

第七讲 · 数据标注

使用大语言模型进行数据标注与插补

欧阳荣鑫¹

¹ 新加坡国立大学传播与新媒体系, 计算传播, 博士候选人

使用因果和计算方法探索信息通讯技术的政治社会影响
如信息摩擦、账户封禁的巨大影响

Email: rongxin@u.nus.edu

2026 年 2 月

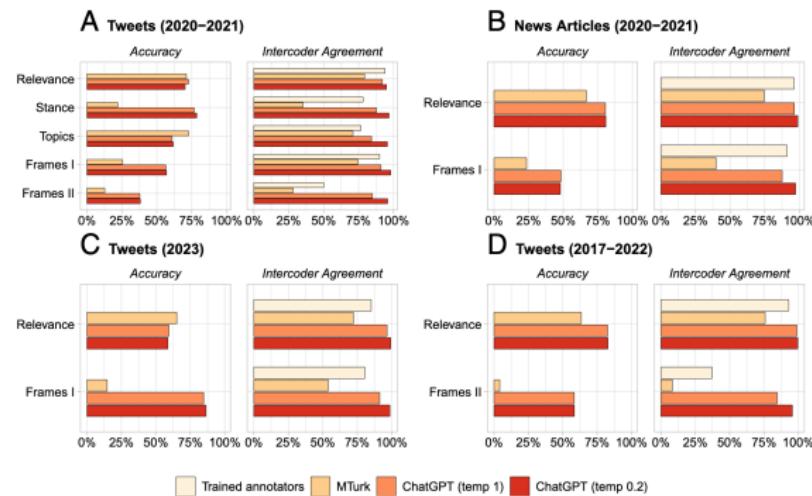
计算社会科学研究 (Lazer et al., 2009) 中,
如何高效地标注大量数据?

引言

- 计算社会科学中的数据包括 “*Spatial data, social networks, and human coding of text and images.*” (Lazer et al., 2020)
- 当数据量巨大时 (正是计算社科的特点之一; 见 Lazer et al., 2009) , 人工标注数据的成本和时间开销曾非常高昂 (Grimmer & Stewart, 2013)
近期自动文本分析进展降低了这些成本 (Barberá et al., 2021) ,
- 幸运的是, 随着转换器模型架构的提出 (Vaswani et al., 2017) , 具备可接受质量的自动文本分析的技术门槛正在进一步降低
- 这些以 GPT 为代表的模型展现出了零样本推理、少样本推理和思维链能力 (Kojima et al., 2023; Törnberg, 2023)
- 因此, 这些模型逐渐被研究中应用于数据标注和缺失值插补任务当中 (见后例)

引言（例）

- PNAS 实验, 大语言模型标注能力超越了人类编码员 (Gilardi et al., 2023)
 - ACL 实验, 麻省理工团队, 展示 LLM 注释给人类提升了报告信心 (Schroeder et al., 2025)



Gilardi et al. (2023)

本期概要

1. 数据标注：情感分析与主题建模 (**tutorial**)
2. 缺失值插补：LLM 预测技术（展示）
3. 提示技术：提高标注质量（展示）
4. 大规模处理：错误处理、进度展示与并行化（展示）
5. 质量评估：标准与方法（模拟实验）
6. 批判性思考：LLM 标注的局限性与风险

引言
○○○○

数据标注
●○

缺值插补
○○○

提示技术
○○○○

规模处理
○○○○○

质量评估
○○○○○

批判思考
○○○

小结练习
○○○

参考文献
○

1. 数据标注

数据标注

生成式语言模型具备强大的能力和多种不同的研究性用途。本节展示生成式语言模型在以下方面的能力：

- **情感分析**: 判断文本的情感极性 (积极/消极/中性)
- **主题建模**: 识别文本中包含的主要主题

核心思路：构建双层提示——系统指令定义分类规则，用户提示输入待分析文本，调用模型进行推理后，直接提取结果。

实验展示

<https://github.com/reycn/llm-annotation-tsinghua>



引言
○○○○

数据标注
○○

缺值插补
●○○

提示技术
○○○○

规模处理
○○○○○

质量评估
○○○○○

批判思考
○○○

小结练习
○○○

参考文献
○

2. 缺失值插补

缺失值插补

LLM 的大规模预训练带来了一个新兴用途——利用其对文本间复杂关系的理解来插补缺失数据 (Ding et al., 2024)。

- 已在推荐系统、社会调查等领域得到研究和使用
- 如果将 LLM 的插补视为表格数据的生成，相关研究展示了语言模型在拟合表格数据上的优越性能 (Borisov et al., 2023)

两种简单易用的插补方法：

- 基于单行的预测：利用 LLM 预测下一个句子的能力
- 基于数据微调的预测：向 LLM 提供示例数据进行微调

实验展示

<https://github.com/reycn/llm-annotation-tsinghua>



引言
○○○○

数据标注
○○○

缺值插补
○○○

提示技术
●○○○

规模处理
○○○○○

质量评估
○○○○○

批判思考
○○○

小结练习
○○○

参考文献
○

3. 提示技术

提示技术

1. **零样本提示** (zero-shot prompting; Kojima et al., 2023) 直接提出任务要求，而不附加示例，考验模型的推理与泛化能力
e.g., ‘请判断以下文本的情感是积极、消极还是中性：{{TEXT}}’
2. **少样本提示** (few-shots; Brown et al., 2020) 通过在输入中提供一小批示例，引导模型学习任务的模式并进行推断
e.g., ‘请判断以下文本的情感是积极、消极还是中性。**例如，我喜欢这个产品’是积极的，‘这个服务很糟糕’是消极的，‘天气还不错’是中性的。现在请判断：** {{TEXT}}’
3. **思维链** (chain-of-thought;s Wei et al., 2022) 提示鼓励模型在回答前进行逐步推理，从而提高复杂问题的准确率
e.g., ‘请判断以下文本的情感是积极、消极还是中性：{{TEXT}}.
Let's think step by step.’

提示技术

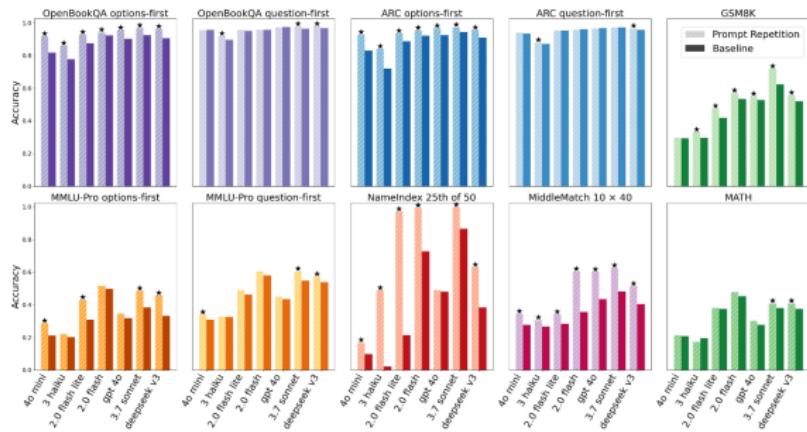
4. 重复 (repetition;

Leviathan et al.,

2025) 重复你的
提示词

e.g., transform
the input from
“<QUERY>” to

“<QUERY><QUERY>”



Leviathan et al. (2025)

提示技术的适用性

- 请注意，与常见的通用性生成模型不同的是，一些推理模型 (reasoning models) 并不推荐用户使用思维链 (CoT) 或少样本提示 (few-shots) 技术。
- e.g., DeepSeek-R1 强调不要对该模型进行少样本提示，因其降低模型表现 (Guo et al., 2025) .

Language mixing

DeepSeek-R1 is at present optimized for Chinese and English, which may result in language-mixing issues when handling queries in other languages. For instance, DeepSeek-R1 might use English for reasoning and responses, even if the query is in a language other than English or Chinese. We aim to address this limitation in future updates. The limitation may be related to the base checkpoint, DeepSeek-V3 Base, which mainly uses Chinese and English, so that it can achieve better results with the two languages in reasoning.

Prompting engineering

When evaluating DeepSeek-R1, we observe that it is sensitive to prompts. Few-shot prompting consistently degrades its performance. Therefore, we recommend that users directly describe the problem and specify the output format using a zero-shot setting for optimal results.

Guo et al. (2025)

引言
○○○○

数据标注
○○

缺值插补
○○○

提示技术
○○○○

规模处理
●○○○○

质量评估
○○○○○

批判思考
○○○

小结练习
○○○

参考文献
○

4. 大规模数据处理

大规模处理的三个技巧

将 LLM 标注技术应用到大规模分析中，需要关注：

1. 错误处理：网络断开、服务过载、模型出错时，应捕捉错误并跳过对应任务，避免整个流程失败
2. 进度展示：当数据量较大时，使用 `tqdm` 进度条观察进度和预估时间
3. 并行化：使用 `pandarallel` 并行化标注过程，充分利用模型服务的吞吐能力

大规模处理的三个技巧: 进度展示

2. 使用 `tqdm` 库为循环添加进度条, 方便监控处理进度。

```
from tqdm import tqdm
for i in tqdm(range(10000)):
    ...
```



A screenshot of a Jupyter Notebook cell. The code imports `tqdm` and uses it to iterate over a range of 10,000. Below the code, there is a progress bar indicating 76% completion. To the right of the progress bar, the output shows the current iteration count (7568), the total number of iterations (10000), and the processing time (00:33<00:10, 229.00it/s).

76% | ██████████ | 7568/10000 [00:33<00:10, 229.00it/s]

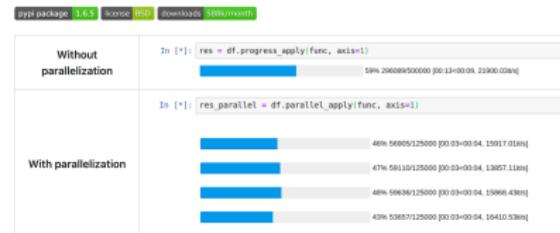
Src. GitHub

大规模处理的三个技巧: 并行化

3. `pandarallel` 可以并行化 pandas 数据框操作, 提升处理速度。

```
1 from pandarallel import
2     pandarallel
3 pandarallel.initialize(
4         progress_bar=True)
5 df.parallel_apply(func)
```

Pandarallel



Src. GitHub

实验展示

<https://github.com/reycn/llm-annotation-tsinghua>



引言
○○○○

数据标注
○○○

缺值插补
○○○

提示技术
○○○○

规模处理
○○○○○

质量评估
●○○○○

批判思考
○○○

小结练习
○○○

参考文献
○

5. 质量评估

评估标准

如何信任 LLM 的标注？我们或许需要对生成式人工智能的结果进行特定场景的再评估。评估受限要解决的是，如何衡量标注结果的质量？主要有两类标准：

1. 是否与事实一致：将标注结果与已知事实对比

- 准确度 (accuracy)、精确度 (precision)、召回率 (recall)
- F1-score、AUROC 等
- 可以参见 Raschka and Mirjalili (2019) 中的相关介绍

2. 是否不同的编码员（模型）一致：衡量不同标注之间的一致性

- 百分比一致性：两个编码员间的简单一致比例
- Krippendorff's α (dubbed " $\alpha - \text{AgreementforCoding}$ "; Krippendorff, 2018)：也适用与多分类多编码员
- 更多的统计量和不同情况的适用性可参见 周翔 (2014), Krippendorff (2018)

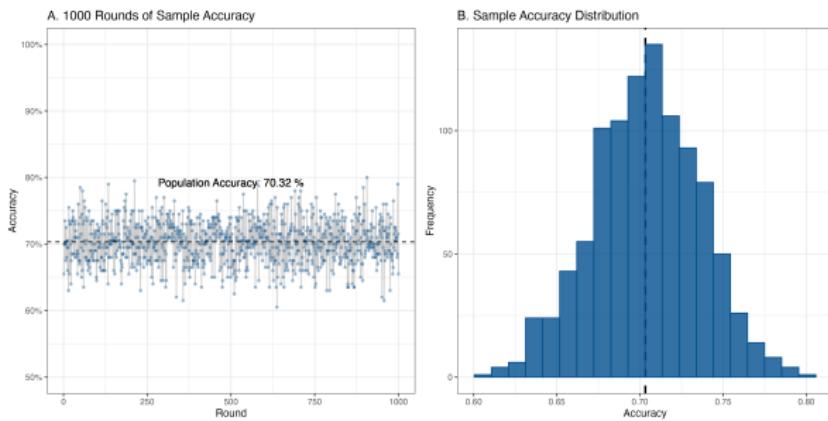
评估方法

如何对大规模标注进行快速、准确的评估？

1. **人类-模型协作评估** (Wang et al., 2024)：将人类辅助纳入验证过程，对低质量标注进行人工重验证
2. **基于众包的人类验证**：使用众包平台（如 MTurk）实现评估规模化
 - 优势：速度快
 - 劣势：成本控制与参与者质量的担忧
3. **抽样验证** (Duan et al., 2025)：通过抽样评估性能，适用于更大规模数据
 - 潜在假设：样本准确度是总体准确度的良好估计

模拟实验：样本准确度是总体准确度的良好估计

- 总体 $N = 10000$, 准确度 $\approx 70.32\%$
- 1000 次放回随机抽样, 每次 $n = 200$
- 面板 A: 每次抽样准确度在总体准确度附近波动
- 面板 B: 样本准确度近似正态分布, 以总体准确度为均值



样本准确度是总体准确度的良好估计

实验展示

<https://github.com/reycn/llm-annotation-tsinghua>



引言
○○○○

数据标注
○○

缺值插补
○○○

提示技术
○○○○

规模处理
○○○○○

质量评估
○○○○○

批判思考
●○○

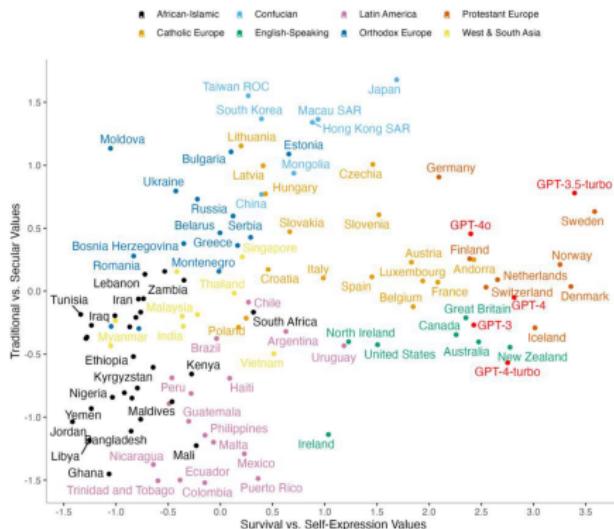
小结练习
○○○

参考文献
○

6. 批判性思考

批判性思考：风险示例

- 在美国，许多重要的政治问题上，大语言模型似乎更加自由、年轻且受过更多教育
(Santurkar et al., 2023)
- 全球范围内，大语言模型表现出文化偏见：“所有模型都展现出与英语国家和新教欧洲国家相似的文化价值观。” (Tao et al., 2024)



Tao et al. (2024)

批判性思考

- 如果现阶段的人工智能尚未达到人类的智力水平，那么将人类能够处理的任务给 AI 处理的最佳方向是？
 - A. 简单任务复杂化
 - B. 复杂任务简单化
- LLM 标注的潜在风险：
 - 模型本身可能存在系统性偏差（bias）
 - 不同模型、不同版本的表现可能不一致
 - 对于需要深层领域知识的任务，LLM 可能不够可靠
 - 评估方法本身的合理性需要审慎考量
- 建议：根据具体任务场景进行特定场景的再评估，而非盲目信任已有实验结论

引言
○○○○

数据标注
○○

缺值插补
○○○

提示技术
○○○○

规模处理
○○○○○

质量评估
○○○○○

批判思考
○○○

小结练习
●○○

参考文献
○

小结

本章小结

- **数据标注**: LLM 在情感分析、主题分类中的应用
- **缺失值插补**: 基于单行预测与微调的两种方法
- **提示技术**: 零样本、少样本、思维链、重复提示
- **大规模处理**: 错误处理、进度展示与并行化
- **质量评估**: 准确度、编码员间信度 (如 Krippendorff's α)
- **批判性思考**: LLM 标注的局限性与风险

课后练习

- 要求：构建一个提示，将输入的新闻文本分类为“政治性新闻”，“娱乐性新闻”，或“其它新闻”
- 可以使用零样本提示、少样本提示或思维链提示等技术
- * Mukerjee et al. (2023) 和 Ouyang et al. (2026) 论文设计思路, 主题分类结果纳入统计分析。

The image displays two side-by-side snippets from the International Journal of Press/Politics. Both snippets are titled "How Social Media Metrics Shape News Production: A Replication Study".
Top Snippet (Left):
Journal of Communication, 2023, 73, 269–272
<https://doi.org/10.1111/jcom.13402>
Advance access publication 22 April 2023
Sage Publications
Bottom Snippet (Right):
The International Journal of Press/Politics
1–18
© The Author(s) 2023
Article reuse guidelines:
http://jpp.sagepub.com/page/author_info/reuse.xhtml
DOI: 10.1177/104113232311628
[journals.sagepub.com](http://jpp.sagepub.com)
Sage

Rongxin Ouyang¹, Subhayan Mukerjee^{1,2,3},
and Tian Yang⁴

Metrics in action: how social media metrics shape news production on Facebook

Subhayan Mukerjee¹, Tian Yang², Yilang Peng^{3,*}

¹Department of Communications and New Media, National University of Singapore, Singapore
²School of Journalism and Communication, The Chinese University of Hong Kong, Hong Kong, China
³Department of Economics, University of Georgia, Athens, Georgia, USA
⁴*Corresponding author: Yilang Peng. Email: yilang peng@uga.edu

Abstract
Social media metrics allow media outlets to get a granular, real-time understanding of audience preferences, and news platforms can use this to decide what content to prioritize in the future. We test this mechanism in the context of Facebook, by comparing traditional and long-tail media news coverage. We find that while the overall effect of audience engagement on future news coverage is significant, there is substantial heterogeneity in how individual outlets respond to different kinds of topics. A handful of light-weight media outlets are more likely to respond to audience engagement metrics than others. This heterogeneity suggests that news platforms have different sensitivities to audience engagement. We also find that how social media platforms have shaped journalistic practices and has implications for the future health of journalism in the United States.

Keywords: audience engagement, Facebook metrics, news producers, print media, social media

Mukerjee et al. (2023), Ouyang et al. (2026)

参考文献 |

- Barberá, P., Boydston, A. E., Linn, S., McMahon, R., & Nagler, J. (2021). Automated text classification of news articles: A practical guide. *Political Analysis*, 29(1), 19–42. <https://doi.org/10.1017/pan.2020.8>
- Borisov, V., Seßler, K., Leemann, T., Pawelczyk, M., & Kasneci, G. (2023, April 22). *Language Models are Realistic Tabular Data Generators*. arXiv: 2210.06280 [cs]. <https://doi.org/10.48550/arXiv.2210.06280>
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., & Askell, A. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901.
- Ding, Z., Tian, J., Wang, Z., Zhao, J., & Li, S. (2024, August 7). *Data Imputation using Large Language Model to Accelerate Recommendation System*. arXiv: 2407.10078 [cs]. <https://doi.org/10.48550/arXiv.2407.10078>
- Duan, Z., Shao, A., Hu, Y., Lee, H., Liao, X., Suh, Y. J., Kim, J., Yang, K.-C., Chen, K., & Yang, S. (2025). Constructing Vec-tionaries to extract message features from texts: A case study of moral content. *Political Analysis*, 1–21. <https://doi.org/10.1017/pan.2025.6>
- Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT outperforms crowd-workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30), e2305016120. <https://doi.org/10.1073/pnas.2305016120>
- Grimmer, J., & Stewart, B. M. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, 21(3), 267–297. <https://doi.org/10.1093/pan/mps028>
- Guo, D., Yang, D., Zhang, H., Song, J., Wang, P., Zhu, Q., Xu, R., Zhang, R., Ma, S., Bi, X., Zhang, X., Yu, X., Wu, Y., Wu, Z. F., Gou, Z., Shao, Z., Li, Z., Gao, Z., Liu, A., ... Zhang, Z. (2025). DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081), 633–638. <https://doi.org/10.1038/s41586-025-09422-z>
- Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2023, January 29). *Large Language Models are Zero-Shot Reasoners*. arXiv: 2205.11916 [cs]. <https://doi.org/10.48550/arXiv.2205.11916>
- Krippendorff, K. (2018). *Content analysis: An introduction to its methodology*. SAGE Publications. Retrieved February 9, 2026, from <https://books.google.com.sg/books?id=nE1aDwAAQBAJ>
- Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., & Gutmann, M. (2009). Computational Social Science. *Science*, 323(5915), 721–723. <https://doi.org/10.1126/science.1167742>

参考文献 II

- Lazer, D., Pentland, A., Watts, D. J., Aral, S., Athey, S., Contractor, N., Freelon, D., Gonzalez-Bailon, S., King, G., Margetts, H., Nelson, A., Salganik, M. J., Strohmaier, M., Vespignani, A., & Wagner, C. (2020). Computational social science: Obstacles and opportunities. *Science*, 369(6507), 1060–1062. <https://doi.org/10.1126/science.aaz8170>
- Leviathan, Y., Kalman, M., & Matias, Y. (2025, December 17). *Prompt Repetition Improves Non-Reasoning LLMs*. arXiv: 2512.14982 [cs]. <https://doi.org/10.48550/arXiv.2512.14982>
- Mukerjee, S., Yang, T., & Peng, Y. (2023). Metrics in action: How social media metrics shape news production on Facebook. *Journal of Communication*, 260–272. <https://doi.org/10.1093/joc/jqad012>
- Ouyang, R., Mukerjee, S., & Yang, T. (2026). How Social Media Metrics Shape News Production: A Replication Study. *The International Journal of Press/Politics*, 19401612251411639. <https://doi.org/10.1177/19401612251411639>
- Raschka, S., & Mirjalili, V. (2019). *Python machine learning: Machine learning and deep learning with Python, scikit-learn, and TensorFlow 2*. Packt publishing ltd. Retrieved February 9, 2026, from <https://books.google.com.sg/books?id=sKXIDwAAQBAJ>
- Santurkar, S., Durmus, E., Ladha, F., Lee, C., Liang, P., & Hashimoto, T. (2023). Whose opinions do language models reflect? *Proceedings of the 40th International Conference on Machine Learning*.
- Schroeder, H., Roy, D., & Kabbara, J. (2025). Just Put a Human in the Loop? Investigating LLM-Assisted Annotation for Subjective Tasks. *Findings of the Association for Computational Linguistics: ACL 2025*, 25771–25795. Retrieved February 8, 2026, from <https://aclanthology.org/2025.findings-acl.1323/>
- Tao, Y., Viberg, O., Baker, R. S., & Kizilcec, R. F. (2024). Cultural bias and cultural alignment of large language models. *PNAS Nexus*, 3(9), pgae346. <https://doi.org/10.1093/pnasnexus/pgae346>
- Törnberg, P. (2023, April 13). *ChatGPT-4 Outperforms Experts and Crowd Workers in Annotating Political Twitter Messages with Zero-Shot Learning*. arXiv: 2304.06588 [cs]. <https://doi.org/10.48550/arXiv.2304.06588>

参考文献 III

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł.,ukasz, & Polosukhin, I. (2017). Attention is All you Need. *Advances in Neural Information Processing Systems*, 30. Retrieved February 8, 2026, from https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fdbd053c1c4a845aa-Abstract.html

Wang, X., Kim, H., Rahman, S., Mitra, K., & Miao, Z. (2024). Human-LLM Collaborative Annotation Through Effective Verification of LLM Labels. *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–21. <https://doi.org/10.1145/3613904.3641960>

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q. V., & Zhou, D. (2022). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Advances in Neural Information Processing Systems*, 35, 24824–24837. Retrieved April 15, 2025, from https://proceedings.neurips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html?ref=https://githubhelp.com

周翔. (2014). 传播学内容分析研究与应用. 重庆大学出版社. Retrieved February 9, 2026, from <https://www.cqup.com.cn/index.php?m=content&a=show&catid=16&id=12415>