Deductive Coding's Role in AI vs. Human Performance [*]

Jeanne McClure
North Carolina State University
Raleigh, NC USA
jmmcclu3@ncsu.edu

Daria Smyslova
North Carolina State University
Raleigh, NC USA
dysmyslo@ncsu.edu

Amanda Hall
North Carolina State University
Raleigh, NC USA
ajhall6@ncsu.edu

Shiyan Jiang
North Carolina State University
Raleigh, NC USA
sjiang24@ncsu.edu

ABSTRACT

This study evaluates the effectiveness of Large Language Models in deductive coding within educational qualitative research, compared to human coders. Employing a mixed-methods approach, LLMs demonstrated an 8.5% higher overall accuracy, with significant improvements in nuanced category identification through advanced prompt engineering. The study advocates a hybrid approach of AI and human cognitive skills for efficient educational data analysis.

Keywords

Large Language Model, qualitative coding, prompting

1. INTRODUCTION

Qualitative education research values the nuanced stories that emerge from the lived experiences of participants, often analyzed through the labor-intensive process of manual coding. In this traditional setup, coders employ their cognitive skills to navigate complex data interpretations, typically utilizing deductive content analysis methods, such as deductive coding. Deductive coding involves applying a predefined set of codes to the data, based on existing theories or the researcher's prior knowledge and experiences [2, 9, 12, 23, 17, 22]. This method contrasts with inductive approaches where categories emerge from the data itself. Despite the meticulous attention to detail that human coders bring to the analysis, their work is fraught with challenges such as maintaining coding integrity, managing subjectivity, and navigating logistical constraints—issues that are further compounded by ambiguities and the incomplete nature of the textual data [17, 20, 14].

The emergence of Large Language Models (LLMs) introduces a promising new methodology that could potentially enhance the depth and precision of data interpretation in qualitative research. These advanced models, equipped with capabilities for In-context Learning (ICL), Chain-of-Thought (COT), and Assertion Enhanced Few-Shot Learning (AEFL), offer innovative approaches to generate contextually informed, nuanced responses [4, 26, 24]. However, the application of these technologies in handling the intricate and ambiguous texts typical of educational research, particularly through the lens of deductive content analysis, remains largely unexplored.

This study hypothesizes that LLMs can achieve accuracy and reliability in deductive coding comparable to or exceeding that of human coders, potentially improving efficiency and scalability in data analysis [15]. By exploring the effectiveness of advanced prompt engineering, this paper aims to investigate how LLMs perform in the context of deductive coding and assess their potential to transform qualitative research methodologies. We address critical research questions concerning their precision, accuracy, and the patterns of misclassification compared to traditional human coding, underlining the need for a collaborative approach that harnesses both human insight and machine efficiency.

2. RELATED WORK

Educational researchers have leveraged traditional AI tools, incorporating natural language processing (NLP) and machine learning techniques, to support deductive content analysis in qualitative research. These tools traditionally utilized specific dictionary approaches or coded documents based on pre-existing examples [18, 13, 21]. Specifically, Inquire, developed by Paredes et al. [18], utilized NLP advancements like word2vec to analyze large textual datasets, thereby enhancing the identification of semantically related passages and facilitating data exploration. This development offered a scalable, cost-efficient alternative to manual analysis, proving beneficial across multiple fields such as psychology, privacy, and well-being. However, the Inquire system faced challenges related to data specificity and reliance on outdated datasets, which underscored the critical importance of dataset relevance. In contrast, recent LLMs studies have shown capability in qualitative deductive coding, synthesizing diverse information to identify underlying themes without direct modeling, thus reducing selection bias and en-

[*] (Does NOT produce the permission block, copyright information nor page numbering). For use with edm_article.cls.

J. McClure, D. Smyslova, A. Hall, and S. Jiang. Deductive coding's role in ai vs. human performance. In B. Paaßen and C. D. Epp, editors, *Proceedings of the 17th International Conference on Educational Data Mining*, pages 809–813, Atlanta, Georgia, USA, July 2024. International Educational Data Mining Society.

https://doi.org/10.5281/zenodo.12729958

hancing research validity [1, 3].

The shift towards LLMs has been marked by the introduction of advanced techniques such as few-shot learning and COT prompting, which have significantly enhanced the predictive accuracy and efficiency of these models [4, 26, 27]. Few-shot learning allows LLMs to utilize a minimal set of examples to achieve agreement in coding complex syntactic structures and analyzing the intricacies of question complexity. The COT method, detailed by Wei et al. [26], involves a structured reasoning process in natural language that guides the model through logical steps towards problem resolution. This technique has proven effective, with LLMs, such as GPT-3, demonstrating a marked reduction in coding time for extensive texts compared to human coders, showcasing its efficiency in handling large datasets [6]. Additionally, the Assertion Enhanced Few-Shot Learning (AEFL) technique introduced by Shariar et al. [24] integrates domain-specific assertions within prompts, enhancing the accuracy of LLMs and minimizing error rates.These advancements collectively signify a progression in LLM capabilities, decisively surpassing traditional machine learning approaches in specific tasks, and heralding a new era of efficiency and effectiveness in qualitative data analysis, particularly in the nuanced tasks of qualitative coding [6, 16].

3. METHODOLOGY

This exploratory study examines a subset of data from high school students engaged in an English Language Arts curriculum that is centered around artificial intelligence [5]. Over three weeks, 28 de-identified students from diverse racial and grade backgrounds engaged in daily 45-minute sessions that spanned eight modules, with particular attention to ”Sentiment Analysis,” ”Features and Models,” and ”All Words.” The primary analysis entailed coding 840 open-ended student responses using an adapted Cognitive engagement framework called ICAP [7], which classifies cognitive engagement into Passive (basic recall, N=266), Active (integration of new information, N=241), and Constructive (creation of new ideas, N=32) categories. The framework has been previously validated in various educational studies for enhancing cognitive engagement through active learning [25, 19]. To ensure robustness the researchers codebook was refined through coding sessions and a validation process where researchers independently applied the guidelines, achieving an inter-rater reliability (IRR) of Cohen’s K = .84 [8].

The participants included two educational graduate student researchers proficient in Bloom’s Taxonomy and cognitive engagement. In addition, GPT-4 was utilized via the Colab Python OpenAI API, setting the temperature to 0 which allowed for the most probable response to maintain uniformity and predictability in precision-critical applications. [25]. The LLM’s performance on the same subset of preprocessed data, which was carefully curated to maintain the integrity of student responses, provided a basis for comparing the coding accuracy of human coders versus advanced AI technology.

3.1 Experiments

The human coders and LLMs conducted a series of three experiments, each designed to progressively build upon the previous one, involving a total of 100 diverse random samples. The first experiment included 25 samples categorized into Passive (10), Active (14), and Constructive (1) engagement levels, utilizing a streamlined COT (Completion-Oriented Task) prompt. This prompt was specially crafted with three steps, incorporating domain-specific details to clarify the engagement levels, and required LLMs to explain their reasoning for each classification.

The second experiment also involved 25 samples and revisited the COT prompt, this time enhanced with Few-Shot learning examples in a tripartite format of Question, Response, and Label, offering four instances per cognitive engagement category. The third experiment expanded to 50 samples, merging elements from the first two experiments and integrating the reasoning component from AEFL [24]. This included a ”Reasoning” step in the Few-Shot format to better align with the reasoning processes defined in the original codebook. Further details on these prompts and the structured approach used in the experiments can be found in the supplemental materials link.

3.2 Evaluation Process

To evaluate the coding performance of both human coders and LLMs, we adopted a mixed-methods approach combining comparative, class-based, and thematic analyses [9]. We conducted quantitative assessments, including precision, recall, and F1 scores, to compare coding accuracy between humans and LLMs in a multiclass context, where these metrics gauge a model’s predictive accuracy, completeness, and balanced performance [10]. Additionally, ANOVA statistical analyses were employed to detect any significant differences in the results of the experiments [11].

4. RESULTS AND DISCUSSION

4.1 RQ1. How do the F1 score, precision, and accuracy of deductive coding decisions made by LLMs compare to those made by human coders?

Table 1: Performance Measures (a) Experiment 1, (b) Experiment 2, (c) Experiment 3

(a) Experiment 1

	Human Coder 1			Human Coder 2			LLM-GPT4		
Type	P	R	F1	P	R	F1	P	R	F1
Passive	.10	.60	.75	.10	.30	.46	.83	.50	.60
Active	.69	.62	.66	.50	.57	.53	.68	.93	.79
Constructive	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

(b) Experiment 2

	Human Coder 1			Human Coder 2			LLM-GPT4		
Type	P	R	F1	P	R	F1	P	R	F1
Passive	.80	.80	.80	.78	.70	.74	.77	.10	.87
Active	.87	.87	.87	.75	.60	.67	.10	.80	.89
Constructive	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

(c) Experiment 3

	Human Coder 1			Human Coder 2			LLM-GPT4		
Type	P	R	F1	P	R	F1	P	R	F1
Passive	.59	.10	.74	.71	.85	.77	.77	.79	.77
Active	.77	.40	.53	.74	.61	.71	.67	.64	.65
Constructive	.33	.20	.25	.33	.20	.25	.29	.40	.33

Initial observations highlighted a common challenge in Constructive engagement, where both LLMs and human coders recorded minimal scores, with LLMs later achieving a 33% F1 score improvement in this category by the third experiment. This demonstrates LLMs' adaptability, particularly when enhanced by reasoning-based inputs (see Table 1).

In terms of Active and Passive engagement, LLMs demonstrated notable performance improvements in the second experiment, achieving an 89% F1 score for Active engagement with a precision of 10%, and in the third experiment, they reached up to 100% precision and an 87% F1 score for Passive engagement. These results underscore LLMs' potential efficiency and superiority in coding tasks, even amidst the inherent ambiguities and complexities discussed by qualitative researchers [22, 17, 20]. The strategic use of ICL, COT, and AEFL prompting techniques has enhanced LLMs' performance, particularly in differentiating between Passive and Active engagements, achieving over 70% in both precision and recall rates in the third experiment.

Moreover, a precision-recall analysis across the experiments revealed consistent patterns of LLMs achieving higher precision than human coders, with performance converging in the third experiment where the coding differences were less pronounced (see Figure 1). Statistical analysis using ANOVA showed no significant differences among the groups (Precision: F =0.033, P = 0.97; Recall: F = 0.15, P= 0.87; F1: F=0.03, P=0.98), suggesting that while LLMs show promise, their application needs careful management to fully realize their potential in enhancing qualitative research methodologies.

4.2 RQ2. What are the specific similarities and differences of using LLMs for deductive coding in qualitative research, as compared to traditional human coding processes?

The comparative analysis between LLMs and human coders in qualitative research offers insights into their respective strengths and limitations within the deductive coding process. While LLMs effectively address some of the inherent subjectivity and logistical constraints of human coding, they also struggle with the complexities of interpreting nuanced student texts, which can exacerbate traditional coding challenges [17, 20, 14]. This aligns with qualitative educational research that values detailed narratives [2, 9, 12].

Throughout the experiments, the coding performance of LLMs and human coders varied significantly:

- **Experiment 1:** Employing basic prompt designs, this experiment revealed distinct error patterns; LLMs and Human Coder 2 showed higher misclassification counts, particularly when categorizing 'Passive' instead of 'Active' engagements, with LLMs mirroring the error rates of Human Coder 2.

- **Experiment 2:** Introduction of few-shot learning aimed to refine the coding process, enhancing LLMs' accuracy. This approach reduced error counts in LLMs significantly in 'Active instead of Passive' categories, aligning more closely with the more accurate Human Coder 1.

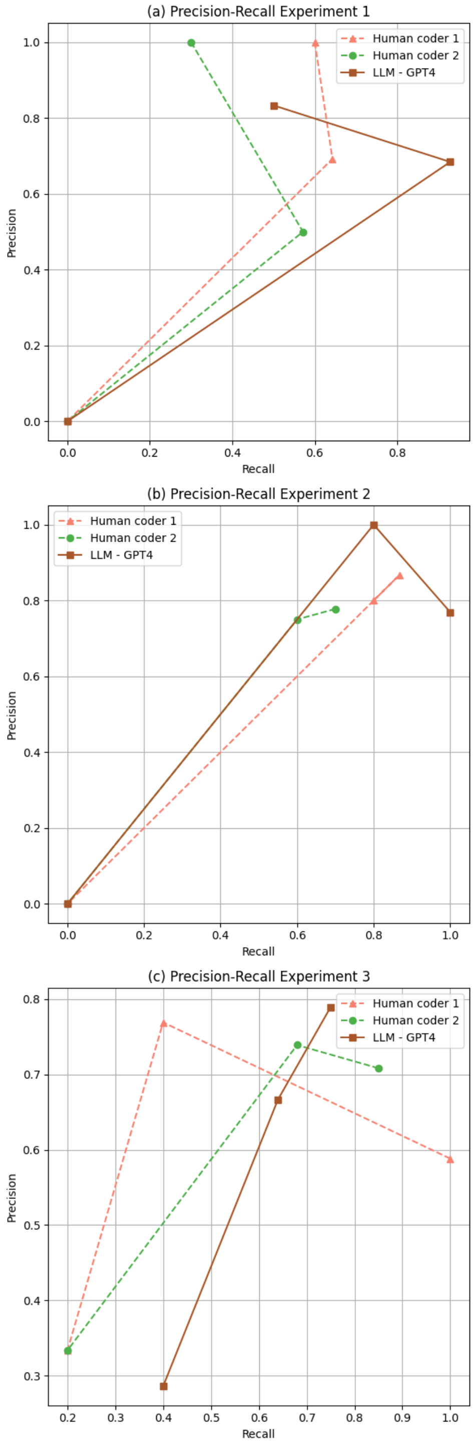

Figure 1: Precision-Recall plots for each experiment comparing the performance of Human coders 1 and 2 with LLM - GPT4.

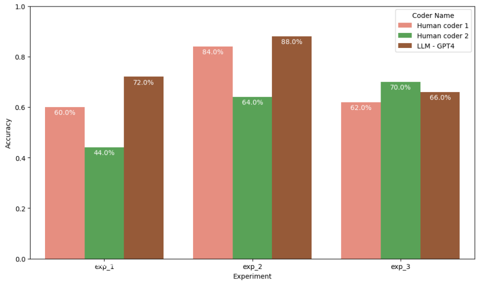

Figure 2: Experiment accuracy scores for two human coders and the LLMs across three experiments.

- **Experiment 3:** This experiment increased complexity by incorporating reasoning tasks into the coding process, resulting in a spike in misclassification rates for both human coders and LLMs, especially notable in Human Coder 1's errors in 'Active instead of Passive' categories, demonstrating the challenges of applying complex coding schemes.

These initial results demonstrate that while LLMs can efficiently process and categorize large datasets, their performance in complex coding scenarios often requires further refinement. Human coders show greater variability in their judgments, which can be a strength in interpreting complex, nuanced responses but also leads to inconsistency. This contrast highlights LLMs' potential as supportive tools in qualitative research, suggesting that a hybrid approach combining human oversight with LLM efficiency could optimize coding outcomes, particularly for complex interpretative tasks.

4.3 Limitation and Conclusion

This study offers crucial insights into the efficacy of LLMs in deductive coding within educational research, benchmarking their performance against traditional human coders. The results substantiate our hypothesis that LLMs, through strategically engineered prompts, can exceed human accuracy and efficiency, achieving an average of 8.5% greater accuracy, particularly when employing complex combination prompts [2, 9, 12, 22, 23]. This emphasizes the importance of prompt engineering in amplifying the capabilities of LLMs, presenting a scalable and efficient approach for educational data analysis [15]. However, the study's findings are bound by certain limitations, including the narrow scope of our dataset and the specific nature of the educational content, which may limit the generalizability of our results to other contexts or disciplines. Additionally, our primary reliance on the current functionalities of GPT-4 and methodologies like COT, Few-Shot, and partial AEFL reasoning could have restricted a more profound exploration of LLMs' potential in qualitative analysis.

The study significantly enhances LLM effectiveness over traditional methods by adopting Shariar et al.'s reasoning strategy [24], highlighting the need for broader coding tasks and more varied datasets. Future research will integrate targeted assertions to improve LLM accuracy and contextual relevance and develop a universal prompt framework for diverse responses [24]. Additionally, we will explore the implications of LLMs in real-time educational settings to enhance collaboration between human cognitive skills and AI's analytical capabilities, advancing the effectiveness of educational data mining for context-aware analysis. This ongoing research promises robust, scalable, and objective methodologies in educational research, showcasing the significant potential of LLMs with proper prompt engineering [18, 1].

5. ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant No. DRL-1949110. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

6. REFERENCES

[1] A. Bewersdorff, C. Hartmann, M. Hornberger, K. Seßler, M. Bannert, E. Kasneci, G. Kasneci, X. Zhai, and C. Nerdel. Taking the next step with generative artificial intelligence: The transformative role of multimodal large language models in science education. *arXiv preprint arXiv:2401.00832*, 2024.

[2] R. Bogdan and S. K. Biklen. *Qualitative research for education.* Allyn & Bacon Boston, MA, 1997.

[3] M. Bowman, S. K. Debray, and L. L. Peterson. Reasoning about naming systems. *ACM Trans. Program. Lang. Syst.*, 15(5):795–825, November 1993.

[4] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[5] J. Chao, B. Finzer, C. P. Rosé, S. Jiang, M. Yoder, J. Fiacco, C. Murray, C. Tatar, and K. Wiedemann. Storyq: A web-based machine learning and text mining tool for k-12 students. In *Proceedings of the 53rd ACM Technical Symposium on Computer Science Education V. 2*, pages 1178–1178, 2022.

[6] R. Chew, J. Bollenbacher, M. Wenger, J. Speer, and A. Kim. Llm-assisted content analysis: Using large language models to support deductive coding. *arXiv preprint arXiv:2306.14924*, 2023.

[7] M. T. Chi and R. Wylie. The icap framework: Linking cognitive engagement to active learning outcomes. *Educational psychologist*, 49(4):219–243, 2014.

[8] J. Cohen. *Statistical power analysis for the behavioral sciences.* Academic press, 2013.

[9] J. W. Creswell and C. N. Poth. *Qualitative inquiry and research design: Choosing among five approaches.* Sage publications, 2016.

[10] Á. B. Jiménez, J. L. Lázaro, and J. R. Dorronsoro. Finding optimal model parameters by discrete grid search. In *Innovations in Hybrid Intelligent Systems*, pages 120–127. Springer, 2008.

[11] H. J. Keselman, C. J. Huberty, L. M. Lix, S. Olejnik, R. A. Cribbie, B. Donahue, R. K. Kowalchuk, L. L. Lowman, M. D. Petoskey, J. C. Keselman, et al. Statistical practices of educational researchers: An analysis of their anova, manova, and ancova analyses. *Review of educational research*, 68(3):350–386, 1998.

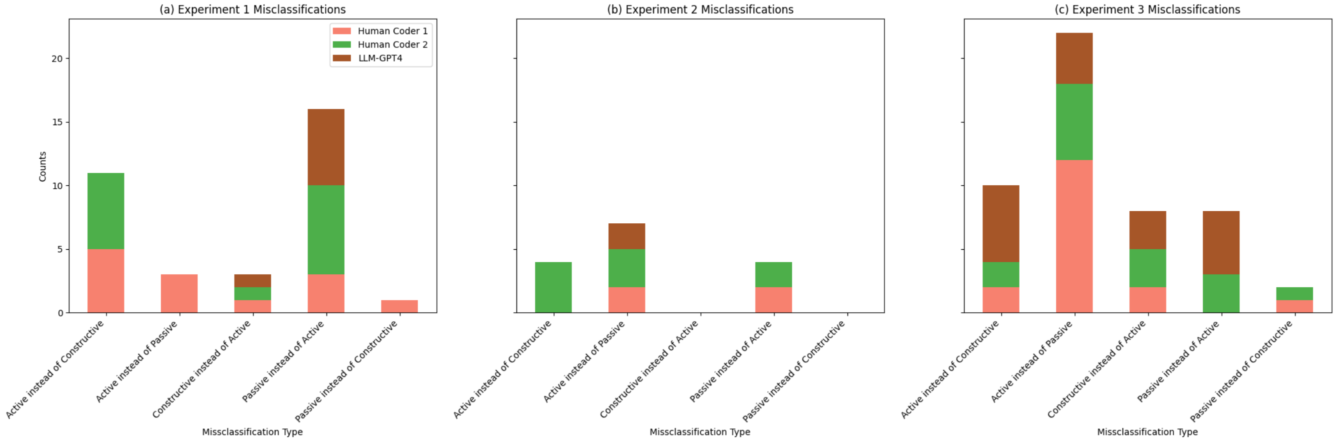

Figure 3: Comparison of misclassification across three experiments between coders

[12] M. Lichtman. *Qualitative research in education: A user's guide*. Routledge, 2023.

[13] J. S. Y. Liew, N. McCracken, S. Zhou, and K. Crowston. Optimizing features in active machine learning for complex qualitative content analysis. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 44–48, 2014.

[14] C. MacPhail, N. Khoza, L. Abler, and M. Ranganathan. Process guidelines for establishing intercoder reliability in qualitative studies. *Qualitative research*, 16(2):198–212, 2016.

[15] E. Mazzullo, O. Bulut, T. Wongvorachan, and B. Tan. Learning analytics in the era of large language models. *Analytics*, 2(4):877–898, 2023.

[16] J. McClure, F. Bickel, C. Tatar, D. Mushi, S. Jiang, and C. Rosé. Mining high school students' cognitive engagement from open-responses on machine learning practices. In *Proceedings of the 7th Educational Data Mining in Computer Science Education (CSEDM) Workshop*, Arlington, TX, United States, 2023.

[17] M. Miles. Qualitative data analysis: A methods sourcebook, by miles, huberman, and saldana, is the latest edi-tion of a longtime favorite of mine. saldana's the coding manual for qualitative researchers is new to me, but the two books work hand in hand. the second edition of miles and huberman's stellar vol-ume, published in 1994, has always been my "go to"book for qualitative data analysis. imagine my sadness when i learned that both miles and huberman had died (1996 and 2001, respectively).

[18] P. Paredes, A. Rufino Ferreira, C. Schillaci, G. Yoo, P. Karashchuk, D. Xing, C. Cheshire, and J. Canny. Inquire: Large-scale early insight discovery for qualitative research. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 1562–1575, 2017.

[19] K. M. Quesnelle, N. T. Zaveri, S. D. Schneid, J. B. Blumer, J. L. Szarek, M. Kruidering, and M. W. Lee. Design of a foundational sciences curriculum: applying the icap framework to pharmacology education in integrated medical curricula. *Pharmacology Research & Perspectives*, 9(3):e00762, 2021.

[20] M. S. Rahman. The advantages and disadvantages of using qualitative and quantitative approaches and methods in language "testing and assessment" research: A literature review. 2020.

[21] T. Rietz and A. Maedche. Cody: An ai-based system to semi-automate coding for qualitative research. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2021.

[22] J. Saldaña. The coding manual for qualitative researchers. *The coding manual for qualitative researchers*, pages 1–440, 2021.

[23] M. Schreier. *Content analysis, qualitative*. SAGE Publications Ltd, 2019.

[24] T. Shahriar, N. Matsuda, and K. Ramos. Assertion enhanced few-shot learning: Instructive technique for large language models to generate educational explanations. *arXiv preprint arXiv:2312.03122*, 2023.

[25] X. Wang, Y. Wang, C. Xu, X. Geng, B. Zhang, C. Tao, F. Rudzicz, R. E. Mercer, and D. Jiang. Investigating the learning behaviour of in-context learning: a comparison with supervised learning. *arXiv preprint arXiv:2307.15411*, 2023.

[26] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

[27] Z. Xiao, X. Yuan, Q. V. Liao, R. Abdelghani, and P.-Y. Oudeyer. 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*, pages 75–78, 2023.

APPENDIX

A. APPENDIX A

Supplemental material at https://osf.io/86h9x/?view$_{o}nly = f5fae0d1c4404ca4913c3535a86c60de$