A Framework for Enhancing Behavioral Science Research with Human-Guided Language Models

Jaelle Scheuerman, Dina Acklin

U.S. Naval Research Laboratory, Stennis Space Center, MS, USA
jaelle.scheuerman@nrlssc.navy.mil, dina.acklin@nrlssc.navy.mil

Abstract

Many behavioral science studies result in large amounts of unstructured data sets that are costly to code and analyze, requiring multiple reviewers to agree on systematically chosen concepts and themes to categorize responses. Large language models (LLMs) have potential to support this work, demonstrating capabilities for categorizing, summarizing, and otherwise organizing unstructured data. In this paper, we consider that although LLMs have the potential to save time and resources performing coding on qualitative data, the implications for behavioral science research are not yet well understood. Model bias and inaccuracies, reliability, and lack of domain knowledge all necessitate continued human guidance. New methods and interfaces must be developed to enable behavioral science researchers to efficiently and systematically categorize unstructured data together with LLMs. We propose a framework for incorporating human feedback into an annotation workflow, leveraging interactive machine learning to provide oversight while improving a language model's predictions over time.

Introduction

The use of computerized assistance for the coding of qualitative research in the field of behavioral science has a long history (Stone and Hunt 1963; Weber 1984). In recent years, natural language processing (NLP) has experienced many breakthroughs with the development of first pretrained transformer models like BERT, and more recently large language models (LLMs) like GPT-4. These models have been leveraged by behavioral science researchers to aid in qualitative coding and analysis due to their ability to generalize across a wide variety of zero shot and few shot classification tasks (Brown et al. 2020). Because configured LLMs are convenient to use through natural language inputs, they are largely accessible to domain experts outside the machine learning field.

Qualitative studies often contain unstructured data collected from open-ended responses, interviews, and other means. There are a number of methods for analyzing qualitative data, but all techniques require researchers to systematically code textual data in order to identify categories or themes across responses (Miles, Huberman, and Saldaña 2020; Corbin and Strauss 2008). This process can be done manually or with the aid of software programs, such as Nvivo[1] and ATLAS.ti[2], that aid in the coding, categorization, sorting, and organization of data prior to analysis. However, because of the complex and subjective nature of the coding process, there may be little commonality between the coding schemes of two independent researchers, and replication is a common concern (Huma and Joyce 2023; Makel et al. 2022).

Machine learning methods have been used in the behavioral sciences with some regularity to enhance the rigor of qualitative analysis by standardizing the coding process, while reducing the burden on researchers to generate new coding schemes when analyzing data from similar protocols and improving interrater reliability. Supervised learning has been notably useful to aid in deductive coding for content analysis, in which labels are generated a priori based on existing research or hypotheses and applied to a body of text. These methods have been demonstrated across a wide range of use cases, such as identifying suicidal ideation (Ji et al. 2018), coding open-ended survey responses (Baumgartner et al. 2021), and analyzing themes and trends in media publications (Dun, Soroka, and Wlezien 2021). While successful, supervised methods require a large corpus of labeled data, which may not exist for niche topics of study and require considerable time to develop. LLMs have been posited as a potential solution to this problem, allowing researchers to quickly perform deductive coding analysis on large datasets with a high degree of interrater reliability. In these use cases LLM's have been successfully deployed with some human oversight to provide few-shot examples to improve later LLM performance (Xiao et al. 2023) or develop and fine tune prompts that act as a proxy for codebooks. Appropriately tuned models have been shown to speed up the coding analysis, being 36 times faster than human research analysts in some scenarios (Chew et al. 2023).

Although LLMs show great promise in facilitating the deductive coding process for qualitative research, they should not be viewed as a replacement for humans, as recently suggested (Byun, Vasicek, and Seppi 2023). Human input is invaluable not only for the expertise needed to craft and

[1] https://lumivero.com/products/nvivo
[2] https://atlasti.com

finetune suitable prompts to generate appropriate coding schemes, but also for the validation of LLM outputs and to monitor for misleading bias and hallucinations. Although models can generate codes and labels based on relationships within the data, the reasoning process used may not speak to a researcher's primary objectives or questions, particularly in the case of deductive coding. Indeed, there has been a great deal of variability in interrater reliability between humans and LLMs, even with humans working alongside models to clarify coding strategies (Chew et al. 2023; Xiao Liu et al. 2022). The range of outcomes demonstrates the continued need for humans to guide and systematically tune LLMs to obtain desirable results. For the remainder of this paper, we will consider potential caveats for using unsupervised LLMs to analyze data collected from behavioral studies and we introduce a potential path forward incorporating human-guided language models into annotation and deductive coding workflows with interactive machine learning.

Challenges to Using Language Models for Deductive Coding and Annotation

Inaccurate and Biased Responses At their core, pretrained language models built upon the transformer architecture are trained on massive datasets to generate the patterns and biases learned from those datasets (Vaswani et al. 2017). Through those patterns, the models are able to generate responses that exhibit knowledge about the world at near human or human level performance over a variety of tasks (Brown et al. 2020). However, the same models also perform quite well in producing plausible sounding responses that range from minor errors to completely inaccurate statements, often called hallucinations (Ji et al. 2023). Further, generated responses exhibit the biases of their underlying text data and care must be taken not to make assumptions that any tools built on these models can be considered completely impartial, rational, or without bias (Caliskan, Bryson, and Narayanan 2017; Schramowski et al. 2022). Without careful inspection, these errors could result in machine-generated coding decisions that do not reflect the intent or objectives of the behavioral research analyst, or could potentially misrepresent the data altogether. For example, Julian Ashwin, Aditya Chhabra, and Vijayendra Rao (2023) demonstrated how the training data from three LLMs (ChatGPT-3.5, Llama-2 and the Llama-2 chat variant) failed to apply the appropriate context when performing coding on data that differed from the original training set (e.g., translated interviews from Rohingya refugees and local Bangladeshi residents). Not only did the LLMs perform poorly in terms of accuracy and precision, but the results were biased and based inappropriately on the demographic characteristics of the interview subjects, rather than the responses from their interviews. This could easily lead to inaccurate conclusions about the interviewees. Just as multiple raters are needed to overcome biases and come to agreed upon categories in deductive coding tasks, a combination of AI raters and human oversight may be required to arrive at reliable and reproducible labels of qualitative data.

Reliability and Reproducibility LLMs also exhibit varying degrees of reliability, with the accuracy and usability of responses significantly influenced by the models, parameters, and prompts employed. Since language models are trained on a wide variety of datasets and architectures, prompts used to query one model may require additional time and effort to validate for a new model. Yet, it may be desirable to test the performance of a variety of different models for a domain task or update to a new model as the technology progresses. In order for behavioral research analysts to maintain reproducibility, great care must be taken to ensure records are kept of the models, prompts, and parameters used to achieve desired results. To eliminate this burden, ML practitioners and AI system developers must make efforts to reduce the resources required to adapt prompts to new models (Dingliwal et al. 2021), preferably through interactive approaches that do not require the specialized efforts of a machine learning expert or prompt engineer (Wen et al. 2023).

Domain Knowledge A key feature of behavioral research tasks such as deductive coding requires the application of theory and domain expertise to the development of an appropriate coding scheme for the selected dataset. Language models remain limited in their ability to generate responses relating to domain areas not well represented in the training set and cannot easily be kept current with changes in real world data. Methods are being employed to overcome this shortcoming with prompt tuning, finetuning, or external databases, (Lester, Al-Rfou, and Constant 2021; Lewis et al. 2020) to bring the necessary context. However, these approaches still require significant time and effort to set up, which is not always appropriate for more specialized tasks and is not a task that most behavioral researchers would be able to independently employ. For these methods to become useful in specialized cases, efforts will need to be made to develop approaches that allow adaptation to new domains or tasks with minimal additional effort on the part of the researcher (Cite Ling et al. 2023).

Human-Guided Language Models with Interactive Machine Learning

To address the challenges described in the previous section, we introduce a method to incorporate online human feedback into a behavioral data coding workflow. Although LLMs can perform well across a wide range of topics for zero- or few-shot classification tasks, they cannot adapt to specific domains outside the training data or new and changing information. Methods such as finetuning, retrieval augmented generation, or composition are employed to help align an LLM with a particular domain or specialized knowledge or requirements (Agrawal et al. 2023; Ranade et al. 2021; Lester, Al-Rfou, and Constant 2021; Bansal et al. 2024). However these require additional efforts in ongoing curation of domain data and updated information.

We propose a framework built upon the principles of interactive machine learning (IML) to create systems to support behavioral research analysts in the coding process. IML is an annotation workflow that pairs human feedback with machine learning and an specialized user-friendly interface to

improve a model's performance over time (Fails and Olsen 2003; Michael, Acklin, and Scheuerman 2020). This approach differs from more conventional implementations of machine learning which rely on offline batch training and may only adapt to changing situations with statistically significant numbers of training examples. Applying IML in the context of language models would aid in data collection and domain alignment, while allowing analysts to tailor models to their specific research needs.

In the following section, we will describe the components of this framework and give an example of how it could be deployed to improve a pre-trained language model's performance for a specialized task like deductive coding. We will conclude by discussing the potential research directions required to deploy this framework to support behavioral scientists in similar data coding tasks.

Framework for Human-Guided Language Models

The framework for human-guided language models provides three high-level components that aid a system in supporting the researcher's workflow for domain-specific tasks such as deductive coding. We highlight some existing research that can be leveraged when building these components and note gaps where additional research is needed. The main components consist of the following: 1) model-agnostic LLM support to achieve effective performance when coding qualitative data, 2) the ability to use iterative feedback online to improve the language model's responses, and 3) a specialized interface for pairing LLM responses with a rater or group of raters in their desired qualitative analysis task.

Model Agnostic Performance The development of LLMs is proceeding rapidly, with new models being released daily and each model possessing its own strengths and weaknesses. It is untenable that researchers outside the machine learning field keep abreast of each new model iteration and update their research to effectively utilize these models. The framework described here is independent of any specific language model. Domain scientists should have the ability to deploy the language model best suited for their task, whether a foundation model accessed via API, or local model potentially finetuned for their specific domain, in order to maintain the desired performance. Currently, this presents many challenges, since time and resources must be spent identifying new prompts and parameters to achieve accurate and reliable performance for each model within a given domain. For specialized domains to realistically make use of LLMs, this process must be improved for users who are not ML practitioners. Some existing efforts have improved workflows for prompt visualization and testing (Hendrik Strobelt et al. 2022) or through prompt optimization strategies that can be incorporated into the user's workflow (Wen et al. 2023).

Online Iterative Feedback The most important component of any interactive machine learning workflow is the ability to provide online iterative feedback that can be immediately be used to improve the performance of the model during a data coding task. Since language models are static after being trained, feedback cannot be used to update the model weights or affect future predictions of the model iteratively in an online manner. However, the text generated by the model is affected by any content sent in the prompt context. This property can be used to provide immediate feedback as contextual input to the model and ultimately effect the output. Existing efforts have used retrieval augmented generation (RAG) to add relevant external context to a prompt to affect the resulting output (Luo et al. 2023). The content is converted to an embedding and stored in a vector database for later retrieval. Before submitting a prompt to a language model, the system can query the database for any content that is similar to the current query. If similar content is found, it is incorporated into the prompt and submitted to the language model. This approach can be extended and applied to corrections provided by the researcher that can be used as model feedback without retraining or finetuning. For example, if the language model misclassifies some data, the research analyst can provide a correction, which can be stored as the triplet (trial description, model response, feedback). The scenario is stored as a vector embedding so that it can easily be compared with future trials waiting to be annotated. If another scenario is similar, then the outcome and feedback from the previous scenario is retrieved and the prompt can be formulated to incorporate those details into the current context.

We explored the effectiveness of using user-supplied corrections to iteratively improve the predictions of a model using RAG to aid in coding qualitative data collected from a behavioral task (see Figure 1). In the task, the participants described the source of various sounds. To code the data, two raters were asked to agree on whether the user's response accurately described the source category of each sound. A language model (Mistral-OpenOrca-7b) was also given instructions to respond yes or no about whether a user-provided response meant the same thing as a particular category of sound. Corrections were provided to the model based on a ground truth responses provided by the raters. If the language model's response differed from the ground truth response, the qualitative description of the sound was stored as a vector embedding, along with the correct response. When similar descriptions were encountered in later trials, the feedback was retrieved from the vector database and the model was instructed to use the correct response. The similarity threshold of the model was set with a parameter representing the minimum similarity required for feedback to be retrieved from the database. It was found that some care is required when setting the similarity parameter. As seen in Table 1, if the minimum similarity parameter was set too low, then irrelevant content was added to the prompt context,

Min. Sim.	0.7	0.8	0.9	No Fdbk
Precision	**86.3**	76.24	74.09	73.6
Recall	77.4	97.5	**95.9**	**95.9**
F1	81.6	**85.6**	83.6	83.3
Accuracy	84.1	**85.5**	82.5	82.1

Table 1: Comparison of model performance with feedback across different min. sim. thresholds and without feedback

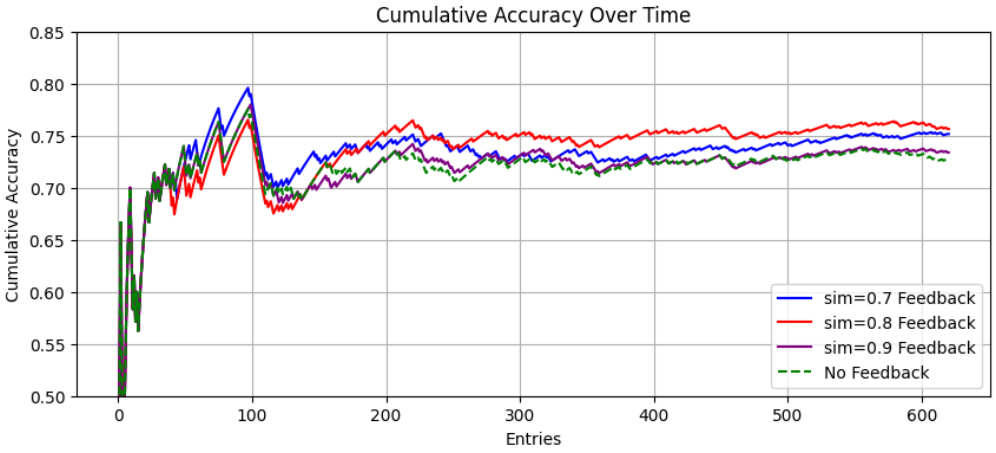

Figure 1: Cumulative accuracy of machine predictions over time, with and without feedback, at different similarity thresholds

leading to higher precision but lower overall accuracy. If the parameter was set too high, then important feedback was not retrieved when needed, leading to overall lower accuracy, precision, and recall. When the minimum similarity value was well calibrated, the model's responses improved from feedback, increasing the accuracy and recall over time, while only lowering the precision slightly. Although this approach relied on a priori ground truth to determine optimal tuning parameters, the methodology for incorporating rater input could be extended to allow users to provide online feedback during a similar coding task. Using this input structure, future research should investigate ways of automatically calibrating the minimum similarity parameter from the analyst's interactions with the annotation interface.

Specialized Interface for LLM-Rater Teams The final component of the framework is the annotation interface that the research analyst uses to validate and correct the output of the language model. Preferably, this would involve sorting the machine responses by how certain the model is in its ability to correctly annotate them. This way, the analyst could prioritize providing feedback for trials that the model is least confident about. That feedback would be incorporated into future prompts for related trials, improving their output, and resulting in fewer corrections that the analyst must make in the future. Because of the real-time nature of the validate and correct feedback loop, it is important that any estimate of confidence is also based on the real-time feedback that was received, and not just the language model's estimate of probability that was generated from the model weights. A method for using reinforcement learning to calibrate uncertainty from iterative online feedback was recently introduced in (Bishof, Scheuerman, and Michael 2023). A similar approach could be applied in the context of pre-trained language models. Gao et al. (2023) describe an interface developed for qualitative analysis that utilizes GPT 3.5 at several stages to aid research analysts throughout the data coding process. For example, during the initial generation stage, researchers can independently provide codes or select a suggested one from GPT. In either case, the researcher can also provide a confidence score for their chosen label to improve subsequent suggestions and aid in later assessments of interrater reliability.

Discussion and Future Directions

In this paper, we reviewed several challenges that could pose significant hurdles to the adoption of LLMs in qualitative research, including inaccurate and biased responses, reliability in model responses, and challenges in incorporating domain knowledge. Currently, the process of finetuning a model with specialized knowledge, engineering effective prompts, and tuning settings requires specialized skills and significant time and effort compared to the current state of the art qualitative coding methods. Overcoming this hurdle will require developing interfaces that support training or tuning a model to a particular domain as part of a research analyst's normal annotation workflow and streamlining prompt requirements across models. We introduced a framework for human-guided language models to support behavioral scientists in performing qualitative analysis tasks like deductive coding, along with guidelines and recommendations for future research and development efforts. We discussed the importance of ensuring that interfaces are model agnostic, giving domain scientists the flexibility to use the model that best suits their needs. Additionally, we explored the benefits of developing effective methods for integrating iterative online feedback to a language model, which necessitates the development of methods to calibrate the language model's confidence predictions for labels as new feedback is obtained. We illustrated one method of using iterative feedback to influence model output, using RAG to provide corrections to a language model completing a qualitative coding task. Finally, we emphasized the need for new specialized interfaces for qualitative coding tasks, with features such as the ability to recommend codes based on analyst feedback, integrate feedback from multiple raters, or act as mediator when raters are not aligned (Gao et al. 2023).

References

Agrawal, G.; Kumarage, T.; Alghami, Z.; and Liu, H. 2023. Can Knowledge Graphs Reduce Hallucinations in LLMs? : A Survey. arxiv:2311.07914.

Bansal, R.; Samanta, B.; Dalmia, S.; Gupta, N.; Vashishth, S.; Ganapathy, S.; Bapna, A.; Jain, P.; and Talukdar, P. 2024. LLM Augmented LLMs: Expanding Capabilities through Composition. arxiv:2401.02412.

Baumgartner, P.; Smith, A.; Olmsted, M.; and Ohse, D. 2021. A Framework for Using Machine Learning to Support Qualitative Data Coding. Preprint, Open Science Framework.

Bishof, Z.; Scheuerman, J.; and Michael, C. J. 2023. Closed-Loop Uncertainty: The Evaluation and Calibration of Uncertainty for Human–Machine Teams under Data Drift. *Entropy*, 25.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models Are Few-Shot Learners. arxiv:2005.14165.

Byun, C.; Vasicek, P.; and Seppi, K. 2023. Dispensing with Humans in Human-Computer Interaction Research. 1–26.

Caliskan, A.; Bryson, J. J.; and Narayanan, A. 2017. Semantics Derived Automatically from Language Corpora Contain Human-like Biases. *Science*, 356(6334): 183–186.

Chew, R.; Bollenbacher, J.; Wenger, M.; Speer, J.; and Kim, A. 2023. LLM-Assisted Content Analysis: Using Large Language Models to Support Deductive Coding. arxiv:2306.14924.

Corbin, J. M.; and Strauss, A. 2008. *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*. SAGE Publications. ISBN 978-1-4129-0644-9.

Dingliwal, S.; Shenoy, A.; Bodapati, S.; Gandhe, A.; Gadde, R. T.; and Kirchhoff, K. 2021. Prompt-Tuning in ASR Systems for Efficient Domain-Adaptation. arxiv:2110.06502.

Dun, L.; Soroka, S.; and Wlezien, C. 2021. Dictionaries, Supervised Learning, and Media Coverage of Public Policy. *Political Communication*, 38(1-2): 140–158.

Fails, J. A.; and Olsen, D. R. 2003. Interactive Machine Learning. In *Proceedings of the 8th International Conference on Intelligent User Interfaces*, IUI '03, 39–45. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-58113-586-2.

Gao, J.; Guo, Y.; Lim, G.; Zhan, T.; Zhang, Z.; Li, T. J.-J.; and Perrault, S. T. 2023. CollabCoder: A GPT-Powered Workflow for Collaborative Qualitative Analysis. *arXiv preprint arXiv:2304.07366*.

Hendrik Strobelt; Albert Webson; Victor Sanh; Benjamin Hoover; Johanna Beyer; Hanspeter Pfister; and Alexander M. Rush. 2022. Interactive and Visual Prompt Engineering for Ad-hoc Task Adaptation With Large Language Models. *IEEE Transactions on Visualization and Computer Graphics*, 1–11.

Huma, B.; and Joyce, J. B. 2023. 'One Size Doesn't Fit All': Lessons from Interaction Analysis on Tailoring Open Science Practices to Qualitative Research. *British Journal of Social Psychology*, 62(4): 1590–1604.

Ji, S.; Yu, C. P.; Fung, S.-f.; Pan, S.; and Long, G. 2018. Supervised Learning for Suicidal Ideation Detection in Online User Content. *Complexity*, 2018.

Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y. J.; Madotto, A.; and Fung, P. 2023. Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, 55(12): 248:1–248:38.

Julian Ashwin; Aditya Chhabra; and Vijayendra Rao. 2023. Using Large Language Models for Qualitative Analysis Can Introduce Serious Bias. *arXiv (Cornell University)*.

Lester, B.; Al-Rfou, R.; and Constant, N. 2021. The Power of Scale for Parameter-Efficient Prompt Tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 3045–3059.

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; Riedel, S.; and Kiela, D. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, 9459–9474. Red Hook, NY, USA: Curran Associates Inc. ISBN 978-1-71382-954-6.

Luo, L.; Li, Y.-F.; Haffari, G.; and Pan, S. 2023. Reasoning on graphs: Faithful and interpretable large language model reasoning. *arXiv preprint arXiv:2310.01061*.

Makel, M. C.; Meyer, M. S.; Simonsen, M. A.; Roberts, A. M.; and Plucker, J. A. 2022. Replication Is Relevant to Qualitative Research. *Educational Research and Evaluation*, 27(1-2): 215–219.

Michael, C. J.; Acklin, D.; and Scheuerman, J. 2020. On Interactive Machine Learning and the Potential of Cognitive Feedback. arxiv:2003.10365.

Miles, M. B.; Huberman, A. M.; and Saldaña, J. 2020. *Qualitative Data Analysis: A Methods Sourcebook*. SAGE. ISBN 978-1-5443-7185-6.

Ranade, P.; Piplai, A.; Joshi, A.; and Finin, T. 2021. CyBERT: Contextualized Embeddings for the Cybersecurity Domain. In *2021 IEEE International Conference on Big Data (Big Data)*, 3334–3342.

Schramowski, P.; Turan, C.; Andersen, N.; Rothkopf, C. A.; and Kersting, K. 2022. Large Pre-Trained Language Models Contain Human-like Biases of What Is Right and Wrong to Do. *Nature Machine Intelligence*, (4): 258–268.

Stone, P. J.; and Hunt, E. B. 1963. A Computer Approach to Content Analysis: Studies Using the General Inquirer System. In *Proceedings of the May 21-23, 1963, Spring Joint Computer Conference*, AFIPS '63 (Spring), 241–256. New York, NY, USA: Association for Computing Machinery. ISBN 9781450378802.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention Is All You Need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, 6000–6010. Red Hook, NY, USA: Curran Associates Inc. ISBN 978-1-5108-6096-4.

Weber, R. P. 1984. Computer-aided content analysis: A short primer. *Qualitative sociology*, 7(1-2): 126–147.

Wen, Y.; Jain, N.; Kirchenbauer, J.; Goldblum, M.; Geiping, J.; and Goldstein, T. 2023. Hard Prompts Made Easy: Gradient-Based Discrete Optimization for Prompt Tuning and Discovery.

Xiao, Z.; Yuan, X.; Liao, Q. V.; Abdelghani, R.; and Oudeyer, P.-Y. 2023. 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*, IUI '23 Companion, 75–78. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701078.

Xiao Liu; Kaixuan Ji; Yicheng Fu; Weng Tam; Zhengxiao Du; Zhou Yang; and Jie Tang. 2022. P-Tuning: Prompt Tuning Can Be Comparable to Fine-tuning Across Scales and Tasks.