# Practical Strategies for Labeling Qualitative Data Using Large Language Models

Marianne Aubin Le Quéré
msa258@cornell.edu
Cornell Tech
New York, New York, USA

Travis Lloyd
tgl33@cornell.edu
Cornell Tech
New York, New York, USA

Madiha Zarah Choksi
mc2376@cornell.edu
Cornell Tech
New York, New York, USA

## ABSTRACT

This extended abstract demonstrates the practical steps required to use large language models (LLMs) to label qualitative text data for HCI research. We present a case study where we used LLMs to label posts about community surveillance made to Nextdoor, a local social media platform. We validate prior work that shows that LLMs can be a feasible way to label qualitative text data, though we note that interrater agreement remains higher between human coders than with the model. We contribute to setting research norms for the community by providing a practical set of recommendations, based on our experience, for deploying these methods to label qualitative text data. Our key considerations include centering user privacy, tactics to balance cost against performance, creating a gold standard dataset of human ratings to test LLM performance, and strategies to improve low-performing codes.

**ACM Reference Format:**
Marianne Aubin Le Quéré, Travis Lloyd, and Madiha Zarah Choksi. 2024. Practical Strategies for Labeling Qualitative Data Using Large Language Models. In . ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

## 1 INTRODUCTION

HCI research has often required manual labeling of data to enable analysis and broad insights. Various forms of content analysis, including news framing analysis, critical discourse analysis, and thematic coding, rely on human judgment to assign labels to a text or document [1, 2, 4, 11, 13]. For example, content analysis has been used to understand how people talk about depression on Twitter [3], analyze European media news frames [14], and code strategies for responding to violent comments online [18]. It has been argued that these types of complex labels can only be assigned by humans [19]. Sometimes, however, manual labeling of data is not sufficient to achieve the goals of a research study, and researchers have explored methods to scale up content analysis. In the era of big data, datasets can be so vast that manual labeling does not sufficiently represent the dataset. Manual coding of qualitative data is also time consuming, and the labor of experts can be expensive. To tackle these problems, researchers have employed crowdworkers to label text at scale [8, 15].

Numerous studies are beginning to showcase that large language models (LLMs) can also provide adequate labels for qualitative data [16]. For example, emerging work demonstrates the application of these methods in the field of communication, such as detecting news outlet credibility and media framings [17, 20]. Some studies have found that using Chat-GPT to annotate text can "outperform" crowdworkers at a significantly lower cost [7]. Researchers are also increasingly developing tools and recommendations to facilitate qualitative labeling of text data with LLMs [5, 6].

In this extended abstract, we address a gap in the literature by centering practical recommendations for how to use an LLM to label qualitative text data. We document our codebook design process and report on the accuracy of the LLM compared with human raters. We synthesize insights into recommendations for researchers who plan to adopt this method in the future.

## 2 METHODS

The case study we present is an analysis of social media posts made to Nextdoor, a local social media platform. Our research goal was to identify a typology for community surveillance posts made to the platform, which we defined as posts that relate to the enforcement of community norms, surveillance, safety, or crime. Our corpus for analysis was a set of 2,019 carefully anonymized Nextdoor posts made in one US city likely to be relevant to community surveillance. To achieve our analysis goal, we first collaboratively developed and iterated on a codebook. We evaluated GPT-4's [12] performance against the codebook using inter-rater reliability (IRR). Our methodology for tagging posts with GPT-4 can roughly be broken out into a feasibility stage and a refining stage.

In the feasibility stage, we tested the basic feasibility of using LLMs to scale up our qualitative coding task. Four researchers constructed an initial codebook, inspired by López and Butler [9] for tagging posts made to local Facebook groups. First, two researchers manually labeled a set of 320 posts according to the codebook. We then evaluated the performance of four LLMs (GPT-4, GPT-3.5, LLAMA2-7B, GPT4AllFalcon) against the hand-labeled posts. At this stage, GPT-4 was the only model that seemed reasonably aligned with our first set of manual labels. The other models were not viable since they often returned codes that were not in the requested format, or hallucinated codes we had not provided. We also conducted prompt engineering where we ran multiple versions of the codebook and compared the outcomes to the manual labels. Through this process we found that setting the model temperature to 0 allowed for reproducibility, adding more context that these were NextDoor posts improved model performance, and we could

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
*Conference'17, July 2017, Washington, DC, USA*
© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

**Table 1: Codebook, code descriptions, and inter-rater agreement measured using unweighted Cohen's Kappa for all categories for 100 community surveillance posts. Human rater 1 and Human rater 2 are labeled as H1 and H2, and GPT-4's performance is labeled as AI.**

| Category | Code | Human 1 x Human 2 $\kappa$ | Human 1 x AI $\kappa$ | Human 2 x AI $\kappa$ | Avg AI $\kappa$ |
|---|---|---|---|---|---|
| **Main post topic** The central matter of interest in a post. | police activity, guns or gunshots, property damage, noise, criminal or suspicious person, theft, unsafe driving, sexual violence or harassment, inanimate object or animal, not applicable | 0.95 | 0.79 | 0.82 | 0.81 |
| **Roles** A function assumed by a person posting on the Nextdoor platform. | community member, administrator, organizer, or moderator, not applicable | 1 | 0.65 | 0.65 | 0.65 |
| **Providing Information** Any post that gives new, potentially beneficial, information to the reader. | yes, no | 0.93 | 0.49 | 0.57 | 0.53 |
| **Explicitly calling for vigilance** A post that explicitly asks people to be cautious, watchful, or be on the lookout for someone | yes, no | 0.96 | 0.66 | 0.70 | 0.68 |
| **Describing a personal experience** Anyone who is sharing an event they witnessed first-hand or that happened to them. | yes, no | 0.91 | 0.81 | 0.77 | 0.79 |
| **Expressing a personal opinion** Someone sharing their viewpoint on something without being prompted or with the intent to convince others. | yes, no | 0.85 | 0.71 | 0.68 | 0.70 |
| **Soliciting information or action** Explicit requests for something, e.g. pictures of an event. | yes, no | 0.94 | 0.73 | 0.73 | 0.73 |
| **Object** The main thing or person that is discussed in a post. | person, group of people, not applicable | 0.97 | 0.26 | 0.25 | 0.26 |
| **Physical Description** The post describes how a person looks e.g. their race, age, or gender. | yes, no | 0.80 | 0.80 | 0.68 | 0.72 |
| **Primary Sentiment** The main emotion a post is likely to evoke in the reader. | positive, neutral, negative | 0.88 | 0.74 | 0.67 | 0.71 |

request all codes at once for a post without sacrificing accuracy. Three members of the research team then coded 100 posts each: 50 posts were overlapping between all three researchers, 50 posts differed. Using crude agreement, we confirmed that GPT-4 coded approximately as well as the three human coders.

We then moved to the refining stage, where we iterated on the codebook design and gradually validated the inter-human IRR as well as the AI-human IRR. We first focused on improving the categories where either the AI-human IRR or the human IRR was lower than 0.6. In particular, for each category, we looked at whether there were specific codes that were frequently mislabeled. Largely, this process required either simplifying the codebook or making it more precise, for example by including more binary categories instead of multi-label categories. More abstract codes like "surveillance" were frequently mislabeled by both the human and AI raters, and were removed or redefined. Occasionally, we had to make edits to the LLM prompt that were not necessary in the human codebook: for example, the human raters easily agreed on what should be deemed "explicitly calling for vigilance," but for GPT-4 to have high agreement in that category, our instructions for GPT-4 define the vigilance code as "explicitly asking people to be cautious or alert."

## 3 RESULTS

In our study, a majority of the labels assigned by GPT-4 yielded a high enough inter-rater reliability (IRR) with the human labels to proceed with further analysis. Nonetheless, the IRR between GPT-4 and human coders was worse than the IRR between human coders across all categories. Using standard measures of agreement for Cohen's Kappa [10], GPT-4 reached an IRR that was almost perfect or substantial (0.81-1; 0.61-0.8) for eight categories, and moderate or fair (0.41-0.6; 0.21-0.4) for three categories. The categories where the AI-human IRR were lower were the "community surveillance post" category (yes/no), the "providing information" category (yes/no), and the "object" category (person, group of people, not applicable). The full results are shown in Table 1.

## 4 RECOMMENDATIONS

Prior research has demonstrated the feasibility of using LLMs for data labeling, but has not provided practical guidance on how to best apply this method. We provide recommendations based on our own experience that we believe will benefit future researchers.

**Privacy first**: When working with user-generated content we must be aware that this data may contain sensitive information, such as personal identifying information (PII). When possible, we recommend using an open-source LLM that you can run on your own host, as this eliminates the need to pass user data to a third party. If this is not possible, data should be scrubbed of all PII before passed to the third party LLM. When using a third party LLM we recommend contacting the LLM service provider about their data retention policies to see if they can provide assurances not to store your data. It is worth noting that using an LLM can benefit user privacy if used in place of sharing data with human coders.

**Establish a gold standard set of human-generated labels**: Before trying to tag an entire dataset, researchers should first test and refine a codebook with only human coders until a reasonable inter-rater reliability is achieved on a sample of the data. Once achieved, the codebook can be iteratively tested and refined with a LLM until there is a reasonable inter-rater reliability between the LLM and human coders. It is also important to test an LLM several times with the same exact input to ensure the IRR is reliable across runs. A low "temperature" value is more likely to produce consistent results from run to run.

**Test various LLM configurations**: When testing LLM performance we recommend evaluating several LLMs and prompts. Prompts should not assume any prior knowledge about the task or data. They should provide the full context necessary to perform a task, just like the instructions that a crowd worker would be given for a similar task.

**Validate LLM output**: Despite clear prompting, LLMs may not always provide output in the expected format. For example, LLMs may hallucinate new labels or refuse to provide a label for content moderation reasons. Regex matching can be used to validate that each LLM response contains a valid label.

**Improve underperforming codes**: There may be categories where LLM and human labels consistently disagree. For these, we recommend iterating on the codebook language for the problematic categories. Approaches that may make categories more legible to an LLM can include simplifying language, providing specific instructions about how to deal with ambiguous cases, or splitting categories with multiple labels into several binary "yes/no" questions, one for each label. It is even possible to prompt the LLM to give an explanation for why it is choosing a specific tag, which may provide clues as to how the codebook language could be reworded to achieve a better result. We note that while these strategies are likely to improve labels to a point where they can be used for further analysis, the expert human labelers continued to have higher IRR, likely due to the iterative process and understanding of contextual information.

**Balance cost and performance**: Depending on the size of your dataset, LLM cost may be an issue. Costs can be reduced by using cheaper, less powerful versions of an LLM, and by prompting an LLM to generate labels for all categories at once, rather than one at a time. When using these techniques the important thing is to always validate on a test set that the LLM achieves acceptable inter-rater reliability with the gold standard human-generated labels.

## 5 CONCLUSION

When hand-labeling data is not feasible from a time or cost perspective, large language models may be an efficient alternative. The research community is still establishing best practices for how to use these emerging tools. We draw on our work to provide recommendations for how researchers can use LLMs for labeling tasks in a robust and repeatable manner.

## REFERENCES

[1] Bernard Berelson. [n. d.]. *Content analysis in communication research.* Free Press. Pages: 220.
[2] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.
[3] Patricia A. Cavazos-Rehg, Melissa J. Krauss, Shaina Sowles, Sarah Connolly, Carlos Rosas, Meghana Bharadwaj, and Laura J. Bierut. [n. d.]. A content analysis of depression-related Tweets. 54 ([n. d.]), 351–357. https://doi.org/10.1016/j.chb.2015.08.023
[4] Paul D'Angelo. [n. d.]. *Doing News Framing Analysis II: Empirical and Theoretical Perspectives.* Routledge. Google-Books-ID: emgPEAAAQBAJ.

[5] Jie Gao, Yuchen Guo, Toby Jia-Jun Li, and Simon Tangi Perrault. [n. d.]. CollabCoder: A GPT-Powered WorkFlow for Collaborative Qualitative Analysis. In *Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing* (New York, NY, USA, 2023-10-14) *(CSCW '23 Companion)*. Association for Computing Machinery, 354–357. https://doi.org/10.1145/3584931.3607500

[6] Simret Araya Gebreegziabher, Zheng Zhang, Xiaohang Tang, Yihao Meng, Elena L. Glassman, and Toby Jia-Jun Li. [n. d.]. PaTAT: Human-AI Collaborative Qualitative Coding with Explainable Interactive Rule Synthesis. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2023-04-19) *(CHI '23)*. Association for Computing Machinery, 1–19. https://doi.org/10.1145/3544548.3581352

[7] Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. [n. d.]. ChatGPT outperforms crowd workers for text-annotation tasks. 120, 30 ([n. d.]), e2305016120. https://doi.org/10.1073/pnas.2305016120 Publisher: Proceedings of the National Academy of Sciences.

[8] Fabienne Lind, Maria Gruber, and Hajo G. Boomgaarden. [n. d.]. Content Analysis by the Crowd: Assessing the Usability of Crowdsourcing for Coding Latent Constructs. 11, 3 ([n. d.]), 191–209. https://doi.org/10.1080/19312458.2017.1317338 Publisher: Routledge _eprint: https://doi.org/10.1080/19312458.2017.1317338.

[9] Claudia A. López and Brian S. Butler. [n. d.]. Consequences of content diversity for online public spaces for local communities. In *Proceedings of the 2013 conference on Computer supported cooperative work* (New York, NY, USA, 2013-02-23) *(CSCW '13)*. Association for Computing Machinery, 673–682. https://doi.org/10.1145/2441776.2441851

[10] Mary L. McHugh. [n. d.]. Interrater reliability: the kappa statistic. 22, 3 ([n. d.]), 276–282. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900052/

[11] Kimberly A. Neuendorf. [n. d.]. *The Content Analysis Guidebook*. SAGE. Google-Books-ID: nMA5DQAAQBAJ.

[12] OpenAI. [n. d.]. GPT-4 Technical Report. https://doi.org/10.48550/arXiv.2303.08774 arXiv:2303.08774 [cs]

[13] Daniel Riffe, Stephen Lacy, Frederick Fico, and Brendan Watson. [n. d.]. *Analyzing Media Messages: Using Quantitative Content Analysis in Research*. Routledge. Google-Books-ID: jTr3DwAAQBAJ.

[14] Ha Semetko and Pm Valkenburg. [n. d.]. Framing European politics: a content analysis of press and television news. 50, 2 ([n. d.]), 93–109. https://doi.org/10.1111/j.1460-2466.2000.tb02843.x _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1460-2466.2000.tb02843.x.

[15] Wouter van Atteveldt, Mariken A. C. G. van der Velden, and Mark Boukes. [n. d.]. The Validity of Sentiment Analysis: Comparing Manual Annotation, Crowd-Coding, Dictionary Approaches, and Machine Learning Algorithms. 15, 2 ([n. d.]), 121–140. https://doi.org/10.1080/19312458.2020.1869198 Publisher: Routledge _eprint: https://doi.org/10.1080/19312458.2020.1869198.

[16] Ziang Xiao, Xingdi Yuan, Q. Vera Liao, Rania Abdelghani, and Pierre-Yves Oudeyer. [n. d.]. Supporting Qualitative Analysis with Large Language Models: Combining Codebook with GPT-3 for Deductive Coding. In *Companion Proceedings of the 28th International Conference on Intelligent User Interfaces* (New York, NY, USA, 2023-03-27) *(IUI '23 Companion)*. Association for Computing Machinery, 75–78. https://doi.org/10.1145/3581754.3584136

[17] Kai-Cheng Yang and Filippo Menczer. [n. d.]. Large language models can rate news outlet credibility. https://doi.org/10.48550/arXiv.2304.00228 arXiv:2304.00228 [cs]

[18] Rachel Young, Stephanie Miles, and Saleem Alhabash. [n. d.]. Attacks by Anons: A Content Analysis of Aggressive Posts, Victim Responses, and Bystander Interventions on a Social Media Site. 4, 1 ([n. d.]), 2056305118762444. https://doi.org/10.1177/2056305118762444 Publisher: SAGE Publications Ltd.

[19] Rodrigo Zamith and Seth C. Lewis. [n. d.]. Content Analysis and the Algorithmic Coder: What Computational Social Science Means for Traditional Modes of Media Analysis. 659, 1 ([n. d.]), 307–318. https://doi.org/10.1177/0002716215570576 Publisher: SAGE Publications Inc.

[20] Yiming Zhu, Peixian Zhang, Ehsan-Ul Haq, Pan Hui, and Gareth Tyson. [n. d.]. Can ChatGPT Reproduce Human-Generated Labels? A Study of Social Computing Tasks. https://doi.org/10.48550/arXiv.2304.10145 arXiv:2304.10145 [cs]