Our evaluation of language knowledge is divided into three aspects: semantics evaluation, syntax evaluation, and pragmatics evaluation, with WordNet (Miller, 1995) serving as the knowledge corpus for language knowledge. WordNet, a comprehensive English lexical database, organizes vocabulary based on sense and establishes a semantic network through word relationships (Miller et al., 1990). Given the extensive scope of syntax, semantics, and pragmatics, we have designed four knowledge corpus-based tasks (tasks 1-4) and four combined natural language understanding (NLU) and natural language generation (NLG) tasks (tasks 5-8) for the knowledge capability evaluation of large

[1] https://www.wikidata.org

language models (LLMs). (i) For semantics evaluation, we establish *(1) Semantic Selection Task* to choose or provide sentences that contain a specific meaning of a polysemous word. *(2) Similar Words Task* to provide near-synonyms for a particular sense of a word. *(3) Idiom Explanation Task* to explain the meanings of idioms. (ii) For syntax evaluation, we establish *(4) Part-of-speech Analysis Task* to analyze the parts of speech and meanings of words within sentences. *(5) Grammatical Correction Task* to correct grammatical errors in sentences. (iii) For pragmatics evaluation, we establish *(6) Specialized Formats Task*, which entail completing writing tasks in specific formats. *(7) Sentiment Analysis Task* to analyze the emotions of characters in sentences. *(8) Writing Style Task* to complete text generation following specific stylistic writing guidelines.

2.2 Data Construction

Ensuring fairness and objectivity is paramount in constructing reliable evaluation methods for LLMs. Our approach to data construction addresses two critical issues: data leakage and prompt sensitivity.

2.2.1 Knowledge-based Question Generation

Previous research has highlighted the severity of data contamination, where models answer questions through memorization rather than genuine knowledge mastery (Marie, 2023; Li, 2023; Sainz et al., 2023). Many existing benchmarks, e.g., CLEVA (Li et al., 2023), keep the evaluation data confidential from users during the assessment process. However, with the abundance of training data for LLMs, such practices cannot guarantee that the data has not been exposed during training. Kola (Yu et al., 2023) addresses this issue by employing continuously evolving data. However, it is foreseeable that such data will soon be utilized in training large models, potentially diminishing this approach's effectiveness.

We propose a knowledge-based question construction method using knowledge from the aforementioned knowledge bases, wherein the knowledge is typically formulated as triplets, denoted as <head, relation, tail>. We design question templates for relation, into which we insert heads from the triplets, thereby creating questions with answers. Additionally, recognizing the unsuitability of this approach for evaluating certain specialized tasks, such as some pragmatic tasks in language knowledge, we use GPT-4 to customize evaluation questions specifically for these types of tasks.

2.2.2 Prompt-based Question Expansion

An ideal large language model should be able to comprehend various prompts for the same knowledge. Nevertheless, current research indicates that the sensitivity of LLMs to prompts significantly influences their performance (Wei et al., 2022). It is reasonable to assert that evaluating LLMs using a single prompt may introduce bias in the evaluation results due to the sensitivity of different LLMs to the prompt. However, most current evaluation methods have not taken this issue into consideration.

To address the aforementioned issues, we propose a prompt-based question expansion method. This method involves the augmentation of a set of question prompts through the utilization of four advanced models, namely Llama-2-70B, Claude-instant, GPT3.5-turbo[2], and GPT-4. For each prompt set, we manually selected 20 prompts that are both universal and diverse. By utilizing the generated prompts, we construct a variety of questions targeting the same knowledge. A detailed description of the question construction process is shown in Appendix A. Furthermore, we explore which models exhibit heightened robustness in responding to prompts based on their performance on these questions.

To assess the sensitivity of different models to various prompts, we computed the entropy for each LLM's responses to different prompts for the same knowledge. Specifically, upon completing multiple-choice questions, we obtained $\Omega = \{\ldots, \omega_j, \ldots, \omega_k\}$ clusters for the n responses and the number of each cluster ω_k is $c(\omega_k)$. We calculate the entropy of the answer distribution as:

$$\text{entropy}(R(q)) = \sum_j \frac{c(\omega_j)}{n} \log\left(\frac{c(\omega_j)}{n}\right) \quad (1)$$

The entropy measures the degree of divergence between the responses for different prompts of the same knowledge. A higher entropy value indicates greater randomness in the answering process, which is associated with the model's uncertainty regarding that particular knowledge.

2.3 Evaluation Criteria

To alleviate biases from individual model evaluations and align assessment results more closely

[2] https://platform.openai.com/overview

with human perception, our evaluation system integrates two key components: multi-model voting evaluation and manual alignment.

2.3.1 Multi-model Voting Evaluation

Owing to the variations in knowledge scope among different models, evaluating results outside a model's knowledge range can introduce biases (Zhao et al., 2023).

We advocate the adoption of collaborative evaluation involving multiple LLMs to mitigate such potential biases. Specifically, the responses from different models to the same question form a set of answers $A = \{a_i\}_{i=1}^{|A|}$ for a question and an evaluation model ensemble $M = \{m_j\}_{j=1}^{|M|}$, we can obtain a set of evaluation results $R = \{(m_1^i), (m_2^i), \ldots, (m_j^i)\}_{i=1}^{|A|}$ for each answer. Then the final evaluation result of each answer is $F = argmax(R)$.

Considering the ultimate goal of LLMs is to cater to human needs, ensuring the generation of evaluation that aligns with human preferences becomes a paramount expectation. We incorporate a manual alignment step into the model evaluation process to achieve this. Specifically, for each type of knowledge, we selected some questions for manual evaluation. Subsequently, we adopt Pearson correlation coefficients between the manual evaluation results set and model evaluation results set R to measure the evaluation performance of different models. Considering that different LLMs excel in different types of knowledge, we selected the top five models of each knowledge type that are most closely aligned with manual evaluation results. The outcome will be determined by a voting process involving these five models. This approach aims to enhance the model's ability to produce evaluations that resonate more closely with human preferences and expectations.

2.4 Scoring Method

Each knowledge has different target views to evaluate LLMs' performance. For example, we have common-frequency, high-frequency, low-frequency, and timeliness to evaluate world knowledge. Since the metrics of different tasks are incomparable and differently sensitive, different results cannot be directly merged. Therefore, we propose to introduce Min-Max normalization to get a unified ranking of LLMs. Specifically, given a model set $M = \{m_j\}_{j=1}^{|M|}$ and the task set $D = \{d_j\}_{j=1}^{|D|}$, we get accuracy matrix a_{ij}. We first compute the Min-Max normalization of a_{ij} as z_{ij}.

$$z_{ij} = 100\frac{a_{ij} - min(a_j)}{max(a_j) - max(a_j)} \tag{2}$$

Then the standardized score S can be calculated as:

$$S_i = avg(z_i) \tag{3}$$

We use S_{QA} to represent the QA question score, and S_{CQ} to represent the Choices Question score.

3 Platform

We have developed an online platform that offers a diverse array of services to the community, as outlined below:

Evaluation process and questions We provide a detailed introduction for our evaluation process in Figure 5 and present a subset of evaluation questions in Figure 7.

Visualizations of evaluation results We show the overall scores (detailed in Figure 6) and metric scores of LLMs across three knowledge assessments (detailed in Figure 2.4), comprehensively analyzing the LLMs' strengths and weaknesses of each type of knowledge.

Submission of Evaluation Model We also invite all participants to engage actively in our evaluations and contribute to the leaderboard.

4 Experiment

4.1 Evaluated Models

In order to promote the development of LLMs, our primary focus for evaluation lies in open-source models with a parameter scale of approximately 10 billion, including ChatGLM3-6b, Baichuan2-13B-Chat (Yang et al., 2023), Baichuan-13B-Chat, Baichuan2-7B-Chat, Qwen-7B-Chat (Bai et al., 2023), Qwen-14B-Chat, Yi-6B-Chat (AI et al., 2024), WizardLM-13B-V1.2 (Xu et al., 2023), Vicuna-7b-v1.5 (Zheng et al., 2024), Vicuna-13B-v1.5 (Zheng et al., 2024), LLaMa2-7b-Chat, LLaMa2-13b-Chat, Mistral-7B-Instruct-v0.2 (Jiang et al., 2023), Mistral-7B-Instruct-v0.1.

4.2 Overall Performance

We report the overall performance in Table 1. S_{CQ} and S_{QA} represent the Choices question score and the QA question score respectively, which are defined in Section 2.4. H represents prompt sensitivity, which is defined in Equation 1. A more detailed

Metrics / Knowledge / LLMs	Commonsense Knowledge			World Knowledge			Language Knowledge		
	S_{CQ}	S_{QA}	H	S_{CQ}	S_{QA}	H	S_{CQ}	S_{QA}	H
Yi-6B-Chat	85.54	19.62	0.49	87.53	50.62	0.34	70.65	92.45	0.21
ChatGLM3-6b	38.88	77.4	0.55	46.63	19.21	0.55	58.56	71.07	0.22
WizardLM-13B-V1.2	24.11	90.59	0.45	27.06	67.13	0.70	71.64	**94.5**	0.30
Baichuan-13B-Chat	16.38	88.84	0.47	12.96	64.7	0.85	59.78	80.61	0.27
Baichuan2-13B-Chat	43.28	57.32	0.53	54.79	47.84	0.58	72.02	69.11	0.25
Baichuan2-7B-Chat	41.95	**93.21**	0.53	68.95	50.53	0.87	71.21	15.17	0.27
Vicuna-13b-v1.5	72.08	24.72	0.44	87.66	**78.06**	0.42	73.02	93.98	0.25
Vicuna-7b-v1.5	64.78	47.41	0.46	52.01	59.25	0.61	51.8	89.67	0.27
LLaMa2-13b-chat	22.3	67.76	0.43	79.6	37.75	0.43	64.2	89.02	0.24
LLaMa2-7b-chat	25.85	65.77	0.47	6.32	0	0.87	52.28	54.16	0.30
Mistral-7B-Instruct-v0.1	68.63	21.01	0.42	27.4	36.71	0.67	62.82	61.67	0.24
Mistral-7B-Instruct-v0.2	78.85	68.99	0.47	**96.64**	74.38	0.31	86.81	89.57	0.22
Qwen-14B-Chat	**96.78**	42.06	0.48	91.44	51.54	0.37	**93.62**	91.61	0.18
Qwen-7B-Chat	54.1	15.49	0.48	68.68	46.67	0.46	80.36	89.05	0.22

Table 1: The overall performance based on three knowledge abilities of the LLMs participating in the ZhuJiu-Knowledge evaluation. S_{CQ}: multiple choice question score, S_{QA}: QA question score, H: entropy.

assessment result can be found on our platform. The results reveal several noteworthy findings:

(1) **Evaluating with QA questions is fairer:** We compare the results of model responses to multiple-choice questions and QA questions across three knowledge abilities, revealing a significant difference. This suggests that LLMs may make correct choices through random selection or co-occurrence frequency calculation, which does not indicate that the LLM has mastered the knowledge. Employing multiple-choice questions as the evaluation method for models can lead to bias.

(2) **LLMs exhibit a preference for their own generated text:** We compare the evaluation results that the model assigned to itself with the evaluations from other LLMs. The specific calculation method and results can be found in Appendix B. The experimental results indicate that the model's self-assigned evaluations are significantly better than the evaluations it receives from other models. This suggests that models exhibit a clear preference for the text they generate, emphasizing the importance of using a voting mechanism for a fair evaluation.

(3) **LLMs is more prompt sensitive to hard question:** Entropy score represents prompt sensitivity, with higher entropy indicating greater model sensitivity. Table 1 shows that LLMs are most sensitive to world knowledge, followed by commonsense knowledge and then language knowledge. This sensitivity ranking aligns with the difficulty of the tasks, as depicted in Figures 2, 3, and 4. LLMs perform the poorest in the world knowledge task, followed by commonsense knowledge and language knowledge. This suggests that prompt sensitivity is higher in challenging tasks, emphasizing the need for carefully designed prompts to improve performance.

5 Conclusion and Future Work

In this work, we presented ZhuJiu-Knowledge, a fairer benchmark for evaluating multiple knowledge types of LLMs. Zhujiu-Knowledge extends the current knowledge evaluation scope to commonsense knowledge, language knowledge, and world knowledge. We introduce a novel data construction methodology that mitigates the risks of data contamination and prompt sensitivity and proposes a novel voting methodology to evaluate generative text. Our comprehensive evaluation of 14 mainstream LLMs provides significant insights into their performance, revealing the strengths and weaknesses of each model of various knowledge abilities. Finally, we provide a comprehensive knowledge evaluation platform for LLMs in the ZhuJiu-Knowledge.

In the future, our objectives include: (1) broadening the scope of the ZhuJiu-Knowledge benchmark to cover a wider array of knowledge assessment dimensions; and (2) enhancing our evaluation platform by integrating more features and improving the user interface, which will facilitate more efficient and user-friendly assessments.

6 Acknowledgements

This work is supported by the National Key Research and Development Program of China (No. 2022ZD0160503), the National Natural Science Foundation of China (No. 62176257). This work is also supported by the Youth Innovation Promotion Association CAS.

References

Mostafa Abdou, Vinit Ravishankar, Artur Kulmizev, and Anders Søgaard. 2022. Word order does matter and shuffled language models know it. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6907–6919.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

01. AI, :, Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Tao Yu, Wen Xie, Wenhao Huang, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng Nie, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai, Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. 2024. Yi: Open foundation models by 01.ai.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2023. A survey on evaluation of large language models.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? *arXiv preprint arXiv:2305.01937*.

Yung-Sung Chuang, Shang-Yu Su, and Yun-Nung Chen. 2020. Lifelong language knowledge distillation.

Ernest Davis. 2023. Benchmarks for automated commonsense reasoning: A survey. *arXiv preprint arXiv:2302.04752*.

Vittoria Dentella, Elliot Murphy, Gary Marcus, and Evelina Leivada. 2023. Testing ai performance on less frequent aspects of language reveals insensitivity to underlying meaning.

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.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, et al. 2023. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *arXiv preprint arXiv:2305.08322*.

Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021. (comet-) atomic 2020: on symbolic and neural commonsense knowledge graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6384–6392.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Marzena Karpinska, Nader Akoury, and Mohit Iyyer. 2021. The perils of using mechanical turk to evaluate open-ended text generation. *arXiv preprint arXiv:2109.06835*.

Yanyang Li, Jianqiao Zhao, Duo Zheng, Zi-Yuan Hu, Zhi Chen, Xiaohui Su, Yongfeng Huang, Shijia Huang, Dahua Lin, Michael R Lyu, et al. 2023. Cleva: Chinese language models evaluation platform. *arXiv preprint arXiv:2308.04813*.

Yucheng Li. 2023. An open source data contamination report for llama series models. *arXiv preprint arXiv:2310.17589*.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Benjamin Marie. 2023. The decontaminated evaluation of gpt-4.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

George A Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J Miller. 1990. Introduction to wordnet: An on-line lexical database. *International journal of lexicography*, 3(4):235–244.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. 2019. Language models as knowledge bases?

Oscar Sainz, Jon Ander Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. 2023. Nlp evaluation in trouble: On the need to measure llm data contamination for each benchmark. *arXiv preprint arXiv:2310.18018*.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, et al. 2022. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

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*.

Albert Webson and Ellie Pavlick. 2021. Do prompt-based models really understand the meaning of their prompts? *arXiv preprint arXiv:2109.01247*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023. Wizardlm: Empowering large language models to follow complex instructions.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Jifan Yu, Xiaozhi Wang, Shangqing Tu, Shulin Cao, Daniel Zhang-Li, Xin Lv, Hao Peng, Zijun Yao, Xiaohan Zhang, Hanming Li, et al. 2023. Kola: Carefully benchmarking world knowledge of large language models. *arXiv preprint arXiv:2306.09296*.

Baoli Zhang, Haining Xie, Pengfan Du, Junhao Chen, Pengfei Cao, Yubo Chen, Shengping Liu, Kang Liu, and Jun Zhao. 2023. Zhujiu: A multi-dimensional, multi-faceted chinese benchmark for large language models.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric P. Xing, Joseph E. Gonzalez, Ion Stoica, and Hao Zhang. 2024. Lmsys-chat-1m: A large-scale real-world llm conversation dataset.

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

Kun Zhou, Yutao Zhu, Zhipeng Chen, Wentong Chen, Wayne Xin Zhao, Xu Chen, Yankai Lin, Ji-Rong Wen, and Jiawei Han. 2023a. Don't make your llm an evaluation benchmark cheater. *arXiv preprint arXiv:2311.01964*.

Kun Zhou, Yutao Zhu, Zhipeng Chen, Wentong Chen, Wayne Xin Zhao, Xu Chen, Yankai Lin, Ji-Rong Wen, and Jiawei Han. 2023b. Don't make your llm an evaluation benchmark cheater.

A Data construction

A.1 Prompt construction

We use GPT-4, GPT3.5, LLaMa70B, and Claude instance to build prompts. The specific instructions are as follows:

Now I would like you to provide me with some templates, and I will give you some examples to illustrate my requirements: OWant represents, as a result, Y or others want For relationships<Event, oWant, Result>, construct the following question: What do PersonalY or others want to do next? <Event, oWant, Result>can be: PersonX admissions PersonY's work, oWant, to be acknowledged PersonX admissions PersonY was wrong, oWant, to apologize PersonX options a child, oWant, to decorate room Please provide me with more questions

Table 2: An example of building multiple prompts

A.2 ConceptNet Question Construction

Table 3 displays a prompt for formulating an LLM query derived from a ConceptNet triple.

Triplet	<hammer , AtLocation , tool box>
Prompt	Where is ______ typically located?
Question	Where is hammer typically located?
Answer	Tool box

Table 3: ConceptNet data construction process

Metrics / Knowledge / LLMs	Conceptnet		
	A_{CQ}	A_{QA}	H
Yi-6B-Chat	43.47	71.29	0.40
TChatGLM3-6b	38.17	73.59	0.63
WizardLM-13B-V1.2	39.62	74.18	0.48
Baichuan-13B-Chat	35.43	74.23	0.54
Baichuan2-13B-Chat	38.21	73.7	0.65
Baichuan2-7B-Chat	39.26	74.85	0.61
Vicuna-13b-v1.5	43.18	73.05	0.50
Vicuna-7b-v1.5	42.13	72.59	0.53
LLaMa2-13b-chat	37.94	73.75	0.49
LLaMa2-7b-chat	39.8	74.69	0.55
Mistral-7B-Instruct-v0.1	42.58	72.41	0.47
Mistral-7B-Instruct-v0.2	44.12	73.28	0.57
Qwen-14B-Chat	43.56	72.33	0.57
Qwen-7B-Chat	38.62	72.05	0.56

Table 4: The specific performance of the LLMs in ConceptNet. A_{CQ}: multiple choice question accuracy, A_{QA}: QA question accuracy, H: entropy.

A.3 ATOMIC Question Construction

During the process of constructing questions, we attempted to replace person variables (e.g., PersonX, PersonY) from ATOMIC with common human names. However, this led the model to treat them as real individuals and refuse to answer. Therefore, we directly use person variables in the questions to refer to human beings, thereby obtaining more objective answers.

Triplet	<PersonX accepts thanks , xReact , happy>
Prompt	______. How does PersonX feel?
Question	PersonX accepts thanks. How does PersonX feel?
Answer	happy

Table 5: ConceptNet data construction process

B Model self-preference value

To measure whether the model exhibits a preference for its own generated results, we use $rank(i, j)$ to represent the ranking assigned by evaluation model i to evaluation model j. We calculated the P as the self-preference value and $G(P)$ as the global self-preference value of models for their generated results in each knowledge domain and present the top five results in Table 7.

$$P(i) = \frac{1}{k-1} \sum_{\substack{j=1 \\ j \neq i}}^{k} (rank(j,i) - rank(i,i)) \tag{4}$$

$$G(P) = \frac{1}{k} \sum_{i=1}^{k} (P(i)) \tag{5}$$

Metrics / Knowledge / LLMs	Conceptnet		
	A_{CQ}	A_{QA}	H
Yi-6B-Chat	51.68	48.96	0.40
Chatglm3-6b	45.52	51.66	0.44
WizardLM-13B-V1.2	36.69	52.18	0.33
Baichuan-13B-Chat	42.95	51.92	0.37
Baichuan2-13B-Chat	47.11	49.71	0.45
Baichuan2-7B-Chat	44.30	41.46	0.43
Vicuna-13b-v1.5	47.19	46.88	0.35
Vicuna-7b-v1.5	46.71	49.97	0.37
LLaMa2-13b-chat	39.69	50.4	0.34
LLaMa2-7b-chat	36.96	48.79	0.38
Mistral-7B-Instruct-v0.1	47.19	47.44	0.33
Mistral-7B-Instruct-v0.2	47.71	51.23	0.40
Qwen-14B-Chat	55.77	49.79	0.39
Qwen-7B-Chat	50.33	47.39	0.39

Table 6: The specific performance of the LLMs in ATOMIC. A_{CQ}: multiple choice question accuracy, A_{QA}: QA question accuracy, H: entropy.

C The Performance of LLMs in World Knowledge and Language Knowledge

We conducted a detailed analysis of the performance of LLMs on world knowledge. As illustrated in Figure 2, the models exhibit significantly better performance on high-frequency knowledge compared to low-frequency knowledge, indicating a certain challenge in mastering low-frequency world knowledge. Additionally, the models exhibit low sensitivity to time-sensitive knowledge, implying a potential confusion of temporal occurrence times.

We also conducted a detailed analysis of the performance of LLMs in the scope of language. As illustrated in Figure 4, compared to other tasks, these models exhibit significantly better performance in pos-analysis tasks, indicating that the models are well-versed in more traditional natural language processing tasks like POS analysis. However, they exhibit lower proficiency in reverse tasks and tasks requiring deeper understanding and application of the knowledge corpus.

Model Name	Total Score	common-frequency	high-frequency	low-frequency	timeliness
01ai/Yi-34B-Chat	**79.85**	47.75	45.5	27.45	37.5
lmsys/vicuna-13b-v1.5	**78.06**	45.5	42.4	37.3	24.45
qwen/Qwen-72B-Chat	**75.3**	42.95	36.6	27.25	49.3
mistralai/Mistral-7B-Instruct-v0.2	**74.38**	44.05	39.6	39.55	20.25
WizardLM/WizardLM-13B-V1.2	**67.13**	44.85	43.7	31.2	16.3
baichuan-inc/Baichuan-13B-Chat	**64.7**	37.8	34.05	33.85	30.65
lmsys/vicuna-7b-v1.5	**59.25**	40.3	34.5	33.4	18.4
meta/llama2-70b-chat	**58.75**	43	37.2	16.75	39.7
qwen/Qwen-14B-Chat	**51.54**	37.65	31.85	31.1	17.05
01ai/Yi-6B-Chat	**50.62**	33.2	32	18.15	44.4
baichuan-inc/Baichuan2-7B-Chat	**50.53**	33.35	32.35	31.3	20.2
baichuan-inc/Baichuan2-13B-Chat	**47.84**	37.15	34.35	23.3	22
qwen/Qwen-7B-Chat	**46.67**	33.9	29.85	31	17.1
meta/llama2-13b-chat	**37.75**	31.85	31.95	24.4	13.95
mistralai/Mistral-7B-Instruct-v0.1	**36.71**	30.15	28.9	25.7	16.5
THUDM/chatglm3-6b	**19.21**	22.7	23.5	20.4	14.7
meta/llama2-7b-chat	**0**	17.5	15.85	16.3	8

Figure 2: The performance of LLMs on common-frequency, high-frequency, low-frequency, and timeliness in world knowledge.

Model Name	Total Score	commonsense_ConceptNet
baichuan-inc/Baichuan2-7B-Chat	**100**	74.85
meta/llama2-7b-chat	**95.51**	74.69
baichuan-inc/Baichuan-13B-Chat	**82.58**	74.23
WizardLM/WizardLM-13B-V1.2	**81.18**	74.18
meta/llama2-13b-chat	**69.1**	73.75
THUDM/chatglm3-6b	**64.61**	73.59
baichuan-inc/Baichuan2-13B-Chat	**61.24**	73.47
mistralai/Mistral-7B-Instruct-v0.2	**55.9**	73.28
lmsys/vicuna-13b-v1.5	**49.44**	73.05
lmsys/vicuna-7b-v1.5	**36.52**	72.59
mistralai/Mistral-7B-Instruct-v0.1	**31.46**	72.41
qwen/Qwen-14B-Chat	**29.21**	72.33
qwen/Qwen-7B-Chat	**21.35**	72.05
01ai/Yi-6B-Chat	**0**	71.29

Figure 3: The performance of LLMs on ConceptNet in commonsense knowledge.

Model Name	Total Score	idiom_explanation	pos_analysis	semantic_selection	similar_words
01ai/Yi-34B-Chat	**96.92**	95.6	98.88	93.98	94.29
openai/chatgpt	**95.48**	85.71	100	97.14	93.75
WizardLM/WizardLM-13B-V1.2	**94.5**	85.19	98.86	97.48	93.94
qwen/Qwen-72B-Chat	**94.05**	86.41	98.8	94.07	94.63
lmsys/vicuna-13b-v1.5	**93.98**	79.2	99.4	97.48	95.36
01ai/Yi-6B-Chat	**92.45**	92.38	98.21	94.08	90.2
qwen/Qwen-14B-Chat	**91.61**	85.96	99.38	89.44	93.02
lmsys/vicuna-7b-v1.5	**89.67**	74.78	100	92.02	93.38
mistralai/Mistral-7B-Instruct-v0.2	**89.57**	72.45	98.87	95.42	94.05
qwen/Qwen-7B-Chat	**89.05**	85.05	98.8	92.31	88.95
meta/llama2-13b-chat	**89.02**	81.9	97.31	90.59	93.41
meta/Llama-2-70b-chat	**81.21**	67.27	96.17	83.67	94.44
baichuan-inc/Baichuan-13B-Chat	**80.61**	79.57	96.92	83.78	86.5
THUDM/chatglm3-6b	**71.07**	77.52	88.58	73.61	90.48
baichuan-inc/Baichuan2-13B-Chat	**69.11**	71.55	92.9	63.76	90
mistralai/Mistral-7B-Instruct-v0.1	**61.67**	20.51	96.79	73.78	93.18
tiiuae/falcon-40b-instruct	**60.64**	53.33	92.53	63.95	86.89
meta/llama2-7b-chat	**54.16**	65.57	92.12	68.18	69.74
baichuan-inc/Baichuan2-7B-Chat	**15.17**	66.06	73.21	30.49	58.87

Figure 4: The four knowledge corpus-based tasks performance of LLMs in language knowledge.

Figure 5: The introduction page of Zhujiu-knowledge platform.

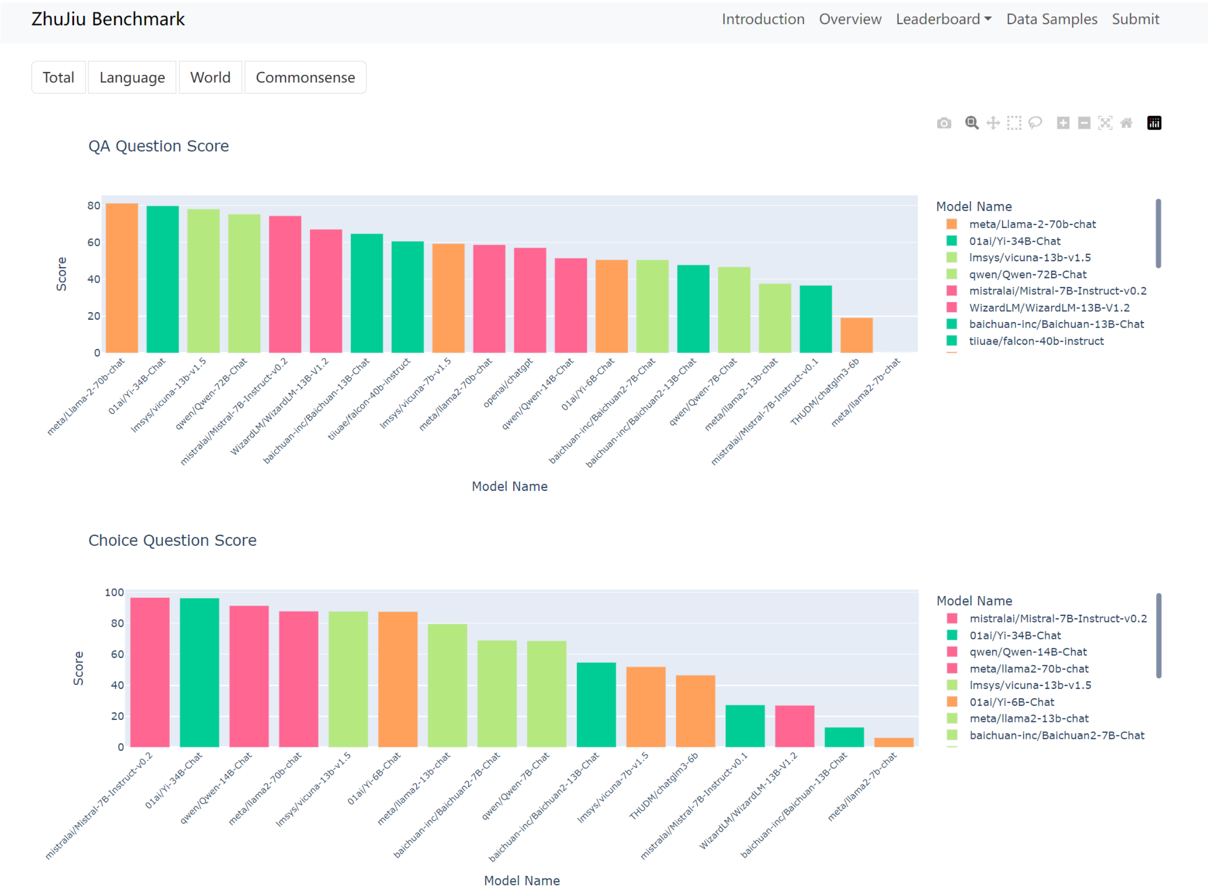

Figure 6: Overview Performance: The total score of LLMs in QA question and Choice question.

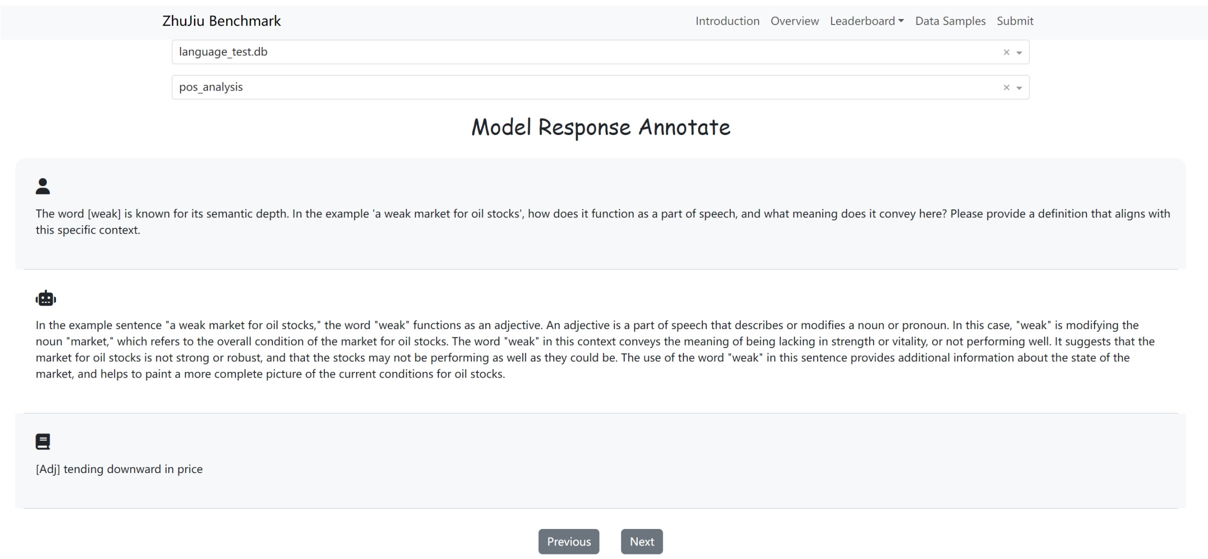

Figure 7: The dataset demonstration page.

evalmodel	commonsense	world	lanauage
top-model1	-3.2	-8.6	6.6
top-model2	0.2	4.4	4.4
top-model3	4.6	0.6	-6.2
top-model4	0.4	-0.75	0.4
top-model5	-1.4	6.8	-1.8
global-self-preference	0.1	0.5	0.6

Table 7: The self-preference value of the top five LLMs in three knowledge evaluations