Performance Evaluation of General-Purpose Chinese Large Language Models in Humanities and Social Sciences

Authors: Zhao Zhixiao, butterfly, Liu Chang, contemplation, Wang Dongbo, Wang Dongbo

Date: 2024-05-08T00:00:00+00:00

Abstract

This paper takes the humanities and social sciences domain as its starting point, conducting comparative model performance evaluations from two aspects: foundational knowledge in humanities and social sciences and academic texts within the field. It aims to provide a systematic evaluation benchmark for large models in the humanities and social sciences domain, serving as a reference for researchers in related fields.

We designed seven evaluation tasks related to the humanities and social sciences domain and selected corresponding metrics. On this basis, we selected currently open-source Chinese large language models with superior performance in the general domain. Domain-specific tasks were completed in a question-answering format by invoking local models, and relevant metrics were selected to conduct quantitative evaluations of their performance in the humanities and social sciences domain.

The evaluation results indicate that among the open-source models selected in this paper, Qwen demonstrates optimal performance in both base and dialogue models, followed closely by Baichuan2, then InternLM, while Atom exhibits the poorest performance. Furthermore, dialogue models demonstrate superior performance compared to base models in most cases.

Full Text

Performance Evaluation of Chinese Universal Large Language Models in the Field of Humanities and Social Sciences

Zhao Zhixiao[12], Hu Die[12], Liu Chang[12], Shen Si[3], Wang Dongbo[12]
[1]College of Information Management, Nanjing Agricultural University, Nanjing 210095
[2]Research Center of Humanities and Social Computing, Nanjing Agricultural University, Nanjing
[3]School of Economics and Management, Nanjing University of Science and Technology, Nanjing

This work is supported by the National Social Science Foundation of China project "Research on the Construction and Application of Cross-language Knowledge Base for Ancient Chinese Classics" (Grant No. 21&ZD331).

Authors: Zhao Zhixiao, Master student; Hu Die, Master student; Liu Chang, Doctoral student; Shen Si, Associate Professor, PhD, Doctoral supervisor; Wang Dongbo, Professor, PhD, Doctoral supervisor, Corresponding author, E-mail: db.wang@njau.edu.cn.

Abstract

[**Purpose/Significance**] This paper evaluates large language model performance in the humanities and social sciences domain from two perspectives: foundational domain knowledge and academic texts. The aim is to provide a systematic evaluation benchmark for large language models specifically tailored to the humanities and social sciences, serving as a reference for researchers in related fields. [**Methods/Processes**] We designed seven evaluation tasks relevant to humanities and social sciences with corresponding metrics. Building upon this, we selected currently open-source, high-performance general-domain Chinese large language models and deployed them locally to complete domain-specific tasks in a question-answering format, quantitatively assessing their performance in humanities and social sciences through selected indicators. [**Results/Conclusions**] Evaluation results demonstrate that among the selected open-source models, Qwen achieves the best performance in both base and chat variants, followed by Baichuan2, with InternLM ranking third and Atom performing the worst. Moreover, in most cases, chat models exhibit superior performance compared to their base counterparts.

Keywords: Humanities and Social Sciences; Large Model Evaluation; Domain Knowledge; Academic Texts

---

1 Introduction

With an increasing number of internet companies and research teams joining the AIGC wave, numerous open-source and commercially viable large language models have been released. Society at large has recognized the opportunities brought by this AI revolution and actively participated in large model research. After half a century of development, humanities and social sciences have achieved deep integration with computer science, giving rise to interdisciplinary fields such as computational humanities and computational social sciences that demonstrate broad development prospects. As an intersection of computer science and humanities/social sciences, this integration aims to apply computer science theories and methods to various humanities and social science domains, enriching research content in these fields [1,2]. With continuous advances in AI technology, the convergence of humanities/social sciences and computer science has entered a new stage. The combination of domain-specific data and theories from humanities and social sciences with AI technology will inevitably drive collaborative development across both domains.

Due to the substantial hardware and data requirements for large models, most research teams currently find it difficult to train general-domain large models from scratch. However, as large model research progresses, training pipelines have gradually converged, and the release of open-source large models enables more research teams to build vertical domain models by combining them with specific domain data. Vertical domain model construction typically involves incremental training of existing models using domain data, enabling them to maintain strong language capabilities while acquiring more vertical domain expertise. Consequently, the selection of base models significantly impacts final model performance. The development of large models has also triggered a wave of evaluation efforts, including evaluations targeting vertical domains, though comprehensive evaluations for humanities and social sciences remain scarce. For most humanities and social sciences scholars, selecting appropriate large language models from numerous open-source general models for their research domains presents considerable difficulty. Whether for applying large language models in humanities and social sciences or for humanities scholars adopting new technologies in the new era, a comprehensive evaluation system is essential. Therefore, this paper conducts an evaluation of Chinese open-source large models specifically for the humanities and social sciences domain—a broad yet distinctive field—aiming to provide a benchmark for evaluating large models that can serve as a reference for relevant domain researchers. For humanities scholars without computer science backgrounds, this paper offers quantitative references for understanding and using large language models, while for computational humanities researchers, the evaluation results can guide the selection of suitable large language models for humanities and social sciences research.

2.1 Interdisciplinary Integration of Humanities/Social Sciences and Artificial Intelligence

Digital humanities has emerged as a new discipline, sparking research enthusiasm in information resource management. Huang Shuiqing et al. [3] reviewed the current disciplinary development of computational humanities, discussing the development status of "Computational X" disciplines through examples such as computational literature, computational linguistics, and computational history. Computer technology has integrated with numerous humanities disciplines, and the data-driven research paradigm has been widely applied in humanities, generating a series of empirical studies grounded in humanities disciplines. Wang Dongbo et al. [4] constructed a domain-specific pre-training model for ancient Chinese text processing based on the Siku Quanshu (Complete Library in Four Sections) data and BERT models, demonstrating superior performance in various ancient text processing tasks. Zhang Wei et al. [5] integrated ancient poetry appreciation with AI technology, combining poetry text vectors with pre-trained models to achieve automatic and efficient extraction of sentiment terms from ancient poetry. Zhang Qi et al. [6] took the *Records of the Grand Historian* as a case study, reorganizing and reformatting complex information in historical texts to construct a multi-dimensional knowledge base and visualization platform, reducing barriers for users reading historical texts. Yu Xuehan et al. [7] integrated neural networks with machine reading comprehension patterns, conducting training and validation on both annalistic and biographical historical corpora. In the digital humanities domain, intelligent processing of ancient texts and mining of traditional Chinese culture have become highly popular research topics, with AI technology driving the revitalization and utilization of ancient texts serving as a model for humanities-AI integration. Beyond this, the intersection of social sciences such as finance and law with AI has become increasingly close. Zhang Ruixiang et al. [8] discussed the evolution and current status of computational jurisprudence research paradigms by examining the disciplinary development path of computational law combined with AI applications in the legal field. Liang Zhu et al. [9] constructed a judgment document recommendation system using news-like factual texts combined with structural content features of legal judgment documents.

Information technology iterations continuously inject new vitality into humanities and social sciences research, driving progress and development. Currently, large language model technology has enabled AI to achieve a qualitative leap. In the era of large models, researchers have already explored the applicability of general large language models in humanities and social sciences. To better adapt to specialized data and personalized task requirements in humanities and social sciences, industry and academia have launched more large language models targeting vertical domains in humanities and social sciences. For instance, in the financial domain, W. Shijie et al. [10] constructed BloombergGPT, a 50B parameter financial domain large model based on massive financial data, filling the gap in financial domain large models. Y. Hongyang et al. [11] proposed

a new open-source framework for financial large models, enhancing domain capabilities through open-source data fine-tuning while reducing training costs, further advancing financial AI development. Subsequent models such as PIXIU [12] and InvestLM [13] represent important attempts at large language models in finance. In the legal domain, LawGPT_{zh} [14] used ChatGPT to clean open-source legal datasets to construct an open-source large model suitable for the legal field. ChatLaw [15] constructed models using extensive legal news, legal forums, statutes, law exam questions, and judgment documents as conversational data based on base models like Ziya-13B. Additionally, domains such as education [16,17], healthcare [18,19], and e-commerce [20,21] have seen corresponding large model achievements, providing reliable support for professional scenario applications.

Overall, AI technology has been applied across various domains of humanities and social sciences. In the large model era, constructing vertical domain models has become increasingly important, with domain dataset construction forming the foundation for in-depth large model research in humanities and social sciences. Building domain-specific datasets inevitably requires participation from domain researchers, while data organization and storage processes constitute important research content in information resource management. Large model training and fine-tuning involve computer science theories and technologies, making large model construction inherently dependent on cross-disciplinary and cross-domain cooperation and communication.

2.2 Related Research on Large Model Evaluation

Current large model evaluations can be categorized into several dimensions: language capability, knowledge reserves, safety, and other aspects. Large language models are most renowned for their powerful language capabilities, enabling them to handle most natural language processing tasks with few-shot or even zero-shot learning. Knowledge reserves stem from the massive data learned during pre-training, allowing models to answer factual questions in domains such as finance, law, and medicine, as well as common-sense questions. Regarding safety, the Interim Measures for the Management of Generative AI Services released in July 2023 explicitly stipulate that providing and using generative AI should comply with laws, regulations, and respect social ethics and moral standards [22]. Other aspects include real-life applications of large models and their use as intelligent agent tools.

Model language capability evaluation can be divided into natural language understanding and natural language generation tasks. Natural language understanding primarily includes sentiment analysis and text classification, while natural language generation mainly encompasses dialogue, summarization, and translation. In large model language understanding evaluation, W. Zengzhi et al. [23] assessed ChatGPT' s ability to understand opinions, emotions, and sen-

timents in text, comparing it with fine-tuned BERT models and state-of-the-art methods to determine ChatGPT' s suitability for sentiment analysis tasks. Z. Wenxuan et al. [24] investigated large model performance across various sentiment analysis tasks, comparing them with small models trained on specific datasets. Results showed that large models perform well on simple tasks but poorly on more complex ones, though they still outperform small models with just a few additional examples. P. Alejandro et al. [25] evaluated large model performance in public affairs document classification, constructing a text classification dataset with 30 categories and building binary classification evaluation data for each category to address sample imbalance issues. In large model language generation evaluation, Z. Wenhao et al. [26] studied large language model translation capabilities, including evaluation, capability stimulation, and performance across different languages, exploring pathways for improving translation capabilities and the impact of training corpus languages.

Regarding knowledge reserves, evaluations include both domain-specific knowledge and general world knowledge. For example, D. Xuanquy et al. [27] analyzed ChatGPT' s performance on high school mathematics multiple-choice questions, finding that ChatGPT struggled with derivatives and spatial geometry but excelled in exponential and logarithmic problems. W. Yiran et al. [28] explored GPT-4' s ability to solve high-difficulty mathematics problems, notably employing multiple approaches to test GPT-4, including MathChat. D. Dat et al. [29] assessed large model performance in genetics, finding ChatGPT' s performance comparable to humans, particularly excelling in memorization-based questions. G. Aidan et al. [30] evaluated ChatGPT' s performance on medical licensing examinations, using two sets of multiple-choice questions to assess ChatGPT and comparing it with GPT-3 and InstructGPT.

In terms of safety, evaluations primarily focus on moral bias and robustness. Z. Jiaxu et al. [31] constructed the CHBias dataset for evaluating Chinese conversational large models, demonstrating that some models still exhibit social bias tendencies. P. Alicia et al. [32] introduced the BBQ benchmark for question-answering bias, covering nine prevalent social biases in the American social context. Evaluation results revealed that providing context can somewhat reduce model stereotypes but cannot eliminate them completely.

Additionally, numerous cross-domain and cross-task large model evaluation benchmarks have been proposed, such as M3Exam [33] constructed by Z. Wenxuan et al. using nine languages from real exam questions, and a multi-level, multi-source multiple-choice evaluation dataset built by H. Yuzhen et al. [34] covering 52 disciplines across four difficulty levels. X. Liang et al. [35] evaluated large language models based on user ratings from large model release platforms, discussing the limitations of evaluating large language models using closed-ended questions.

In summary, many current large language model evaluations are based on ChatGPT, primarily because ChatGPT has become the benchmark for large models, and its strong instruction-following capability facilitates evaluation. Some stud-

ies have also used ChatGPT to evaluate other large models' outputs, which proves feasible for smaller models. On the other hand, many evaluation benchmarks employ numerous multiple-choice questions, which facilitates quantitative evaluation results. However, closed-ended tasks may not fully demonstrate model performance. In contrast, open-ended tasks face challenges in achieving accuracy, objectivity, and convenience simultaneously in quantitative evaluation.

---

3 Evaluation System Design

In recently introduced large model evaluation tasks, multiple-choice questions have become a significant component. The rationale is that evaluating models through multiple-choice questions enables relatively objective and rapid acquisition of scores in specific aspects. However, multiple-choice questions primarily assess domain knowledge reserves and instruction-following capabilities, somewhat neglecting text generation abilities. To comprehensively evaluate large model performance in humanities and social sciences, this paper approaches from two angles—domain knowledge and academic texts—selecting 13 high-performance open-source general-domain Chinese large models and designing seven tasks for evaluation. The overall framework is illustrated in Figure 1 [Figure 1: see original paper].

3.1 Model Selection

During the initial model selection phase, we surveyed well-recognized large model evaluation leaderboards, including OpenLLM [36], SuperCLUE [35], C-eval [34], and CLiB [37], selecting high-performance open-source Chinese large models as candidates. Survey results revealed that parameter scale significantly impacts model performance, with most current open-source Chinese models concentrated around 7B parameters. Therefore, this paper uniformly selected large models at the billion-parameter scale for evaluation. Specific model information is presented in Table 1 . We selected 13 models, including six base models and seven chat models. For some models, such as ChatGLM-6B, only chat versions are currently open-sourced, so only the chat variant was included. Compared to base models, chat models, fine-tuned on general conversational data, possess stronger instruction-following capabilities and can better understand user questions and respond appropriately. Based on most general-domain model evaluation results, chat models typically exhibit superior performance in dialogue-form tasks. Therefore, this paper hypothesizes that chat models will also demonstrate superior performance in vertical domains, comparing chat and base models separately to clarify performance differences. Regarding evaluation task formats, since base models have weaker instruction-following capabilities, we emphasized post-processing of base model outputs for tasks with strict formatting requirements, such as multiple-choice and classification tasks.

Table 1 Chinese Open Source Large Models | Model Name | Organization | |———|———-| | FlagAlpha Atom-7B [38] | FlagAlpha | | Baichuan-7B [39] | Baichuan Inc. | | Baichuan2-7B [40] | Baichuan Inc. | | Chinese-Alpaca-7B [41] | ymcui | | InternLM-7B [42] | Shanghai AI Laboratory | | Qwen-7B [43] | Alibaba Cloud | | Atom-7B-Chat [38] | FlagAlpha | | Baichuan2-7B-Chat [40] | Baichuan Inc. | | ChatGLM-6B [44] | Zhipu AI | | ChatGLM2-6B [45] | Zhipu AI | | InternLM-7B-Chat [42] | Shanghai AI Laboratory | | Phoenix-Inst-Chat-7B [46] | Chinese University of Hong Kong (Shenzhen) | | Qwen-7B-Chat [43] | Alibaba Cloud |

3.2 Task Construction

For humanities and social sciences domain knowledge, we constructed three evaluation tasks: multiple-choice questions, terminology explanation, and open-ended Q&A. Data for multiple-choice and open-ended Q&A tasks were sourced from humanities and social sciences knowledge competition questions, including university student humanities knowledge competitions and domain-specific competitions in philosophy, law, and economics from the past three years. To ensure broader and more balanced subject coverage, we sampled data at a 4:2:2:2 ratio from these sources. Multiple-choice questions primarily cover basic knowledge in philosophy, history, Chinese cultural conventions, law, and economics. Open-ended Q&A focuses on humanities and social sciences discussion questions, including explanations of specialized knowledge and enumerations of domain-specific world knowledge, assessing large language models' mastery of domain knowledge and ability to answer in structured points—for example, "In what aspects does the applied trend of contemporary humanities manifest?"Terminology explanation data were sourced from research outcomes of the Interdisciplinary Research Base for Terminology and Translation at Nanjing University' s School of Foreign Studies [47], containing terminology texts and explanations across ten disciplines including management, education, economics, and archaeology. We extracted 20 comprehensive and complete explanations from each discipline, totaling 200 data points as the terminology explanation test corpus.

For humanities and social sciences academic texts, we constructed four evaluation tasks: disciplinary classification of paper abstracts, rhetorical move recognition in abstracts, title generation, and academic text translation. Paper abstracts, titles, and their corresponding disciplines and move annotations were sourced from CSSCI-indexed journal paper abstracts and their subject classifications. Considering that CSSCI currently classifies papers into 26 disciplines, which would be too many for models to handle effectively, we sampled 20 data points per discipline per task from 10 disciplines, totaling 200 data points each for abstract classification, move recognition, and title generation tasks. Text translation data were obtained and processed similarly to terminology explanation data. Given that current Chinese large models possess bilingual capabilities, the text translation task evaluates models more comprehensively through Chinese-English and English-Chinese translation.

After collecting the required evaluation data, we constructed prompt instructions for different tasks, combining relevant tasks with expected outputs. For multiple-choice questions, to reduce subsequent evaluation difficulty, we needed to ensure models output only the correct option. Simply inputting questions and options would often result in irrelevant content, complicating evaluation. Additionally, assigning roles to models during prompting can improve output quality [48]. Considering the domain-specific nature of this study, we incorporated relevant role-setting language in prompt construction. We then tested small data samples across models to maximize the number of models producing required outputs, constructing instructions as shown in Table 2 .

Table 2 Examples of Prompts | Task | Prompt Template | |—|————-| | Open Q&A | As a humanities and social sciences researcher, answer the following open-ended question in humanities and social sciences as comprehensively and systematically as possible: {input} | | Multiple-choice | As a humanities and social sciences researcher, answer the following multiple-choice question by outputting only the option (A, B, C, or D) without explanation or additional content: {input} | | Terminology Explanation | As a humanities and social sciences researcher, explain the following domain terminology in as much detail as possible: {input} | | Title Generation | Based on the following humanities and social sciences abstract text, provide the most likely title, outputting only the title without additional content: {input} | | Text Translation | Translate the following {cultural studies} domain English (Chinese) text into Chinese (English), outputting only the translation result: {input} | | Disciplinary Classification | Based on the following humanities and social sciences abstract text, determine which discipline this abstract belongs to from ['Philosophy' , 'History' , 'Law' , 'Political Science' , 'Economics' , 'Sociology' , 'Education' , 'Psychology' , 'Management' , 'Linguistics'], outputting only the category name without additional content: {input} | | Move Recognition | Based on the following humanities and social sciences abstract text, determine which part of the abstract this text belongs to from ['Results' , 'Methods' , 'Purpose' , 'Limitations' , 'Conclusions'], outputting only the category name without additional content: {input} |

After constructing basic instructions, we determined the model prompting mode. The powerful text generation capabilities of large models partly stem from their strong in-context learning abilities. Therefore, to maximize output quality for certain tasks, we can provide a few examples to help models better understand instructions and produce more desirable content, i.e., few-shot learning [49]. However, not all tasks and models are suitable for few-shot mode. Based on the constructed instructions, we tested 0-shot, 1-shot, and 3-shot modes across different tasks and models. We minimized the number of examples while ensuring models could produce ideal outputs, finalizing the example counts for each task as shown in Table 3 .

Table 3 Mode of Prompts | Task | Prompt Mode | |—|———-| | Multiple-choice | 0-shot | | Terminology Explanation | 0-shot | | Open Q&A | 0-shot

| | Disciplinary Classification | 0-shot | | Move Recognition | 0-shot | | Title Generation | 3-shot | | Text Translation | 1-shot |

3.3 Evaluation Metrics

The development of generative models has posed significant challenges for model evaluation. Building upon current natural language processing evaluation metrics, this paper selected corresponding metrics for each specific task to enable objective and quantitative evaluation of model-generated content as much as possible. The evaluation metrics and calculation methods adopted are described below.

(1) Accuracy
Accuracy is simple and convenient to calculate, suitable for multiple-choice and classification tasks. However, it requires high standardization of model outputs. During our evaluation, some models scored lower due to non-standard output content. For multiple-choice and classification tasks, after obtaining model outputs, we manually corrected results, modifying cases where answers were correct but contained extraneous content.

(2) Precision, Recall, and F1-score
Precision, recall, and F1-score are commonly used metrics in natural language processing tasks, calculated based on confusion matrices. For each category, all predictions can be divided into four types: True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). The corresponding formulas are:

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1\text{-}Score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

Considering the non-standardization issues of generative model outputs, we employed weighted averaging for multi-class task metrics to obtain final scores. Weighted averaging uses the proportion of samples in each category relative to the total sample size as weights to calculate average metrics, effectively mitigating the impact of non-standard model outputs. The weighted average formula is as follows, where x_i represents the proportion of samples predicted as category i relative to the entire dataset, and f_i represents the F1-score for category i:

$$Wx = \sum x_i \times f_i$$

(3) BLEU

BLEU [50] is a series of metrics based on n-gram precision for measuring similarity between generated and reference texts. We adopted BLEU as the evaluation metric for text translation. The specific formula is:

$$BLEU\text{-}n = \frac{\sum_{S\in\{Candidate\}}\sum_{gram_n\in S} Count_{match}(gram_n)}{\sum_{S\in\{Candidate\}}\sum_{gram_n\in S} Count(gram_n)}$$

where n represents n-gram length, and $Count_{match}$ represents the maximum number of n-grams co-occurring in both generated and reference texts. The denominator represents the total number of n-grams in the generated text.

(4) ROUGE

ROUGE [51] is a series of metrics based on n-gram recall for measuring similarity between generated and reference texts. We adopted ROUGE-N and ROUGE-L as evaluation metrics for title generation, terminology explanation, and open Q&A. The ROUGE-N formula is:

$$ROUGE\text{-}N = \frac{\sum_{S\in\{Reference\}}\sum_{gram_n\in S} Count_{match}(gram_n)}{\sum_{S\in\{Reference\}}\sum_{gram_n\in S} Count(gram_n)}$$

ROUGE-L is based on Longest Common Subsequence (LCS) concepts, calculating P, R, and F values for the maximum common subsequence between two sentences:

$$R_{lcs} = \frac{LCS(X,Y)}{m}$$

$$P_{lcs} = \frac{LCS(X,Y)}{n}$$

$$F_{lcs} = \frac{(1+\beta^2)R_{lcs}\times P_{lcs}}{R_{lcs}+\beta^2 P_{lcs}}$$

where m and n represent the lengths of reference and generated texts, respectively. In calculations, β is set to a large value. Like ROUGE-N, ROUGE-L primarily references recall.

(5) chrF

chrF [52] is similar to BLEU but operates at the character level, primarily calculating F-score. The formula is:

$$chrF_\beta = (1+\beta^2)\frac{chrP\times chrR}{\beta^2\times chrR + chrP}$$

where *chrP* is the percentage of characters in the generated text that belong to the reference text, and *chrR* is the percentage of characters in the reference text that belong to the generated text. chrF is also applied to text translation evaluation. To ensure accurate BLEU and chrF calculations, we used methods provided by sacreBLEU [53].

(6) MAUVE

MAUVE [54] assesses the similarity between AI-generated and human-generated text from a KL divergence perspective. This metric relies on autoregressive language models. According to relevant experiments, MAUVE's correlation with human evaluation increases with model parameter size. In this paper, we used the GPT2-large [55] model as the autoregressive language model for MAUVE calculation, applying it to evaluate open Q&A and terminology explanation.

In specific calculations, the open-source evaluate [56] library from HuggingFace provides convenient pathways and methods for metric calculation. After preparing models and relevant metric calculation code, metrics can be loaded and computed through the evaluate library. For final comparison convenience, we unified all metric scores to a 100-point scale to obtain final task-specific scores.

---

4 Evaluation Results and Analysis

This evaluation employed PyTorch-2.0.1 as the deep learning framework, with model inference based on transformers-4.30.1. For hardware, we used a single NVIDIA RTX A6000 48GB GPU for model inference, with NVIDIA driver version 535.146.02 and CUDA version 12.2. During experiments, considering that reference text lengths for all tasks did not exceed 512 tokens, we set the maximum generation length to 512 to prevent excessively long outputs. Model output parameters are detailed in Table 4 ; unlisted parameters used default values.

Table 4 Parameters of Model Output | Parameter | Value | |——-|——-| | max_{length} | 512 | | min_{length} | - | | do_{sample} | False | | no_{{{repeat}}*{{ngram}}*}{size} | - | | top_k | - | | top_p | - | | temperature | - |

4.1.1 Multiple-Choice Questions

Multiple-choice questions are a common task in large model evaluation. Considering the weak instruction-following capability of base models, we post-processed base model outputs to improve evaluation accuracy, modifying cases where models produced non-standard but correct outputs to the correct option. In practice, most base models struggled to output options as required, with some simply repeating questions, while chat models generally produced more normal results. Although chat models occasionally included explanations, their performance far exceeded that of base models. Among all models, Baichuan2-7B-Chat and

InternLM-7B-Chat demonstrated superior output standardization, perfectly following instructions to output only options across all 100 test questions. Detailed performance of each model on the multiple-choice task is shown in Table 5 .

Table 5 Indicators of Single Choices | Model | Score (Accuracy) | |—-|————-| | Atom-7B | - | | Baichuan-7B | - | | Baichuan2-7B | - | | Chinese-Alpaca-7B | - | | InternLM-7B | - | | Qwen-7B | - | | Atom-7B-Chat | - | | Baichuan2-7B-Chat | - | | ChatGLM-6B | - | | ChatGLM2-6B | - | | InternLM-7B-Chat | - | | Phoenix-Inst-Chat-7B | - | | Qwen-7B-Chat | - |

Most chat models demonstrated superior performance, with Baichuan2-7B-Chat achieving the highest accuracy among chat models, while Qwen-7B achieved the highest accuracy among base models. Notably, Atom-7B-Chat performed worse than some base models on multiple-choice tasks. Upon examining model outputs, we found that Atom-7B-Chat frequently produced empty outputs in multiple-choice tasks, a problem that persisted despite repeated experiments, likely due to data quality issues during the instruction fine-tuning process for this chat model.

4.1.2 Terminology Explanation and Open Q&A

Terminology explanation and open Q&A evaluate large models' domain knowledge and text generation capabilities. Unlike multiple-choice questions, these tasks have no format requirements for outputs and require no post-processing, though evaluating open-ended generation tasks presents greater challenges. Detailed performance of each model on these tasks is shown in Table 6 .

Table 6 Indicators of Terminology Definition and Open Q&A | Model | Terminology Explanation (ROUGE-L/MAUVE) | Open Q&A (ROUGE-L/MAUVE) | |—-|————————-|————-| | Atom-7B | - | - | | Baichuan-7B | - | - | | Baichuan2-7B | - | - | | Chinese-Alpaca-7B | - | - | | InternLM-7B | - | - | | Qwen-7B | - | - | | Atom-7B-Chat | - | - | | Baichuan2-7B-Chat | - | - | | ChatGLM-6B | - | - | | ChatGLM2-6B | - | - | | InternLM-7B-Chat | - | - | | Phoenix-Inst-Chat-7B | - | - | | Qwen-7B-Chat | - | - |

Overall, open Q&A scores were significantly lower than terminology explanation scores. Moreover, the performance gap between base and chat models was smaller for these tasks. The conversational capabilities demonstrated by current generative models essentially stem from text continuation abilities. Since the question formats and instructions for terminology explanation and open Q&A tasks closely resemble daily conversations, and many models' pre-training data includes Q&A content, base models can also exhibit decent conversational and Q&A abilities. Open Q&A scores are lower because open questions use more varied vocabulary and have greater answer diversity, affecting score calculation when comparing against standard answers.

Among base models, Baichuan2-7B achieved the best scores, while among chat models, Baichuan2-7B-Chat and Qwen-7B-Chat achieved the best scores for

terminology explanation and open Q&A, respectively. In open Q&A tasks, Chinese-Alpaca-7B frequently produced empty outputs that persisted despite repeated experiments, while Atom-7B series models, though not producing empty outputs, generated non-linguistic content such as repeated numbers and punctuation. Considering practical applicability, we did not post-process these outputs before metric calculation, resulting in significantly lower scores for Chinese-Alpaca-7B and Atom-7B series models.

4.2.1 Academic Text Disciplinary and Move Classification

For classification tasks, even with example samples, base models struggled to output in correct formats. Similar to multiple-choice tasks, we post-processed base model outputs for classification tasks to better reflect actual model performance. After post-processing, detailed scores for each model on classification tasks are shown in Table 7 .

Table 7 Indicators of Categories Tasks | Model | Disciplinary Classification (Accuracy) | Move Classification (Accuracy) | |—-|————————-|————————-| | Atom-7B | - | - | | Baichuan-7B | - | - | | Baichuan2-7B | - | - | | Chinese-Alpaca-7B | - | - | | InternLM-7B | - | - | | Qwen-7B | - | - | | Atom-7B-Chat | - | - | | Baichuan2-7B-Chat | - | - | | ChatGLM-6B | - | - | | ChatGLM2-6B | - | - | | InternLM-7B-Chat | - | - | | Phoenix-Inst-Chat-7B | - | - | | Qwen-7B-Chat | - | - |

Overall, Baichuan-7B and Qwen-7B achieved the best scores among base models, while InternLM-7B-Chat achieved the best score among chat models. In most cases, move classification scores were lower than disciplinary classification scores. Observation revealed that most models only output 2-3 move categories, failing to cover all provided categories. Additionally, Baichuan2-7B and Qwen-7B base models outperformed their chat counterparts in disciplinary classification, partly due to post-processing of base model outputs and partly because chat models' performance degraded on non-conversational tasks like disciplinary classification after conversational fine-tuning.

4.2.2 Title Generation

Text summarization has long been a traditional task in natural language processing. For academic texts, generating summaries from full texts requires increasing maximum output length and resolving parsing issues with figures and formulas, presenting high difficulty. To evaluate large models' text organization capabilities, we employed title generation as an alternative to abstract generation, assessing models' text summarization abilities. Although models occasionally output irrelevant content, we did not perform further post-processing since our evaluation metric ROUGE emphasizes recall calculation. Detailed scores for each model on title generation are shown in Table 8 .

Table 8 Indicators of Title Generation | Model | ROUGE-1 | ROUGE-2 | ROUGE-L | |—-|——|——|——| | Atom-7B | - | - | - | | Baichuan-7B | - | - | -

| | Baichuan2-7B | - | - | - | | Chinese-Alpaca-7B | - | - | - | | InternLM-7B | - | - | - | | Qwen-7B | - | - | - | | Atom-7B-Chat | - | - | - | | Baichuan2-7B-Chat | - | - | - | | ChatGLM-6B | - | - | - | | ChatGLM2-6B | - | - | - | | InternLM-7B-Chat | - | - | - | | Phoenix-Inst-Chat-7B | - | - | - | | Qwen-7B-Chat | - | - | - |

Qwen-7B-Chat and Qwen-7B achieved the best scores among chat and base models, respectively, with Qwen-7B even slightly outperforming Qwen-7B-Chat. Combined with Qwen-7B' s performance on previous tasks, it is evident that Qwen models incorporated conversational data during pre-training, significantly enhancing base model performance and making their conversational abilities clearly superior to other base models.

4.2.3 Academic Text Translation

Machine translation is another traditional natural language processing task. With the continuous development of large language models, their superior performance in machine translation has shifted research paradigms, moving from Encoder-Decoder architectures to Decoder-only architectures. We evaluated large models' translation capabilities for humanities and social sciences academic texts using Chinese-English parallel corpora containing academic terminology. Detailed scores for each model on text translation are shown in Table 9 .

Table 9 Indicators of Translation | Model | Chinese-English (BLEU/chrF) | English-Chinese (BLEU/chrF) | |—-|————-|————-| | Atom-7B | - | - | | Baichuan-7B | - | - | | Baichuan2-7B | - | - | | Chinese-Alpaca-7B | - | - | | InternLM-7B | - | - | | Qwen-7B | - | - | | Atom-7B-Chat | - | - | | Baichuan2-7B-Chat | - | - | | ChatGLM-6B | - | - | | ChatGLM2-6B | - | - | | InternLM-7B-Chat | - | - | | Phoenix-Inst-Chat-7B | - | - | | Qwen-7B-Chat | - | - |

In text translation tasks, Qwen-7B and Baichuan2-7B-Chat achieved the best scores among base and chat models, respectively. Comparing scores across the two sub-tasks, Chinese-English translation scores were significantly higher than English-Chinese translation scores. Based on examination of evaluation data and model outputs, we attribute this to two factors. First, text translation metrics primarily calculate based on n-gram counts in generated texts, and English texts (word count) are shorter than Chinese texts (character count), resulting in higher scores for English outputs. Second, upon inspecting model outputs for English-Chinese translation tasks, we found some models output English text instead of Chinese translations, likely because they failed to understand instructions and simply continued the given English text.

Across all tasks, chat models generally demonstrated superior performance compared to base models for most models, though some exceptions existed. For instance, Qwen models exhibited superior performance in text translation for both base and chat variants, with the base model slightly outperforming the chat model. Conversely, Atom chat models performed poorly on title generation, even worse than their base counterparts. Additionally, for disciplinary classification tasks, post-processed Baichuan2 and Qwen base models signifi-

cantly outperformed their chat versions, indicating these models possess certain knowledge reserves in humanities and social sciences, though chat models cannot always follow instructions completely. Overall, consistent with our initial hypothesis, even in the vertical domain of humanities and social sciences, chat models almost universally outperform base models when results are not post-processed.

In summary, we evaluated six base models and seven chat models across seven tasks. To obtain a more intuitive performance comparison, we calculated final comprehensive scores using model rankings across different tasks, as metric values varied significantly across tasks. Specifically, for base models, the top-ranked model in each task received six points, decreasing sequentially to the last place. Final comprehensive scores are presented in Table 10 .

Table 10 Final Scores of Models | Model Type | Domain Knowledge Score | Academic Text Score | |———|————————|————————| | Atom-7B | - | - | | Baichuan-7B | - | - | | Baichuan2-7B | - | - | | Chinese-Alpaca-7B | - | - | | InternLM-7B | - | - | | Qwen-7B | - | - | | Atom-7B-Chat | - | - | | Baichuan2-7B-Chat | - | - | | ChatGLM-6B | - | - | | ChatGLM2-6B | - | - | | InternLM-7B-Chat | - | - | | Phoenix-Inst-Chat-7B | - | - | | Qwen-7B-Chat | - | - |

Qwen models achieved the highest scores for both base and chat variants. Notably, Qwen-7B base model demonstrated a significant lead in title generation and text translation, even rivaling some chat models. Based on Qwen-7B' s performance across tasks, it is evident that Qwen-7B incorporated conversational data during pre-training in some form, enabling superior performance on common natural language processing tasks. Additionally, in multiple-choice tasks, Baichuan2-7B-Chat and InternLM-7B-Chat almost perfectly followed instructions to output only options—a challenging task for generative language models –likely due to special instruction data added during chat model construction.

Overall, chat models generally outperformed base models across tasks. Instruction fine-tuning, as a crucial intermediate step in large model development, primarily guides the knowledge and data injected during pre-training, enabling models to better understand natural language and fully leverage their capabilities. Thus, current vertical domain model construction should emphasize both domain dataset building to ensure sufficient domain knowledge for model enhancement and the model' s own conversational abilities to enable full capability utilization.

---

5 Conclusions and Prospects

This paper constructed a series of large model evaluation tasks targeting humanities and social sciences domain knowledge and academic texts, aiming to provide references for humanities and social sciences researchers and facilitate interdisciplinary integration between humanities/social sciences research and

AI technology. Through evaluating 13 base and chat models, results show that Qwen series models exhibit the most superior performance in humanities and social sciences, followed by Baichuan2 series models, with InternLM ranking third and Atom performing the worst. Additionally, analysis of model outputs revealed that a key factor affecting base model scores is their tendency to output excessive irrelevant content and fail to understand instruction meanings. This demonstrates that teaching models to understand instructions is crucial for enhancing large model performance—a necessary second stage in training qualified large models. Training a qualified large model requires three stages: (1) Massive text pre-training, injecting large amounts of text data into models for learning to ensure sufficient knowledge reserves for addressing domain-specific questions; (2) Conversational data instruction fine-tuning, which uses multi-task instruction learning to teach models how to utilize knowledge injected during pre-training, enabling more diverse capabilities; (3) Reinforcement learning from human feedback, aligning model values with human preferences to make outputs more consistent with human expectations. Currently, most open-source chat models represent products of the second stage. Our evaluation results show that large models often lack not domain knowledge but conversational ability.

However, some exceptionally well-performing models also present another challenge for large model evaluation. Due to the closed nature of current model training data, we cannot determine whether superior performance stems from inherent capabilities or training data specialization. This raises two urgent issues: model interpretability and more comprehensive and accurate evaluation systems. Since the development of neural networks, interpretability has remained a difficult problem to overcome, and the emergence of large models has further enriched this challenge. The origins of large models' "emergent" capabilities, how to effectively explain large models, and large model safety have become new research paths for model interpretability [57]. Although this paper proposes an evaluation system for assessing large model domain capabilities based on humanities and social sciences, limitations remain. Specifically, evaluation metrics need improvement, as generated content is difficult to evaluate through quantitative metrics alone. Expanding and enriching evaluation methods and metrics is necessary for further development of large model evaluation.

Since ChatGPT' s release in 2022, large model research has gradually formed widely recognized research paths and paradigms over more than a year. Currently, large model research must address three main issues: algorithms, computing power, and data. At the algorithm level, model architectures currently show a trend of Decoder-only architecture as mainstream with a few other frameworks developing separately, gradually converging. At the computing power level, hardware limitations prevent many ordinary research teams from joining large model research, prompting researchers to explore methods to reduce training costs. The application of LoRA [58] to large models has enabled more researchers to participate in this wave of large model research. At the data level, as AI technology barriers gradually lower, data quality has become the foundation of large model performance. As information resource management

researchers, we must leverage our professional advantages to excel in data engineering, including data collection, organization, storage, and service, exploring more pathways and methods throughout the entire data construction process. We must also emphasize interdisciplinary integration, applying technological methods to various domains to drive the transition from the big data era to the era of big knowledge and big intelligence.

References

[1] Huang Shuiqing, Liu Liu, Wang Dongbo. The development and outlook of computing humanities[J]. Scientific Information Research, 2021, 3(4): 1-12.

[2] Ding Botao. Comparison and analysis of computational social science related concepts[J]. Information and Documentation Services, 2018(6): 60-67.

[3] Huang Shuiqing, Liu Liu, Wang Dongbo. The connotation, system and opportunity of computational humanities[J]. Library & Information, 2023(1): 1-11+153+145.

[4] Wang Dongbo, Liu Chang, Zhu Zihe, et al. Construction and application of pre-trained models of Siku Quanshu in orientation to digital humanities[J]. Library Tribune, 2022, 42(6): 31-43.

[5] Zhang Wei, Wang Hao, Deng Sanhong, et al. Sentiment term extraction and application of Chinese ancient poetry text for digital humanities[J]. Journal of Library Science in China, 2021, 47(4): 113-131.

[6] Zhang Qi, Wang Dongbo, Huang Shuiqing, et al. Multi-dimensional knowledge reorganization and visualization of history books: Based on Records of the Grand Historian[J]. Journal of the China Society for Scientific and Technical Information, 2022, 41(02): 130-141.

[7] Yu Xuehan, He Lin, Wang Xianqi. Research on event extraction from ancient books based on machine reading comprehension[J]. Journal of the China Society for Scientific and Technical Information, 2023, 42(3): 316-326.

[8] Zhang Ruixiang, Zhao Zhixiao. Concepts, explorations and trends of computational jurisprudence from the perspective of artificial intelligence[J]. Library & Information, 2023(1): 39-47.

[9] Liang Zhu, Shen Si, Ye Wenhao, et al. Automatic recommendation of judgment documents based on structural content features[J]. Journal of the China Society for Scientific and Technical Information, 2022, 41(2): 167-175.

[10] Wu S, Irsoy O, Lu S, et al. Bloomberggpt: a large language model for finance[J]. arXiv preprint arXiv:2303.17564, 2023.

[11] Yang H, Liu X Y, Wang C D. FinGPT: Open-source financial large language models[J]. arXiv preprint arXiv:2306.06031, 2023.

[12] Xie Q, Han W, Zhang X, et al. PIXIU: a large language model, instruction data and evaluation benchmark for finance[J]. arXiv preprint arXiv:2306.05443, 2023.

[13] Yang Y, Tang Y, Tam K Y. InvestLM: a large language model for investment using financial domain instruction tuning[J]. arXiv preprint arXiv:2309.13064, 2023.

[14] LawGPT_{zh}[EB/OL]. [2023-10-11]. https://github.com/LiuHC0428/LAW-GPT.

[15] Cui J X, Li Z J, Yan Y, et al. ChatLaw: Open-Source legal large language model with integrated external knowledge bases[J]. arXiv preprint arXiv:2306.16092, 2023.

[16] Blcuicall/taoli[EB/OL]. [2024-01-21]. https://github.com/blcuicall/taoli

[17] ECNU-ICALK/EduChat[EB/OL]. [2024-01-21]. https://github.com/icalk-nlp/EduChat.

[18] Scutcyr/BianQue[EB/OL]. [2024-01-21]. https://github.com/scutcyr/BianQue.

[19] SCIR-HI/Huatuo-Llama-Med-Chinese[EB/OL]. [2024-01-21]. https://github.com/SCIR-HI/Huatuo-Llama-Med-Chinese.

[20] Li Y, Ma S, Wang X, et al. EcomGPT: Instruction-tuning large language models with chain-of-task tasks for e-commerce[J]. arXiv preprint arXiv:2308.06966, 2023.

[21] IMOSR/MediaGPT[EB/OL]. [2024-01-21]. https://github.com/IMOSR/MediaGPT.

[22] Interim measures for the management of generative artificial intelligence services[EB/OL]. [2023-10-10]. https://www.gov.cn/zhengce/zhengceku/202307/content_{6891752}.htm.

[23] Wang Z, Xie Q, Ding Z, et al. Is ChatGPT a good sentiment analyzer? A preliminary study[J]. arXiv preprint arXiv:2304.04339, 2023.

[24] Zhang W, Deng Y, Liu B, et al. Sentiment analysis in the era of large language models: A reality check[J]. arXiv preprint arXiv:2305.15005, 2023.

[25] Pena A, Morales A, Fierrez J, et al. Leveraging large language models for topic classification in the domain of public affairs[J]. arXiv preprint arXiv:2306.02864, 2023.

[26] Zhu W, Zhou H, Gao C, et al. Research development of machine translation and large language model[C]//Proceedings of the 22nd Chinese national conference on computational linguistics (Volume 2: Frontier Forum). 2023: 30-39.

[27] Dao X Q, Le N B. Investigating the effectiveness of ChatGPT in mathematical reasoning and problem solving: Evidence from the Vietnamese national high school graduation examination[J]. arXiv preprint arXiv:2306.06331, 2023.

[28] Wu Y, Jia F, Zhang S, et al. An empirical study on challenging math problem solving with GPT-4[J]. arXiv preprint arXiv:2306.01337, 2023.

[29] Duong D, Solomon B D. Analysis of large-language model versus human performance for genetics questions[J]. European Journal of Human Genetics, 2023: 1-3.

[30] Gilson A, Safranek C W, Huang T, et al. How does ChatGPT perform on the United States medical licensing examination? The implications of large language models for medical education and knowledge assessment[J]. JMIR Medical Education, 2023, 9(1): e45312.

[31] Zhao J, Fang M, Shi Z, et al. CHBias: Bias evaluation and mitigation of Chinese conversational language models[J]. arXiv preprint arXiv:2305.11262, 2023.

[32] Parrish A, Chen A, Nangia N, et al. BBQ: A hand-built bias benchmark for question answering[C]//Findings of the Association for Computational Linguistics: ACL 2022. Dublin, Ireland: Association for Computational Linguistics, 2022: 2086-2105.

[33] Zhang W, Aljunied SVM, Gao C, et al. M3Exam: A multilingual, multimodal, multilevel benchmark for examining large language models[J]. arXiv preprint arXiv:2306.05179, 2023.

[34] Huang Y, Bai Y, Zhu Z, et al. C-eval: A multi-level multi-discipline Chinese evaluation suite for foundation models[J]. arXiv preprint arXiv:2305.08322, 2023.

[35] Xu L, Li A, Zhu L, et al. SuperCLUE: A comprehensive Chinese large language model benchmark[J]. arXiv preprint arXiv:2307.15020, 2023.

[36] Open LLM Leaderboard - A HuggingFace Space by HuggingFaceH4[EB/OL]. [2023-10-10]. https://huggingface.co/spaces/HuggingFaceH4/open_{{{llm}}}_{{leaderboard}}}.

[37] Chinese-llm-benchmark[EB/OL]. (2023-10-10)[2023-10-10]. https://github.com/jeinlee1991/chinese-llm-benchmark.

[38] FlagAlpha/Llama2-Chinese[EB/OL]. [2023-10-07]. https://github.com/FlagAlpha/Llama2-Chinese.

[39] Baichuan-Inc/Baichuan-7B: A large-scale 7B pretraining language model developed by Baichuan-Inc.[EB/OL]. [2023-10-07]. https://github.com/baichuan-inc/Baichuan-7B.

[40] Baichuan-Inc/Baichuan2: A series of large language models developed by Baichuan-Inc[EB/OL]. [2023-10-07]. https://github.com/baichuan-inc/Baichuan2.

[41] Ymcui/Chinese-LLaMA-Alpaca[EB/OL]. [2023-10-07]. https://github.com/ymcui/Chinese-LLaMA-Alpaca.

[42] InternLM/InternLM[EB/OL]. [2023-10-07]. https://github.com/InternLM/InternLM.

[43] QwenLM/Qwen[EB/OL]. [2023-10-07]. https://github.com/QwenLM/Qwen.

[44] THUDM/ChatGLM-6B[EB/OL]. [2023-10-07]. https://github.com/THUDM/ChatGLM-6B.

[45] THUDM/ChatGLM2-6B[EB/OL]. [2023-10-07]. https://github.com/THUDM/ChatGLM2-6B.

[46] FreedomIntelligence/LLMZoo[EB/OL]. [2023-10-07]. https://github.com/FreedomIntelligence/LLMZoo.

[47] Wei Xiangqing. Exploring the Construction of Chinese-English Terminology Knowledge Base in Humanities and Social Sciences: Concepts and Methods[M]. 1st ed. Nanjing: Nanjing University Press, 2022.

[48] ChatGPT Prompt Engineering for Developers[EB/OL]. [2023-10-07]. https://www.deeplearning.ai/short-courses/chatgpt-prompt-engineering-for-developers/.

[49] Liu P, Yuan W, Fu J, et al. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing[J]. ACM Computing Surveys, 2023, 55(9): 1-35.

[50] Papineni K, Roukos S, Ward T, et al. BLEU: A method for automatic evaluation of machine translation[C]//Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. 2002: 311-318.

[51] Lin C Y. ROUGE: A package for automatic evaluation of summaries[C]//Text Summarization Branches Out. Barcelona, Spain: Association for Computational Linguistics, 2004: 74-81.

[52] Popovic M. chrF: Character n-gram F-score for automatic MT evaluation[C]//Proceedings of the Tenth Workshop on Statistical Machine Translation. 2015: 392-395.

[53] Post M. A call for clarity in reporting BLEU scores[C]//Proceedings of the Third Conference on Machine Translation: Research Papers. Brussels, Belgium: Association for Computational Linguistics, 2018: 186-191.

[54] Pillutla K, Swayamdipta S, Zellers R, et al. MAUVE: Measuring the gap between neural text and human text using divergence frontiers[J]. Advances in Neural Information Processing Systems, 2021, 34: 4816-4828.

[55] Radford A, Wu J, Child R, et al. Language models are unsupervised multitask learners[J]. OpenAI Blog, 2019, 1(8): 9.

[56] Evaluate Metric[EB/OL]. [2023-10-08]. https://huggingface.co/evaluate-metric.

[57] Zhao H, Chen H, Yang F, et al. Explainability for large language models: A survey[J]. arXiv preprint arXiv:2309.01029, 2023.

[58] Hu E J, Shen Y, Wallis P, et al. LoRA: Low-rank adaptation of large language models[J]. arXiv preprint arXiv:2106.09685, 2021.

Author Contributions: Zhao Zhixiao: Experimental process implementation, paper writing; Hu Die: Experimental process refinement, paper writing; Liu Chang: Experimental process refinement, paper revision; Shen Si: Paper revision; Wang Dongbo: Overall framework and research direction formulation.

Note: Figure translations are in progress. See original paper for figures.