Lecture 7: Data Annotation
Data Annotation and Imputation Using Large Language Models

Rongxin Ouyang[1]

[1] Department of Communications and New Media, National University of Singapore
PhD Candidate in Computational Communication
Exploring political and social impacts of information and communication technology using causal and computational methods
Email: rongxin@u.nus.edu

February 2026

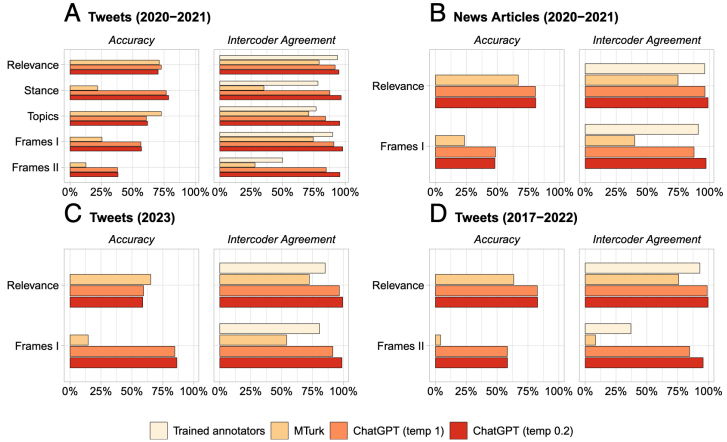
In computational social science research (Lazer et al., 2009) ,
how can we efficiently annotate large amounts of data?

Introduction

- Data in CSS includes “Spatial data, social networks, and **human coding of text** and images.” (Lazer et al., 2020)
- When data volume is large – which is one of the key characteristics of CSS (Lazer et al., 2009) – the cost and time overhead of manual annotation have been very high (Grimmer & Stewart, 2013) . Recent advances in automated text analysis have reduced these costs (Barberá et al., 2021) .
- Fortunately, with the proposal of the Transformer architecture (Vaswani et al., 2017) , the technical threshold for acceptable-quality automated text analysis is further lowering.
- Models exemplified by GPT demonstrate zero-shot reasoning, few-shot reasoning, and chain-of-thought capabilities (Kojima et al., 2023; Törnberg, 2023)
- Therefore, applied in annotation and imputation (examples below)

Introduction (Examples)

- PNAS experiment: Large language models outperform human annotators (Gilardi et al., 2023)
- ACL experiment by MIT team: LLM annotations boost human confidence in assessments (Schroeder et al., 2025)



Lecture Overview

1. **Data Annotation: Sentiment Analysis and Topic Modeling (tutorial)**
2. Missing Value Imputation: LLM Prediction Techniques (demonstration)
3. Prompting Techniques: Improving Annotation Quality (demonstration)
4. Large-Scale Processing: Error Handling, Progress Tracking, and Parallelization (demonstration)
5. **Quality Evaluation: Standards and Methods (simulation)**
6. **Critical Thinking: Limitations and Risks of LLM Annotation**

1. Data Annotation

Data Annotation

Generative language models possess powerful capabilities and diverse research applications. This section demonstrates generative language models' abilities in the following domains:

- **Sentiment Analysis**: Determining the sentiment polarity of text (positive/negative/neutral)
- **Topic Modeling**: Identifying the main topics contained in text

Core approach: Create a two-layer prompt – system instructions define classification rules, user prompts input the text to be analyzed, invoke the model for reasoning, then directly extract results.

Demonstration

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

2. Missing Value Imputation

Missing Value Imputation

Large-scale pretraining of LLMs brings an emerging use case – leveraging their understanding of complex relationships between texts to impute missing data (Ding et al., 2024) .

- Has been researched and applied in recommendation systems, social surveys, and other domains
- When viewing LLM imputation as table data generation, related research shows superior performance of language models in fitting tabular data (Borisov et al., 2023)

Two simple and practical imputation methods:

1. **Row-based prediction**: Leveraging LLM's ability to predict the next sentence
2. **Fine-tuning-based prediction**: Providing example data to LLM for fine-tuning

Demonstration

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

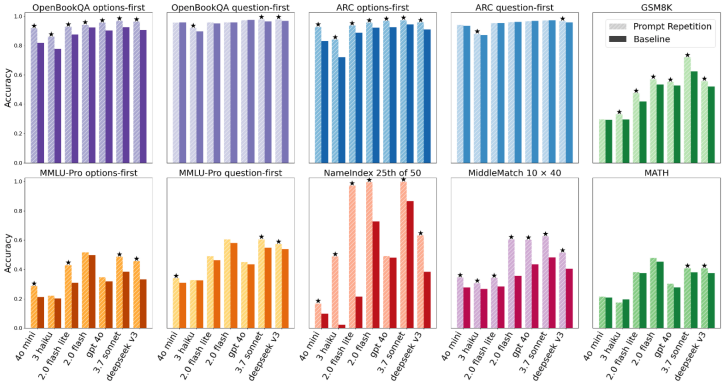
3. Prompting Techniques

Prompting Techniques

1. **Zero-shot prompting** (Kojima et al., 2023) : Directly stating task requirements without examples, testing the model's reasoning and generalization ability
   e.g., "Determine whether the following text is pos., neg., or neu.: {{TEXT}}"
2. **Few-shot prompting** (Brown et al., 2020) : Providing a small batch of examples in the input to guide the model in learning task patterns and making inferences
   e.g., "Determine whether the following text sentiment is positive, negative, or neutral. For example, xxx is positive, etc. Now determine: {{TEXT}}"
3. **Chain-of-thought** (Wei et al., 2022) : Prompting that encourages the model to perform step-by-step reasoning before answering
   e.g., "Determine whether the following text sentiment is positive, negative, or neutral: {{TEXT}}. Let's think step by step."

Prompting Techniques

4. **Repetition** (repetition; Leviathan et al., 2025) :
   Repeat your prompt
   e.g., transform the input from "<QUERY>" to "<QUERY><QUERY>"



Leviathan et al. (2025)

Applicability of Prompting Techniques

- Note that unlike common generative models, some reasoning models do not recommend users employing chain-of-thought (CoT) or few-shot prompting techniques.
- e.g., DeepSeek-R1 emphasizes not using few-shot prompting on this model as it reduces model performance (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)

4. Large-Scale Data Processing

Three Tricks for Large-Scale Processing

When applying LLM annotation techniques to large-scale analysis, pay attention to:

1. **Error handling**: When network drops, service overloads, or model errors occur, catch errors and skip corresponding tasks to avoid entire pipeline failure
2. **Progress tracking**: With large data volumes, use `tqdm` progress bars to monitor progress and estimate time
3. **Parallelization**: Use `pandarallel` to parallelize the annotation process and fully leverage model service throughput

Three Tricks for Large-Scale Processing: Progress Tracking

2. Use the `tqdm` library to add progress bars to loops for easy progress monitoring.

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

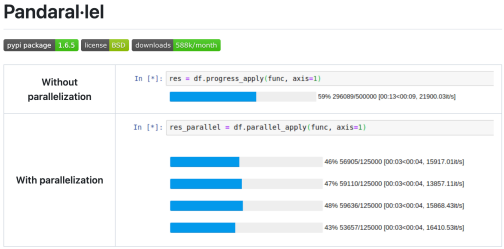
```
76%|                                   | 7568/10000 [00:33<00:10, 229.00it/s]
```

Src. GitHub

Three Tricks for Large-Scale Processing: Parallelization

3. `pandarallel` can parallelize `pandas` dataframe operations, improving processing speed.

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



Src. GitHub

Demonstration

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

5. Quality Evaluation

Evaluation Standards

To what extent we can trust LLM annotations? Two criteria for evaluation:

1. **Consistency with facts**: Comparing annotation results with known facts
   - Accuracy, Precision, Recall
   - F1-score, AUROC, etc.
   - See relevant discussions in Raschka and Mirjalili (2019)
2. **Consistency among different annotators (models)**: Measuring agreement among different annotations
   - Percentage agreement: Simple agreement proportion between two annotators
   - Krippendorff's α (dubbed "$\alpha - AgreementforCoding''$; Krippendorff, 2018) : Also applicable to multi-class multi-annotator cases
   - More statistics and applicability across different scenarios can be found in 周翔 (2014), Krippendorff (2018)

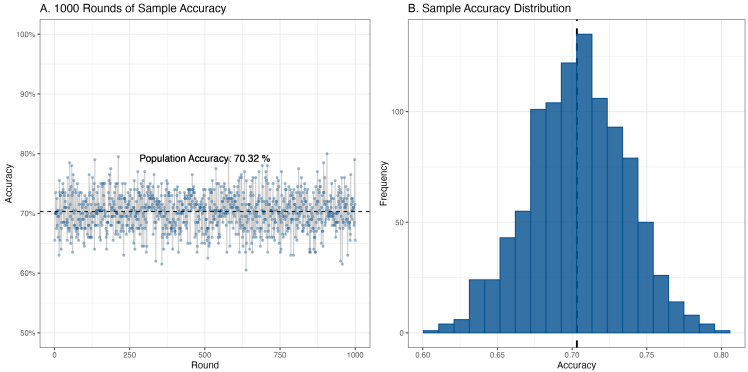
Evaluation Methods

How to quickly and accurately evaluate large-scale annotations?

1. **Human-Model Collaborative Evaluation** (Wang et al., 2024) : Incorporating human assistance into the validation process, with manual re-verification for low-quality annotations
2. **Crowdsourced Human Validation**: Using crowdsourcing platforms (e.g., MTurk) to scale evaluation
   - Advantages: Fast
   - Disadvantages: Cost control and concerns about participant quality
3. **Sampling Validation** (Duan et al., 2025) : Evaluating performance through sampling, applicable to larger-scale data
   - Underlying assumption: Sample accuracy is a good estimate of population accuracy

Simulation Experiment: Sample Accuracy as a Good Estimate of Population Accuracy

- Population $N = 10000$, accuracy $\approx 70.32\%$
- 1000 rounds of random sampling with replacement, $n = 200$ per round
- Panel A: Sample accuracy fluctuates around population accuracy
- Panel B: Sample accuracy approximately follows



Sample Accuracy as a Good Estimate of Population Accuracy

Demonstration

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

6. Critical Thinking

Critical Thinking: Risk Examples

- In the United States, on many important political issues, large language models appear more liberal, younger, and better educated (Santurkar et al., 2023)
- Globally, large language models exhibit cultural biases: "All models demonstrate cultural values similar to English-speaking countries and Protestant Europe." (Tao et al., 2024)

Critical Thinking

- If current artificial intelligence has not yet reached human-level intelligence, what is the best direction for assigning tasks humans can handle to AI?
  - A. Complexifying simple tasks
  - B. **Simplifying complex tasks**
- Potential risks of LLM annotation:
  - The model itself may have systematic biases
  - Different models or versions may have inconsistent performance
  - For tasks requiring deep domain knowledge, LLMs may not be reliable enough
  - The rationality of evaluation methods themselves requires careful consideration
- Recommendation: Conduct **scenario-specific re-evaluation** based on the specific task context, rather than blindly trusting existing experimental conclusions

Summary

Chapter Summary

- **Data Annotation**: Application of LLM in sentiment analysis and topic classification
- **Missing Value Imputation**: Two methods based on row-based prediction and fine-tuning
- **Prompting Techniques**: Zero-shot, few-shot, chain-of-thought, and repetition prompting
- **Large-Scale Processing**: Error handling, progress tracking, and parallelization
- **Quality Evaluation**: Accuracy and inter-annotator agreement (e.g., Krippendorff's α)
- **Critical Thinking**: Limitations and risks of LLM annotation

Homework Exercise

- **Task:** Construct a prompt to classify input news text into “Political News”, “Entertainment News”, or “Other News”
- You may use zero-shot prompting, few-shot prompting, or chain-of-thought prompting techniques
- ** See design approaches in papers by Mukerjee et al. (2023) and Ouyang et al. (2026); incorporate topic classification results into statistical analysis.*

Replication Study

The International Journal of Press/Politics
1–19
© The Author(s) 2026
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/19401612251414199
journals.sagepub.com/home/hij

Sage

How Social Media Metrics Shape News Production: A Replication Study

Rongxin Ouyang[1], Subhayan Mukerjee[1,2,3], and Tian Yang[4]

Journal of Communication, 2023, 73, 260–272
https://doi.org/10.1093/joc/jqad012
Advance access publication 22 April 2023
Original Article

International Communication Association

OXFORD

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

Subhayan Mukerjee[1], **Tian Yang**[2], **Yilang Peng**[3,*]

[1]Department of Communications and New Media, National University of Singapore, Singapore
[2]School of Journalism and Communication, , The Chinese University of Hong Kong, Hong Kong, China
[3]Department of Financial Planning, Housing and Consumer Economics, University of Georgia, Athens, Georgia, USA
*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 may therefore be used to decide what content to prioritize in the future. We test this mechanism in the context of Facebook, by using topic modeling and longitudinal data analysis on a large dataset comprising all posts published by major media outlets used by American citizens in 2016–2019. 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 right-wing media outlets are more likely to respond to audience engagement metrics than other outlets, but with partisan politics topics and not with entertainment-oriented content. Our research sheds new light on 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 production, partisan media, social media

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

References I

Barberá, P., Boydstun, 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

References II

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

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., Ladhak, 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

References 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/3f5ee243547dee91fbd053c1c4a845aa-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