Can Large Language Model Comprehend Ancient Chinese? A Preliminary Test on ACLUE

Yixuan Zhang **Haonan Li**
Mohamed bin Zayed University of Artificial Intelligence, UAE
haonan.li@mbzuai.ac.ae

Abstract

Large language models (LLMs) have showcased remarkable capabilities in understanding and generating language. However, their ability in comprehending ancient languages, particularly ancient Chinese, remains largely unexplored. To bridge this gap, we present ACLUE, an evaluation benchmark designed to assess the capability of language models in comprehending ancient Chinese. ACLUE consists of 15 tasks cover a range of skills, spanning phonetic, lexical, syntactic, semantic, inference and knowledge. Through the evaluation of eight state-of-the-art LLMs, we observed a noticeable disparity in their performance between modern Chinese and ancient Chinese. Among the assessed models, ChatGLM2 demonstrates the most remarkable performance, achieving an average score of 37.4%. We have made our code and data public available.[1]

1 Introduction

The study of ancient languages provides valuable insights into the past civilizations' thoughts, languages, societies, and histories (Zhiming, 1990; Woodard, 2008; Bouchard-Côté et al., 2013). Ancient China, as one of the oldest civilizations, has left a significant impact on contemporary societies including Japan, Korea, and Vietnam. However, existing research in ancient Chinese language processing have primarily focused on specific time periods or genres (Yan et al., 2016; Xie et al., 2019; Liu et al., 2020; Hu et al., 2021; Tian et al., 2021). Typically, the previously proposed models require customized fine-tuning for particular tasks.

Recently, the significant advancements made in large language models (LLMs) underscore their remarkable proficiency across a range of tasks, showcasing their potential in performing various tasks without the need for fine-tuning (Brown et al., 2020; Scao et al., 2022; Touvron et al., 2023; Muennighoff et al., 2022; Zeng et al., 2023). These models encapsulate extensive knowledge and sophisticated reasoning capabilities. Notably, the emergence of ChatGPT (OpenAI, 2023) and Chinese-oriented LLMs such as ChatGLM (Zeng et al., 2023), has accentuated their remarkable ability in comprehending and generating modern language. However, due to the lack of ancient language benchmarks, the abilities of LLMs in handling ancient language remains largely unexplored.

We present the Ancient Chinese Language Understanding Evaluation (ACLUE), an evaluation benchmark consisting of 15 tasks. These tasks are derived from a combination of manually curated questions from publicly available resources, and automatically generated questions from classical Chinese language corpora. The range of questions span from the Xia dynasty (2070 BCE) to the Ming dynasty (1368 CE), covering a broad temporal range. Similar to the well-established LLM benchmarks such as ARC (Clark et al., 2018) and MMLU (Hendrycks et al., 2021), ACLUE adopts multiple-choice question format for all tasks. This ensures simplicity and uniformity in evaluating models, accommodating variations in different training or fine-tuning procedures and prompting methodologies.

In our preliminary experiments, we assessed the performance of 8 advanced LLMs, where the Chinese LLM ChatGLM2 demonstrates the best performance with an average accuracy of 37.4%, slightly surpassing ChatGPT. However, considering the baseline accuracy of 25% from random guessing and the average accuracy of around 50% achieved by the same models on contemporary modern Chinese benchmarks such as AGIEval (Zhong et al., 2023) and CMMLU (Li et al., 2023), we believe there is still ample room for improvement in the proficiency of existing LLMs in understanding ancient Chinese.

[1] https://github.com/isen-zhang/ACLUE

Proceedings of the Ancient Language Processing Workshop associated with RANLP-2023, pages 80–87,
held in Varna Bulgaria, Sept 8, 2023.
https://doi.org/10.26615/978-954-452-087-8_009

2 ACLUE Benchmark

ACLUE consists of 15 tasks that encompassing lexical, syntactic, semantic, inference, and general knowledge of ancient Chinese. The details of the tasks are provided in Appendix A, where basic statistics can be found in Table 2, and examples of each task are listed in Table 3. The questions cover a wide range of genres, including poetry, prose, classical novels, couplets, historical records, and biographies, spanning the period from 2070 BCE to 1368 CE. Among the 15 tasks, 8 were automatically generated using existing corpora or datasets, 5 were collected from freely available standard tests, and 2 were directly sourced from other work. Each task includes 100 to 500 questions, exceeding the number required for testing a human participant.

ACLUE serves as an evaluation suite for LLMs ability in understanding ancient Chinese without task-specific fine-tuning. To ensure fair comparison among different models trained with varying approaches, all tasks are formatted into multiple-choice questions with four choices, of which only one is correct. The task details and dataset construction process are elaborated in this section.

2.1 Lexical Tasks

We create three lexical tasks using the ancient Chinese corpus, which includes over 50,000 word sense annotations and 3,000 named entity annotations (Shu et al., 2021).

Polysemy resolution aims to understand the different senses or meanings of words. Two types of questions are created: one asks which character in a given sentence carries a particular meaning, while the other requires identifying the meaning of a character within the sentence.

Homographic character resolution focuses on recognizing homographic characters in ancient Chinese texts. Homographic characters, also known as "通假字" (tōng jiǎ zì) in Chinese, are substitutions of characters in ancient Chinese texts with others that have similar pronunciation or appearance.

Named entity recognition focuses on identifying named entities (e.g., names of people, places, dynasties, etc.) in ancient Chinese texts. Two types of questions are created: one type asks for the specific entity type of a given entity within a contextual sentence, while the other type asks in which context a Chinese word represents an entity.

2.2 Syntactic and Semantic Tasks

Sentence segmentation is a task that involves choosing the correct segmentation of a given sentence. Since ancient Chinese lacks punctuation marks, accurate sentence segmentation becomes crucial for analyzing syntax and semantics of a sentence. We create the task by sampling sentences from the Classical-Modern Chinese Corpus,[2] which provides labeled sentence segmentation. To create false options, we manipulate the original punctuation marks by moving, adding, or deleting them.

Couplet prediction involves predicting the most likely second line of a Chinese couplet based on a given first line. Chinese couplet, also known as "对联" (duì lián), is a traditional form of poetic expression consisting of two lines of verse. The two lines are expected to match in terms of meaning, rhyme, and other poetic elements. We construct this task using a couplet dataset.[3]

Poetry context prediction is a task constructed using the Chinese-poetry corpus.[4] The objective of this task is to select the most likely next or previous sentence given a specific sentence from a poem.

2.3 Inference

Poem quality estimation task is constructed based on dataset proposed by Yi et al. (2018), which consists of 173 Chinese quatrains, with each one being rated for fluency, coherence, and meaningfulness on a scale of 0 to 5 by human expert. We randomly select four poems and create questions asking models to identify the best or worst poem based on a specific criterion. To ensure clear distinctions, we maintain a minimum score differences of 2 between the correct option and the other options. The task aims to evaluate the ability of models to compare the quality of Chinese quatrains.

Reading comprehension is based on the AGIEval dataset (Zhong et al., 2023). It contains a subset of Chinese Gaokao questions. We select questions that contains ancient Chinese text from this subset.

Poetry sentiment analysis involves predicting the sentiment of an entire poem or parts of a poem, determining whether it is positive, neutral, or negative. We utilize a dataset proposed by Shao et al. (2021), which contains 5,000 poems. Each poem

[2] https://github.com/NiuTrans/Classical-Modern

[3] https://github.com/wb14123/couplet-dataset

[4] https://github.com/chinese-poetry

以下是关于 古代文学知识 的单项选择题，请直接给出正确答案的选项。
Here are some multiple-choice questions about Ancient Chinese literature, please provide the correct answer choice directly.

题目：下列诗句中，属于杜牧咏史诗的是：
Question: Among the following lines of poetry, the one that belongs to Du Mu's historical poem is:
A. 旧时王谢堂前燕，飞入寻常百姓家
In former times, the swallows in front of the halls of Wang and Xie flew into the homes of ordinary people
B. 长空澹澹孤岛没，万古销沉向此中
The vast sky engulfed the desolate island, and for eternity it sank into this place.
C. 千寻铁锁沉江底，一片降幡出石头
Thousands of chains sank to the bottom of the river, and a stone emerged with a descending flag
D. 三百年间同晓梦，钟山何处有龙盘
For three hundred years, the same dream awakened at dawn, where on Zhongshan Mountain can a dragon coil
答案是： (Answer:)

Figure 1: An examples from ACLUE. English translations are provided for better readability.

and its individual sentences are labeled with fine-grained sentiment categories, including negative, implicit negative, neutral, implicit positive, and positive sentiments. We merge implicit negative and implicit positive labels with their respective categories to address ambiguity.

Poetry appreciation is manually curated from openly accessible online resources.

2.4 Knowledge-intensive Tasks

Ancient Chinese knowledge tasks cover various subjects, including **ancient Chinese medical**, **ancient Chinese literature**, **traditional Chinese culture**, and **ancient Chinese phonetics**. To create these tasks, we collected relevant questions from various online open resources. Additionally, we extracted a subset of questions from the CMMLU dataset (Li et al., 2023), which consist of questions at the high-school level in current Chinese education. This selection allows us to form the tasks of **basic ancient Chinese**.

3 Experiment

To provide an overview of the language ability of existing open-sourced LLMs on ancient Chinese, we assess 8 models including 4 multilingual models: ChatGPT (OpenAI, 2023), LLaMA (Touvron et al., 2023), Falcon (Almazrouei et al., 2023), BLOOMZ (Muennighoff et al., 2022), and 4 Chinese models: ChatGLM (Du et al., 2022), Baichuan,[5] ChatGLM2 (Zeng et al., 2023), and MOSS (OpenLMLab, 2023). Details about these models are introduced in Appendix C.

For models optimized to function as chatbots, such as ChatGPT and ChatGLM, we generate output and use regular expressions to extract the answer key. For other models, we directly obtain the probability of the next tokens after the prompt and selected the one with the highest probability among the answer keys (i.e., 'A','B','C','D'). We employ both zero-shot (do not provide examples) and in-context five-shot (provide few examples) evaluation. An example of evaluation instance is shown in Figure 1.

3.1 Results

Table 1 shows the zero-shot performance of all models. The five-shot results are similar to the zero-shot results, suggesting that models can comprehend the task without additional demonstrations. Overall, the Chinese model ChatGLM2 demonstrates the best performance, with an average accuracy of 37.4%. Moreover, its performance on almost all tasks is above the random guessing (25%). The multilingual model ChatGPT achieves a slightly lower accuracy of 36.9%, compared to ChatGLM2, yet it maintains relatively consistent performance in terms of standard deviation.

Regarding specific tasks, we have several findings: (1) BLOOMZ exhibits exceptional performance in *couplet prediction* (T5), achieving an accuracy of 60.2%. This accuracy is nearly double that of most other models, possibly due to BLOOMZ's training set, xP3, having overlaps with our data source. Similar, ChatGLM2 may have been exposed to the original texts used for *sentence segmentation* (T4) and *poetry appreciation* (T9), which explains its proficient performance in these tasks. (2) All models face challenges in the *homographic character resolution* (T2), with performance close to random guessing. This issue likely arises because the auto-regressive training objective does not emphasize understanding of homographic concepts. (3) *Reading comprehension* (T8) poses a considerable challenge for all models due to the extreme long length of the question (nearly 1,000 tokens on average). Specifically, BLOOMZ, LLaMA, and Baichuan are significantly affected, exhibiting lower performance on this task compared to their average across other tasks. This observation suggests that these models may lack adequate support for processing very long input.

[5] https://github.com/baichuan-inc/baichuan-7B

Model	Lexical			Syntactic	Semantic		Inference				Knowledge					Overall
	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	T15	
ChatGLM2	**45.4**	24.4	34.8	**46.4**	39.8	24.6	28.3	29.7	**42.7**	**52.6**	28.9	**50.7**	34.6	43.8	**35.0**	**37.4**$_{\pm 8.9}$
ChatGPT	41.8	20.6	**41.2**	43.0	45.4	27.4	**39.7**	**39.6**	38.8	47.8	29.3	43.4	34.6	33.8	27.0	36.9$_{\pm 7.6}$
BLOOMZ	45.2	22.4	35.6	32.2	**60.2**	27.2	31.5	17.8	26.2	45.2	29.7	44.1	**39.3**	**44.4**	29.0	35.3$_{\pm 10.7}$
ChatGLM	39.6	19.4	39.4	36.6	37.2	23.4	30.8	32.7	30.1	43.8	29.3	36.8	30.8	40.6	27.0	33.2$_{\pm 6.6}$
Falcon	40.4	**28.8**	21.2	32.6	37.2	**31.4**	36.9	22.8	31.1	43.8	**30.5**	30.1	30.3	36.9	26.0	32.0$_{\pm 6.0}$
Baichuan	31.6	26.4	22.0	33.0	37.2	27.8	30.3	16.8	25.2	38.2	27.3	36.0	37.0	41.9	31.0	30.8$_{\pm 6.5}$
LLaMA	36.4	22.2	26.4	33.0	29.6	29.6	31.5	18.8	24.3	41.8	24.5	23.5	29.4	29.4	31.0	28.8$_{\pm 5.6}$
MOSS	30.6	27.6	25.8	24.0	30.0	25.0	29.8	27.7	21.4	30.8	26.5	22.1	24.6	22.5	26.0	26.3$_{\pm 3.0}$

Table 1: Zero-shot average accuracy of all models. The overall results are averaged (with standard deviation) over all tasks. T1: Polysemy resolution, T2: Homographic character resolution, T3: Named entity recognition, T4: Sentence segmentation, T5: Couplet prediction, T6: Poetry context prediction, T7: Poetry quality estimation, T8: Reading comprehension, T9: Poetry appreciation, T10: Poetry sentiment analysis, T11: Basic ancient Chinese, T12: Traditional Chinese culture, T13: Ancient Chinese medical, T14: Ancient Chinese literature, T15: Ancient Chinese phonetics.

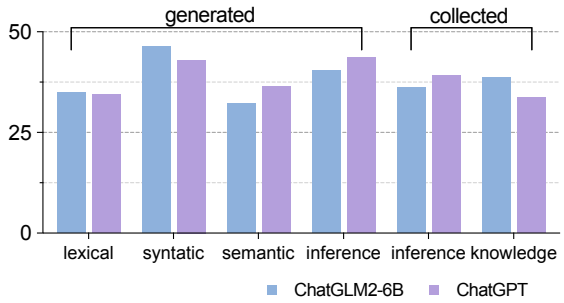

Figure 2: The performance of ChatGPT and ChatGLM2 on ACLUE of different categories.

Based on data origin, we divide the tasks into two categories: auto-generated and manually collected. In Figure 2, we compare the performance of ChatGPT and ChatGLM2, the best multilingual and Chinese models, respectively. We find that while ChatGLM2 exhibits superior overall performance on ACLUE, its dominance only observed in the auto-generated syntactic tasks and collected knowledge categories. More comparison results are provided in Appendix B.

In terms of data quality and reliability, auto-generated questions within ACLUE were slightly less intricate than collected questions, but the difference was not significant. This suggests that the auto-generated questions hold reasonable potential for effectively evaluating models' grasp of ancient Chinese language.

4 Related Work

A lot of research has been conducted on various aspects of ancient Chinese language processing, encompassing topics such as ancient Chinese to modern Chinese translation (Liu et al., 2020), Chinese couplets generation (Yan et al., 2016; Yuan et al., 2019; Qu et al., 2022), Classic Chinese poem generation (Yi et al., 2017; Yang et al., 2018; Guo et al., 2019; Xie et al., 2019; Zhao et al., 2022; Ma et al., 2023), and ancient Chinese sentence segmentation (Han et al., 2018; Hu et al., 2021), as well as general language model pre-training (Tian et al., 2021). However, many of these studies focus on specific types or literary formats that were popular during specific time periods.

Recently, large language models have demonstrated remarkable language understanding and generation capabilities (Brown et al., 2020; Scao et al., 2022; Almazrouei et al., 2023). Researchers have began to evaluate these LLMs based on their performance across a wide range of tasks (Touvron et al., 2023; Muennighoff et al., 2022; OpenAI, 2023). However, the absence of a comprehensive evaluation benchmark poses a challenge in assessing the performance of LLMs in ancient language understanding. Existing ancient Chinese evaluation datasets either have a narrow focus on specific tasks, limiting the scope of evaluation, or require model fine-tuning prior to evaluation. In contrast, ACLUE provides a natural support for evaluation under zero-shot and in-context learning settings, making it more compatible with LLMs.

5 Conclusion

We propose ACLUE, the first evaluation benchmark for ancient Chinese language understanding. Our preliminary evaluation of 8 large language models reveals that, despite their exceptional performance in modern language understanding, they struggle with even basic tasks in ancient Chinese. Through analysis, we illustrate that the auto-generated questions possess similar difficulty levels to those found in actual school tests.

References

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Heslow, Julien Launay, Quentin Malartic, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. Falcon-40B: an open large language model with state-of-the-art performance.

Alexandre Bouchard-Côté, David Hall, Thomas L. Griffiths, and Dan Klein. 2013. Automated reconstruction of ancient languages using probabilistic models of sound change. *Proc. Natl. Acad. Sci. USA*, 110(11):4224–4229.

Tom B. Brown, Benjamin Mann, Nick Ryder, et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the AI2 reasoning challenge. *CoRR*, abs/1803.05457.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

Zhipeng Guo, Xiaoyuan Yi, Maosong Sun, Wenhao Li, Cheng Yang, Jiannan Liang, Huimin Chen, Yuhui Zhang, and Ruoyu Li. 2019. Jiuge: A human-machine collaborative chinese classical poetry generation system. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28 - August 2, 2019, Volume 3: System Demonstrations*, pages 25–30. Association for Computational Linguistics.

Xu Han, Hongsu Wang, Sanqian Zhang, Qunchao Fu, and Jun S. Liu. 2018. Sentence segmentation for classical chinese based on LSTM with radical embedding. *CoRR*, abs/1810.03479.

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

Renfen Hu, Shen Li, and Yuchen Zhu. 2021. Knowledge representation and sentence segmentation of ancient chinese based on deep language models. *Journal of Chinese Information Processing*, 35(4):8–15.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2023. Cmmlu: Measuring massive multitask language understanding in chinese.

Dayiheng Liu, Kexin Yang, Qian Qu, and Jiancheng Lv. 2020. Ancient-modern chinese translation with a new large training dataset. *ACM Trans. Asian Low Resour. Lang. Inf. Process.*, 19(1):6:1–6:13.

Jingkun Ma, Runzhe Zhan, and Derek F. Wong. 2023. Yu sheng: Human-in-loop classical chinese poetry generation system. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics. EACL 2023 - System Demonstrations, Dubrovnik, Croatia, May 2-4, 2023*, pages 57–66. Association for Computational Linguistics.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M. Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2022. Crosslingual generalization through multitask finetuning. *CoRR*, abs/2211.01786.

OpenAI. 2023. Gpt-4 technical report.

OpenLMLab. 2023. Moss.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.

Qian Qu, Jiancheng Lv, Dayiheng Liu, and Kexin Yang. 2022. Coupgan: Chinese couplet generation via encoder-decoder model and adversarial training under global control. *Soft Comput.*, 26(15):7423–7433.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilic, et al. 2022. BLOOM: A 176b-parameter open-access multilingual language model. *CoRR*, abs/2211.05100.

Yizhan Shao, Tong Shao, Minghao Wang, Peng Wang, and Jie Gao. 2021. A sentiment and style controllable approach for chinese poetry generation. In *CIKM '21: The 30th ACM International Conference on Information and Knowledge Management, Virtual Event, Queensland, Australia, November 1 - 5, 2021*, pages 4784–4788. ACM.

Lei Shu, Yiluan Guo, Huiping Wang, Xuetao Zhang, and Renfen Hu. 2021. 古汉语词义标注语料库的构建及应用研究(the construction and application of Ancient Chinese corpus with word sense annotation). In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 549–563, Huhhot, China. Chinese Information Processing Society of China.

Huishuang Tian, Kexin Yang, Dayiheng Liu, and Jiancheng Lv. 2021. Anchibert: A pre-trained model

for ancient chinese language understanding and generation. In *International Joint Conference on Neural Networks, IJCNN 2021, Shenzhen, China, July 18-22, 2021*, pages 1–8. IEEE.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Roger D Woodard. 2008. *The ancient languages of Europe*. Cambridge University Press.

Zhuohan Xie, Jey Han Lau, and Trevor Cohn. 2019. From shakespeare to li-bai: Adapting a sonnet model to chinese poetry. In *Proceedings of the The 17th Annual Workshop of the Australasian Language Technology Association, ALTA 2019, Sydney, Australia, December 4-6, 2019*, pages 10–18. Australasian Language Technology Association.

Rui Yan, Cheng-Te Li, Xiaohua Hu, and Ming Zhang. 2016. Chinese couplet generation with neural network structures. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Cheng Yang, Maosong Sun, Xiaoyuan Yi, and Wenhao Li. 2018. Stylistic chinese poetry generation via unsupervised style disentanglement. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 3960–3969. Association for Computational Linguistics.

Xiaoyuan Yi, Ruoyu Li, and Maosong Sun. 2017. Generating chinese classical poems with RNN encoder-decoder. In *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data - 16th China National Conference, CCL 2017, - and - 5th International Symposium, NLP-NABD 2017, Nanjing, China, October 13-15, 2017, Proceedings*, volume 10565 of *Lecture Notes in Computer Science*, pages 211–223. Springer.

Xiaoyuan Yi, Maosong Sun, Ruoyu Li, and Wenhao Li. 2018. Automatic poetry generation with mutual reinforcement learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3143–3153, Brussels, Belgium.

Shengqiong Yuan, Luo Zhong, Lin Li, and Rui Zhang. 2019. Automatic generation of chinese couplets with attention based encoder-decoder model. In *2nd IEEE Conference on Multimedia Information Processing and Retrieval, MIPR 2019, San Jose, CA, USA, March 28-30, 2019*, pages 65–70. IEEE.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Zhiyuan Liu, Peng Zhang, Yuxiao Dong, and Jie Tang. 2023. GLM-130b: An open bilingual pre-trained model. In *The Eleventh International Conference on Learning Representations*.

Jiaqi Zhao, Ting Bai, Yuting Wei, and Bin Wu. 2022. Poetrybert: Pre-training with sememe knowledge for classical chinese poetry. In *Data Mining and Big Data - 7th International Conference, DMBD 2022, Beijing, China, November 21-24, 2022, Proceedings, Part II*, volume 1745 of *Communications in Computer and Information Science*, pages 369–384. Springer.

Bao Zhiming. 1990. Language and world view in ancient china. *Philosophy East and West*, 40(2):195–219.

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. *CoRR*, abs/2304.06364.

A Data details

The Table 2 listed the Chinese, category, and origin of the tasks in ACLUE, and the Table 3 provides examples for each task.

B Further analysis

The performance comparison of all LLMs on different data origins is illustrated in Figure 3. Evaluating the LLMs' performance on auto-generated questions versus manually collected questions in ACLUE, we found that while the generated questions were less intricate than the collected ones, the difference was not significant. This indicates a comparable level of difficulty between the two types of questions. Among all the models, only ChatGLM2 demonstrated better performance on collected questions compared to auto-generated questions, which may indicate exposure to the original question texts used in ACLUE.

C Models being Evaluated

BLOOMZ is derived from BLOOM through fine-tuning on a crosslingual task mixture (xP3), which is an instruction-following dataset. BLOOMZ exhibits competitive performance with models that have a larger number of parameters across various non-generation tasks.

Task	Total Q.	Avg. len	Task (zh)	Category	Origin
Named entity recognition	500	138	古汉语命名体识别	lexical	generated
Polysemy resolution	500	116	古文单字多义	lexical	generated
Homographic character resolution	500	137	通假字	lexical	generated
Sentence segmentation	500	210	古文断句	syntatic	generated
Couplet prediction	500	62	对联预测	semantic	generated
Poetry context prediction	500	77	古诗词上下句预测	semantic	generated
Poetry sentiment analysis	500	60	诗词情感分类	inference	generated
Poem quality estimation	406	118	古诗词质量评估	inference	generated
Ancient Chinese medical	211	38	医古文	knowledge	collected
Ancient Chinese literature	160	44	古代文学知识	knowledge	collected
Traditional Chinese culture	136	59	国学常识	knowledge	collected
Poetry appreciation	103	258	古诗词曲鉴赏	inference	collected
Basic ancient Chinese	249	52	基础古汉语知识	knowledge	collected
Reading comprehension	101	982	古文阅读理解	inference	collected
Ancient Chinese phonetics	101	50	古音学	knowledge	collected

Table 2: ACLUE task overview. We list the total number of questions (Total Q.), average question length counted in Chinese characters (Avg. len), task names in Chinese, task type, and data origin type.

Baichuan-7b is an open-source large-scale pre-trained model developed by Baichuan Intelligence. Built on the Transformer architecture, it adopts the same model design as LLaMA. This 7-billion-parameter model was trained on approximately 1.2 trillion tokens using proprietary Chinese-English bilingual corpora, with optimization focused on Chinese.

ChatGLM-6B is bidirectional dense model pre-trained using the General Language Model (GLM) algorithm developed by Tsinghua University. It supports bilingual (Chinese and English) language processing. ChatGLM is a version of GLM that has been supplemented with supervised fine-tuning, feedback bootstrap, and reinforcement learning with human feedback, specifically optimized for Chinese question answering (QA) and dialogue tasks.

ChatGLM2-6B is the second generation of ChatGLM. It uses the hybrid objective function of GLM, and has undergone pre-training with 1.4T bilingual tokens and human preference alignment training. It offers enhanced performance and an expanded context length of 32K. With efficient inference using Multi-Query Attention technology, it achieves efficient inference with higher speed and lower memory usage.

ChatGPT is a GPT model developed by OpenAI and fine-tuned using reinforcement learning from human feedback (RLHF). As a commercial product, specific details about its model size, training data, and training process are not disclosed.

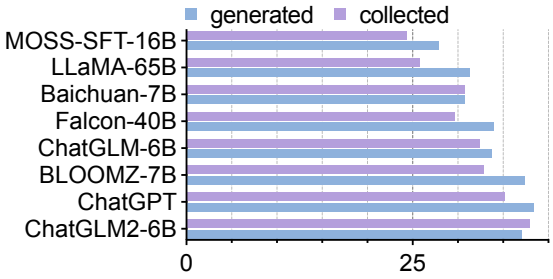

Figure 3: The performance comparison of LLMs on ACLUE across different data origins.

LLaMA-65B is an auto-regressive language model proposed by Meta. It incorporates several structural improvements over the vanilla transformer and is trained on a mixture of publicly available data sources. LLaMA has demonstrated comparable or even superior performance to models that are ten times its size.

Falcon-40B is a decoder-only model created by TII and trained on 1,000B tokens of RefinedWeb (Penedo et al., 2023) data. Due to the high quality of its training data, Falcon-40B performs competitively with LLaMA-65B on various benchmarks.

MOSS is an open-source Chinese language model proposed by Fudan University. It matches ChatGPT in terms of training scale and alignment techniques. MOSS-SFT is initialized with CodeGen and further pre-trained on 100B Chinese tokens and 20B English tokens. The SFT (supervised fine-tuned) version of MOSS-SFT enables the model to follow instructions in multi-turn dialogues.

ID	Task	Example
T1	单字多义 Polysemy resolution	下列选项中对“此神农之所以[长]，而尧舜之所以章也。”这句话中的“长”字理解正确的是( ) **A.** 首领　　B.排行第一, 长子　　C.长处, 专长　　D. 长大，成年
T2	通假字 Homographic character resolution	列选项中[]内的“红”字是通假字的是( ) **A.** 吾已食禄，又夺园夫女[红]利。　　B. 晓看[红]湿处，花重锦官城 C. [红] 芳满院参差折，绿醑盈杯次第衔。　　D. 竹缘浦以被绿，石照涧而映[红] 。
T3	命名体识别 Named entity recognition	下列选项中[]内的“阳”字代表了地名的是( ) **A.** 夫人授兆丹书真文、月中玉。　　B. 令飞升上造洞[阳]之宫。 C. 今朝日[阳]里，梳落数茎丝。　　D. 晓发碧水[阳]，暝宿金山寺。
T4	古文断句 Sentence segmentation	以下选项断句正确的是( ) **A.** 史记/曰/秦使武安君白起攻赵/赵发兵拒秦/秦大破赵於长平/ B. 史记/曰秦/使武安君白起攻赵/赵发兵拒秦/秦大破赵於长平 C. 史记/曰/秦使武安君白起攻赵/赵发兵拒秦秦大/破赵於长/平 D. 史/记曰秦使武安君白起攻赵/赵发兵拒秦秦大/破赵於长平
T5	对联 Couplet prediction	“兔去龙来，交替人间春好景”的下联最可能是( ) **A.** 莺歌燕舞，和谐社会岁祥光。　　B. 香遗书案，传家苦读育春风。 C. 赛龙夺锦，万人江岸闹端阳。　　D. 情牵天下，凭谁设榻效陈蕃。
T6	古诗词上下句预测 Poetry context prediction	“何秣候明，便可一横江。”的上一句是( ) **A.** 江藉草作寒食，雨梨花思故。　　B. 孰知文有忌，情至自生哀。 C. 樽前唱醉翁曲，歌花舞催。　　D. 千村落呼客，山南北花吹香。
T7	古诗词质量评估 Poem quality estimation	下列古诗词前后文连贯性最差的是( ) **A.** 阴雨难侵牖|春虫足哺儿|年年秋报喜|牛女有佳期 B.富贵良非愿|林泉毕此生|酒因随量饮|诗或偶然成 C.久不闻山歌|南风五月多|牧童呼伴侣|吹笛下西坡 D.今日骐阁|当年鹦鹉洲|寄书愁不达|书达得无愁
T8	古文阅读理解 Reading comprehension	下列对原文有关内容的理解和分析，表述不正确的一项是( ) 谢贞，字元正，陈郡阳夏人，晋太傅安九世孙也。父藺，正员外郎，... ... 察因启曰：“贞有一子年六岁。”即有敕长给衣粮。（节选自《陈书·列传第二十六》，有删改）。【注】惠连：谢惠连，南朝宋文学家。 A. 谢贞天性聪慧，小时候读过不少典籍，有的读过就能背诵，有的粗通大意；他八岁时写的诗就深得长辈称赞。 B. 谢贞受府长史周确委托，为他撰写辞让都官尚书的表文。陈后主读过之后，怀疑该表文不是周确亲笔所作。 **C.** 谢贞非常孝顺，小时候祖母因病难以进食，他便也不进食；父亲去世他悲痛欲绝，之后，奉养母亲未曾间断。 D. 母亲去世后，谢贞一心守丧，极度悲痛，骨瘦如柴，令人叹息。他忧病而死后，后主下令长期供他儿子吃穿。
T9	古诗词曲鉴赏 Poetry appreciation	下列对这首诗的赏析，不正确的一项是( ) 《幽居初夏》陆游。湖山胜处放翁家，槐柳阴中野径斜。水满有时观下鹭，草深无处不鸣蛙。鐸龙已过头番笋，木笔犹开第一花。叹息老来交旧尽，睡来谁共午瓯茶。 A. 首句“湖山”二字总冒全篇，勾勒环境，笔力开张，巧妙地从山光水色中引出“幽居”。 **B.** 首句概言**“湖山胜处”**，颔联写湖，是近处宽处静景；颈联写庭院周围，是远处细处动态。 C. 诗中写放翁心中郁结与柳宗元《小石潭记》中写“以其境过清”时的心境相似。 D. 本诗前三联写景，尾联结情，景情相衬，描写与抒情紧密关联，脉络清晰。
T10	诗词情感分类 Poetry sentiment analysis	古诗词“庭前芍药妖无格|池上芙蕖净少情|唯有牡丹真国色|花开时节动京城“的整体情感是( ) **A.** 积极的　　B. 消极的　　C. 中性的　　D. 无法判断
T11	国学常识 Basic ancient Chinese	"近朱者赤，近墨者黑"所蕴含的道理和下列哪句话最相似？( ) A. 青出于蓝，而胜于蓝。　　**B.** 蓬生麻中，不扶而直。 C. 公生明，偏生暗。　　D. 三天打鱼两天晒网
T12	古汉语知识 Traditional Chinese culture	下列句中，含有双宾语的一句是( ) A. 夫何之有?　　**B.** 重之而之。　　C. 兔不可得，而身宋笑。　　D. 甚矣，汝之不惠!
T13	医古文 Ancient Chinese medical	以下除（）之外，都有病愈之义。 A. 已　　B. 起　　**C.** 性　　D. 差
T14	古代文学知识 Ancient Chinese literature	杜甫《春望》中的“感时花溅泪，恨别鸟惊心”所反映的是( ) A. 早年的读书和漫游生活。　　B. 困居长安十年时的感受。 **C.** **“安史之乱”**时的国恨家愁。　　D. 晚年漂泊西南的客旅生活。
T15	古音学 Ancient Chinese phonetics	下列字在古代的声母、调类、等和开合口标注错误的是( ) **A.** 温（影母平声二等开）　　B. 权（群母平声三等合） C. 空（溪母平声一等合）　　D. 狂（群母平声三等合）

Table 3: ACLUE tasks examples.