CCL25-Eval任务7系统报告：学而不思则罔?

郑陈锐[1], 朱奕澄[1], 王欣雨[1], 姜伟麟[1], 吴会腾[2]

[1]华东师范大学中文系
[2]山东大学法学系
{10194800486,10210110437,10220110478,10240110028}@stu.ecnu.edu.cn
202100171091@mail.sdu.edu.cn

摘要

以Deepseek-R1为代表，“思考”普遍被认为是一种提高大语言模型性能的方法。在CCL25-Eval“争鸣”中文阅读理解任务下，本文分别探索了“思考”和“非思考”两种模型在这项任务下的潜力。具体来说，在古代文学知识理解任务中，本文构建了古汉语特定领域的知识数据集，用大模型蒸馏了思考数据集，整理了高质量思考数据集，在这些数据基础之下同样lora微调，发现思考模型虽然性能有巨大提升，但依旧比不上原本的非思考模型。最后，开源并提交了基于Qwen2.5的SongPanda模型。

关键词： 大语言模型；中文阅读理解；推理模型

System Report for CCL25-Eval Task 7: Learning Without Thinking Leads to Confusion?

Chenrui Zheng[1], Yicheng Zhu[1], Xinyu Wang[1], Weilin Jiang[1], Huiteng Wu[2]

[1]Department of Chinese, East China Normal University
[2]School of Law, Shandong University
{10194800486,10210110437,10220110478,10240110028}@stu.ecnu.edu.cn
202100171091@mail.sdu.edu.cn

Abstract

Represented by Deepseek-R1, “thinking” is generally considered a method to enhance the performance of large language models. Under the CCL 2025 “Zhengming” Chinese language understanding task, this paper explores the potential of both “thinking” and “non-thinking” models. Specifically, for the ACLUE task, we construct a domain-specific knowledge dataset for ancient Chinese, distill a thinking dataset using Deepseek-R1, and curate a high-quality thinking dataset. With these datasets, we perform LoRA fine-tuning and find that although the thinking model achieves significant performance gains, it still underperforms the original non-thinking model. Finally, we open-source and submit the SongPanda model based on Qwen2.5.

Keywords: Large Language Model , Chinese language understanding , Reasoning Model

1 引言

特定领域的中文文学文本（比如古汉语文本），由于其语言的复杂性、多义性及浓厚的文化背景，给大语言模型（LLMs）带来了巨大的挑战。研究者多致力于评测、提高LLMs在

特定领域的表现力。ALP2025，作为NAACL子会议，评测了其古汉语命名实体识别能力(李斌等, 2025)；Chen et al. (2025)综述了LLMs在中医领域的应用，评测了2023-2024年发表的10项中医大语言模型。如果不重新从头训练模型，提高它的表现力，大致通过指令工程（prompt engineering）、微调（finetuning）、检索增强生成（RAG）实现。聚焦到微调上，在这一过程之中，机器如何真正"学习"（如何真正掌握特定领域的知识框架），以及"思考"是否有助于"学习"，这两个问题显得非常关键。

首先，关于"如何学"（即微调如何注入知识）。具体实现的流程，研究者主要模仿OpenAI的范式，预训练（Pre-training）、监督微调（Supervised Fine-Tuning,SFT）、强化学习(Reinforcement Learning,RL)(Long Ouyang et al., 2022)。比如Ye et al. (2024)采用领域持续预训练（Continual Pre-training,CPT）、监督微调（SFT）、直接偏好优化（Direct Preference Optimization,DPO）三阶段训练方法，构建特定数据集，提出了Qilin-Med医疗大语言模型；Lu et al. (2024)探索了CPT、SFT、DPO在材料科学等领域的策略；Yao et al. (2025)在构建wenyanGPT，专为文言文任务设计的大型语言模型时，同样基于LLaMA3-8B-Chinese 进行CPT和SFT。

其次，"思考"在研究上很大程度上被认为是对复杂任务，尤其是推理任务非常有效的。Wei et al. (2022)展示了思维链（CoT）在数学、常识和符号推理任务中的显著提升。在GSM8K数学问题基准测试中，540亿参数的PaLM模型通过仅8个CoT示例即可超越微调后的GPT-3。随后，Deepseek通过强化学习直接训练模型生成CoT，展现出自验证、反思等能力(Deepseek-AI et al., 2024)。然而，最近的研究，如苹果团队则对LLM 的推理能力提出了质疑，并提出DeepSeek-R1、o3这类模型实际上根本没有进行推理，只是很擅长记忆模式罢了(Shojaee et al., 2025)。Li et al. (2024)也同样指出思考会成为一种负担，在很多任务之中，更多的思考不一定带来好的结果，知道何时以及如何思考可能比简单地增加思考量更为重要。

1.1 争鸣评测任务描述

Zhengming（争鸣）作为CCL25-Eval评测任务之一，正是在此背景提出的评测框架。[1]争鸣评测基准提供了关于中国语言理解的七个任务和他们对应的数据集，包括现代文学批评倾向（CritBias）、现代文学批评挖掘（CritPred）、古代文学知识理解（ACLUE）、文学阅读理解（ReadCom）、文学命名实体识别（LitNRE）、文学作品风格预测（AuthIDE）和文学语言风格转换（ClaTrans）。前两个任务是域外测试，后五个则是提供用以微调。按照数据集的文本载体来分类，主要有古代文本（ACLUE、ClaTrans）和现代文本（CritBias、CritPred、ReadCom、LitNRE、AuthIDE）。

研究先基于Qwen2.5-7b-instruct和Deepseek-V3重新测试了baseline、然后又在每个供微调任务的训练数据集中各自随机抽取2k条，总共1w条数据用以SFT、以及先在包含一些基础的国学常识电子书来进行Pre-training，再重复SFT，先得出了以下结果:

Task	Qwen	Deepseek	Qwen-SFT	Qwen-pre1-SFT
古代文学知识理解	0.59	**0.69**	0.65	0.67
文学作品风格预测	0.78	0.86	0.94	**0.96**
文学命名实体识别	0.22	0.42	0.58	**0.59**
现代文学批评挖掘	0.86	**0.87**	0.81	0.83
现代文学批评倾向	0.60	**0.72**	0.62	0.63
文学阅读理解	0.62	0.80	0.88	**0.93**

Table 1: Qwen2.5-7b/Deepseek-V3重新测量的baseline和SFT、Pre-training+SFT初步结果

需要说明的是：1、文学语言风格转换（ClaTrans）任务暂时没有参与评测；2、各项数据集任务的计算方法均与主办方一致，只是为了方便对比，结果指标只用ACC或EM表示，方便对比。3、本实验全部基于llamafactory框架，lora微调，下不再赘述。

从上面的结果指标中，发现：1、域外测试的现代文学批评倾向、现代文学批评挖掘任务结果基本没有提升，或很小幅度提升；2、参与微调的现代文本的任务在Pre-training+SFT之后，

[1]https://github.com/isShayulajiao/CCL25-Eval-ZhengMing

结果指标均上升显著，以Qwen2.5 7b为基座模型微调后的结果，均超过671b的Deepseek；3、古汉语任务（ACLUE）也有所提升，但是依旧小于Deepseek的0.69。

因为古代文学/古文字学知识是本团队的优势，所以我们将研究重点放在“古代文学知识理解”（对应数据集为ACLUE）任务之上。

2 实验方法

2.1 古代文学知识理解任务描述

Ancient Chinese Language Understanding Evaluation (ACLUE) 是一个面向古代汉语的评估基准[2]，旨在帮助评估大型语言模型在古代汉语上的表现。基准由15 个任务组成，分别是：词汇（古文单字多义，通假字，古汉语命名体识别），句法（古文断句），语义（对联，古诗词上下句预测），推理（古诗词质量评估，古文阅读理解，古诗词曲鉴赏，诗词情感分类），知识（古汉语知识，国学常识，医古文，古代文学知识，古音学）。可见，古代文学知识理解任务非常复杂，包含了很多古汉语特定领域知识，而7b小模型的知识数据还是比较少，从前面的研究来看，在微调过程之中，单纯的SFT比较难“注入”这些特定领域知识，因此还需要CPT来提高表现，最初我们试验的时候选择的CPT数据量较少，且没有精心整理，构建合适的古汉语领域知识数据集在这个任务的提高上显得非常关键。

2.2 构建古汉语领域知识库

研究最终整理了一份合适的古汉语领域知识数据集，分别包含三部分知识：古诗词赏析（来自“古诗文网”赏析集）、古文字（来自《说文解字》《汉语大字典》《文字学概要》）、古代文化（来自“中国古代文学史”“音韵学教程”课程笔记、以及王力《中国古代文化常识》）。数据组成比例大致为1:1:1，最终包含11,216,638个token。

3 实验结果分析

基于以上准备，结果如下：

Task	Deepseek	Qwen-pre1-SFT	Qwen-pre2-SFT
古代文学知识理解	0.69	0.67	**0.71**
文学作品风格预测	0.86	**0.96**	0.94
文学命名实体识别	0.42	0.59	**0.66**
现代文学批评挖掘	**0.87**	0.83	0.80
现代文学批评倾向	**0.72**	0.63	0.62
文学阅读理解	0.80	**0.93**	0.92

Table 2: Qwen-pre2-SFT（SongPanda）结果与其他对比

发现：古代文学知识理解有很大的提升（尤其是考虑到Deepseek也只有0.69的准确率），可见在CPT的过程之中，模型确实有效学习到了专业知识。

研究还意外发现现代文的命名实体识别任务也提高很多，可能是学习这些数据集之后模型对文本的理解力也确实提高了不少。最后模型命名为SongPanda，开源在huggingface（ningzhuo/SongPanda）。

3.1 思考还是不思考?

因为古代文学知识理解存在很多的推理任务，我们猜想Cot的方法会在这项任务上有所提升，所以也用了具有reasoning模式的模型评测了这项任务。基于Deepseek-V3，研究蒸馏了随机挑选的8k条train数据，得出了对应包含Cot过程的数据集。模型选择上，首先使用了Deepseek-R1-Distill-Qwen-7B用以微调[3]。在评测任务结束之前，阿里发布了Qwen3(Yang et al., 2025)，作为混合推理的模型，所以我们也将Qwen3-8b纳入评测，结果如下：

[2]https://huggingface.co/datasets/tyouisen/ACLUE
[3]https://huggingface.co/Deepseek-ai/Deepseek-R1-Distill-Qwen-7B

Task	distill	distill-SFT	Qwen3-think	Qwen3-SFT	Qwen3-pre-SFT
古代文学知识理解	0.30	0.44	0.50	0.61	**0.65**

Table 3: 推理模型的古代文学知识理解结果（distill即Deepseek-R1-Distill-Qwen-7B）

可见，虽然准确率有显著提高，依旧不如非思考模型微调后的水平。我们检查了蒸馏后的Cot数据，发现这些合成数据，在思考链上存在各种程度的错误，所以分题型对这些思维链进行重新校对和审核。研究将古代文学知识理解任务重新合并分成以下七类：1、考查古汉语字义；2、考查通假字、古今字，或者形声字、会意字等相关问题；3、古汉语断句；4、诗歌或对联上下句预测；5、古诗词情感鉴赏；6、文言文阅读理解；7、文化常识知识（比如古医学）。对于每种题型，都试图归纳出一种可重复、可理解的思考过程模版。“可理解”是针对机器而言的，比如在判断“诗歌上下句预测”题型，专业的思路应该是首先根据字数（五言、七言），再来根据平仄，但是机器数数的能力太差了，经常会把五个字说出七个字(Yehudai et al., 2024)；同时也不识平仄，所以在撰写思维链的时候尽量回避这些判断标准。

例如，对于“判断通假字”题型，归纳的思路如Figure1所示(陈剑, 2015; 裘锡圭, 2013)：

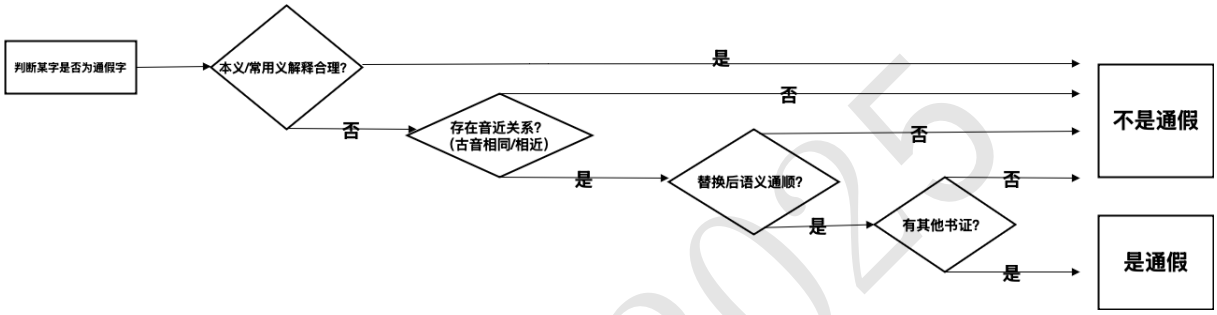


Figure 1: 判断通假字题型思路

受限于时间和精力，研究最后也只整理了三百条Cot数据，用这300条古代文学知识理解数据和其他所有任务（这些任务因为本身题目不难，蒸馏合成的Cot数据质量也比较高）混合微调（每项任务也只筛选300-500条数据，以控制合适的混合比例），结果如下：

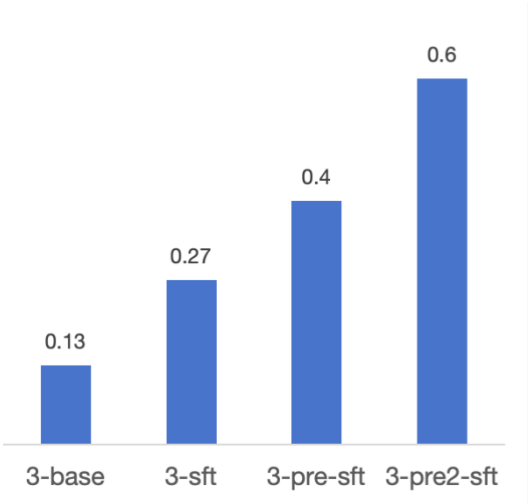
Task	Qwen	Qwen-pre2-SFT	Qwen3-pre2-SFT
古代文学知识理解	0.59	**0.71**	0.62
文学作品风格预测	0.78	**0.94**	0.89
文学命名实体识别	0.22	**0.66**	0.57
现代文学批评挖掘	0.86	0.80	**0.86**
现代文学批评倾向	0.6	0.62	**0.65**
文学阅读理解	0.62	**0.92**	0.77

Table 4: Qwen3微调后的各项结果对比

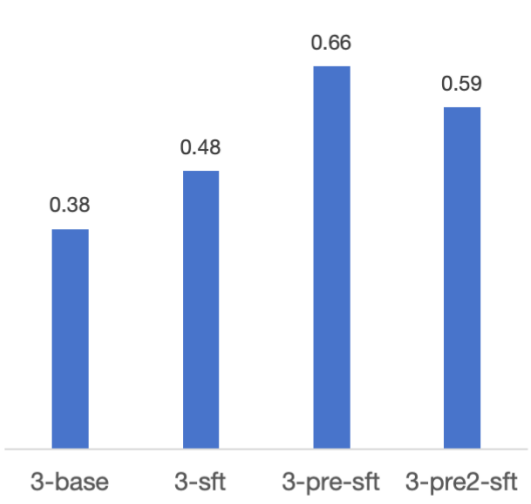
补充说明：“pre2”即是用上文提到整理的资料库进行CPT。

最后的结果还是基本上都比“SongpPanda”低，不过还是观察到了，在思考模式下，之前一直停滞的域外任务，现代文学批评倾向、现代文学批评挖掘二者结果均有所提升。

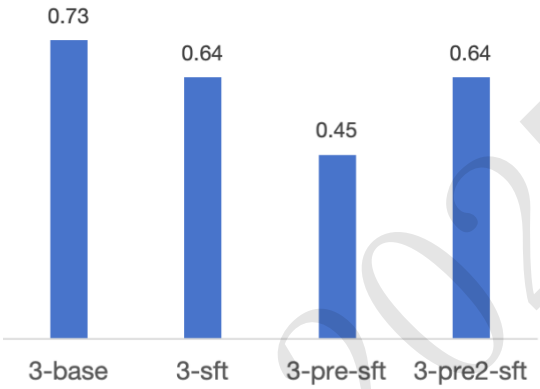
把目光放回古代文学知识理解任务，为了观察在CPT和SFT之后，模型分别发生了什么变化，研究将古代文学知识理解的七类任务利用大模型，把test数据批量打上1-7的tag，分别为：1、考查字词字义；2、考查通假字、古今字，或者形声字、会意字等相关问题；3、古汉语断句；4、诗歌或对联上下句预测；5、古诗词理解与鉴赏；6、文言文阅读理解；7、其他的国学常识。想要观察经过训练之后，哪些题型的准确率上升，哪些下降。见Figure2中结果，发现：



(a) tag2题目正确率变化



(b) tag4题目正确率变化



(c) tag5题目正确率变化

Figure 2: 不同题型准确率变化对比

tag2（判断古汉语选段中，某些字是否是通假字或其他）准确率提高非常迅速，相对的，tag4（上下句判断）、tag5（古诗情感鉴赏）的题目准确率却停滞或者下降。

挑选了test的一道错题（见Figure3），观察具体思考过程：

[ACLUE1833]
“登大坟以远望兮，聊以舒吾忧心”中“坟”义为（）
A、上古典籍 B、洲中坟墓 C、山崖 D、水中高地

Figure 3: 古代文学知识理解错题例之一

这题的答案是D，错误项一般是C，以下是Qwen3-pre-SFT的部分思考过程：“坟，更常用来指土堆或高地，而不是山崖或水中高地。.....所以综合考虑，坟，在这里应该是指山崖，也就是选项C。”在这里Qwen3推理前后自相矛盾，这似乎表明“思考”本身的问题。Anthropic在最新的研究报告中也指出[4]，Claude在某些任务具备长远规划能力，甚至还会为了迎合人类而编造推理过程。从这个推理过程来看，模型推理的前后自我矛盾，应该也是模型能力本身的受限，即便出现aha moment，模型发现了错误，可能也只会重复纠结（因为它一开始就把CD都排除了“不是山崖或水中高地”）。Deepseek-R1对这道题的思考过程则如下：“再查一下具体出

[4]https://transformer-circuits.pub/2025/attribution-graphs/methods.html

处，这句话出自屈原的《九章·哀郢》，原文中的登大坟以远望兮，聊以舒吾忧心中的坟，王逸的《楚辞章句》注释为，水中高者曰坟，也就是水中的高地，所以D选项是正确的。”这种题实际上是思考本身难以解决的，最重要的还是模型本身的知识储备能力。所以看起来，相对于作为形式的“思考”，可能更重要的是“学习”本身。

4 总结

通过对争鸣评测任务的深入研究，我们发现在古汉语理解任务中，传统的知识注入方法（CPT+SFT）仍然是最有效的。虽然“思考”模式在某些领域任务上展现出潜力，但在知识密集型的古汉语理解任务中，扎实的知识基础比复杂的推理过程更为重要。我们最终提交的SongPanda模型通过精心构建的古汉语知识库获得了良好的性能表现，这表明在特定领域任务中，“学习”确实比“思考”更为关键。

5 致谢

本次测试使用的Deepseek-V3是华东师范大学自搭的ecnuplus（地址：https://chat.ecnu.edu.cn/），感谢华师大平台。另外，本实验全部都在“趋动云”提供的算力下训练，感谢赞助支持。最后，也感谢CCL25-Eval-ZhengMing的负责老师、工作人员，以及全体审稿老师的宝贵意见。

参考文献

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388*.

Bin Li, Bolin Chang, Ruilin Liu, Xue Zhao, Si Shen, Lihong Liu, Yan Zhu, Zhixing Xu, Weiguang Qu, and Dongbo Wang. 2025. Overview of EvaHan2025: The First International Evaluation on Ancient Chinese Named Entity Recognition. *In Proceedings of the Second Workshop on Ancient Language Processing*, pages 156–164, The Albuquerque Convention Center, Laguna. Association for Computational Linguistics.

Emmanuel Ameisen, Jack Lindsey, Adam Pearce, Wes Gurnee, Nicholas L. Turner, Brian Chen, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, Andrew Persic, Zhenyi Qi, T. Ben Thompson, Sam Zimmerman, Kelley Rivoire, Thomas Conerly, Chris Olah, and Joshua Batson. 2025. Circuit Tracing: Revealing Computational Graphs in Language Models. *Transformer Circuits Thread*. https://transformer-circuits.pub/2025/attribution-graphs/methods.html

Deepseek-AI et al. 2024. Deepseek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *arXiv preprint arXiv:2501.12948*.

Gilad Yehudai, Haim Kaplan, Asma Ghandeharioun, Mor Geva, and Amir Globerson. 2024. When Can Transformers Count to n? *arXiv preprint arXiv:2407.15160*.

Jason Wei et al. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *arXiv preprint arXiv:2201.11903*.

Parshin Shojaee, Iman Mirzadeh, Keivan Alizadeh, Maxwell Horton, Samy Bengio, and Mehrdad Farajtabar. 2025. The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity. *arXiv preprint arXiv:2506.06941*.

Qichen Ye, Jun Liu, Dingdong Chong, Pengju Zhou, Yining Hua, Fenglin Liu, Meiling Cao, Zhengyi Wang, Xin Cheng, Zhongyu Lei, Zhenhua Guo. 2024. Qilin-Med: Multi-stage knowledge injection advanced medical large language model. *arXiv preprint arXiv:2310.09089*.

第二十四届中国计算语言学大会论文集，第271页-第277页，济南，中国，2025年8月11日至14日。

Qiguang Chen, Libo Qin, Jinhao Liu, Dengyun Peng, Jiannan Guan, Peng Wang, Mengkang Hu, Yuhang Zhou, Te Gao, and Wanxiang Che. 2025. Towards Reasoning Era: A Survey of Long Chain-of-Thought for Reasoning Large Language Models. *arXiv preprint arXiv:2503.09567*.

Wei Lu et al. 2024. Fine-tuning Large Language Models for Domain Adaptation. *arXiv preprint arXiv:2409.03444*.

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

Xiaomin Li, et al. 2024. When thinking fails: The pitfalls of reasoning for instruction-following in LLMs. *arXiv preprint arXiv:2505.11423*.

Xinyu Yao, Mingjun Wang, Bingxuan Chen, Xin Zhao. 2025. WenyanGPT: A Large Language Model for Classical Chinese Tasks.

Zhe Chen, Hui Wang, Chengxian Li, Chunxiang Liu, Fengwen Yang, Dong Zhang, Alice Josephine Fauci, Junhua Zhang. 2025. Large language models in traditional Chinese medicine: a systematic review. *Acupuncture and Herbal Medicine*, 5(1):57–67.

陈剑. 2015. 《释殷墟甲骨文里的“远”“迄”及有关诸字》导读. 载黄天树、沈培、陈剑、郭永秉读解：中西学术名篇精读·裘锡圭卷，中西书局，上海.第255页.

裘锡圭. 2013. 文字学概要. 商务印书馆，北京，第174—197页.