Revolutionizing Qualitative Human-Robot Interaction Research by Using GPT Models for Inductive Category Development*

Clarissa Sabrina Arlinghaus[1], Charlotte Wulff[2], and Günter W. Maier[3]

***Abstract*— Coding qualitative data is essential but time-consuming. This late-breaking report presents a new method for developing inductive categories utilizing GPT models. We examined two different GPT models (gpt-3.5-turbo-0125 and gpt-4o-2024-05-03) and three temperature settings (0, 0.5, 1), each with ten repetitions. The generated categories were fairly consistent across settings, although higher temperatures included less relevant aspects. The agreement for GPT-generated category assignments exceeded that of human coders, with the best performance observed at temperature setting 0. Thus, we recommend using a GPT model with the temperature setting 0 to create and assign inductive categories for qualitative data.**

I. Introduction

Qualitative research is an essential aspect of human-robot interaction (HRI) studies [1]. Therefore, many HRI researchers collect qualitative data. Responses from participants to open-ended questions or during interviews convey valuable insights about users' thoughts, impressions, feelings, or attitudes. To analyze this data, specific methods must be applied, such as qualitative content analysis according to Mayring [2] or the Grounded Theory [3].

Several approaches involve the coding of statements [2], [3], [4]. Typically, this is done by two researchers [5], although having more coders would be beneficial [6]. However, forming multi-coder teams is challenging to achieve. It is already an existing problem, that plenty of qualitative data remains underutilized due to the time-consuming nature of qualitative analyses and the limited resources and time available to researchers [7].

In addition, nowadays there is an increasing prevalence of HRI online studies due to the effectiveness of video interactions [8], [9]. Online studies allow researchers to quickly and cost-efficiently reach a larger audience [10] but then there is also even more qualitative data to be analyzed. Even if there is only one minor open-ended question included in an online questionnaire, qualitatively analyzing hundreds of free-text responses can be overwhelming. Consequently, it is not surprising that HRI online studies with more than 200 [11], [12], 300 [13] or even 700 [14] participants focus on quantitative measures. Nevertheless, such large sample sizes could also be highly beneficial for qualitative research questions. In our view, a lot of potential remains unused, as the extensive effort required for large qualitative analysis is usually too challenging.

To address this issue, we aim to develop a new method that simplifies the coding process. This will lower the barrier to analyzing qualitative data, resulting in less data being lost.

The introduction of ChatGPT marked a significant technological advancement, making artificial intelligence (AI) accessible to the general population [15]. This led us to the idea that GPT models could facilitate qualitative research, specifically in the coding of statements. We have found studies demonstrating that Large Language Models (LLMs), such as BERT [7] and GPT-3.5 [16], [17], are suitable for deductive coding. GPT-3.5 has achieved results comparable to human coders [16]. However, the potential of LLMs for inductive coding yet remains under-explored. Building on the promising findings from LLM-aided deductive coding [7], [16], [17], we aim to find out whether LLMs are also suitable for inductive coding.

In our late-breaking work consisting of two pre-registered studies [19], [20], we assess the potential of GPT models to support qualitative research through inductive coding. Focusing on inductive category development (cf. [2]), we request two GPT models (gpt-3.5-turbo-0125 and gpt-4o-2024-05-03) to create content categories based on diverse statements (Study 1). After that, we let these two GPT models assign the created categories back to the statements (Study 2). This procedure, visualized in Figure 1, is designed to evaluate the ability of GPT models to inductively code qualitative data.

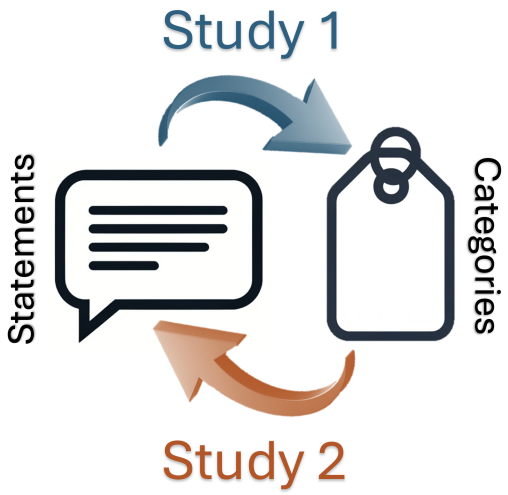

Fig. 1. Visualization of the Two Studies

*SAIL is funded by the Ministry of Culture and Science of the State of North Rhine-Westphalia under the grant no NW21-05A.

[1] Bielefeld University, Bielefeld, North-Rhine Westphalia, Germany clarissa_sabrina.arlinghaus@uni-bielefeld.de

[2] Bielefeld University, Bielefeld, North-Rhine Westphalia, Germany charlotte.wulff@uni-bielefeld.de

[2] Bielefeld University, Bielefeld, North-Rhine Westphalia, Germany ao-psychologie@uni-bielefeld.de

We will evaluate outputs from two GPT models with different temperatures based on established quality criteria in qualitative research (cf. [28]). Our goal is to derive recommendations for optimal settings when using GPT models for coding statements.

This is an exploratory examination of GPT models' potential to support qualitative research. We do not formulate any hypotheses in advance but instead investigate the following pre-registered research questions [19], [20]:

RQ1: Can GPT models be used for coding statements into categories?

RQ2: Which GPT model is the most suitable for coding statements into categories?

RQ3: What role does the temperature play?

By evaluating this novel AI method for qualitative analysis, we aim to assess its suitability as a coding assistant. If GPT models prove to be effective, it could revolutionize coding processes in qualitative research.

II. Methods

We practice Open Science. Therefore, we will upload all files (e.g., py, txt, xslx) in the Open Science Framework (OSF). During the IEEE RO-MAN conference, we will present our completely new method as late-breaking work and would be delighted if it is well-received and utilized by others. We permit other researchers to adapt and use our Python codes for their research, but kindly ask them to cite us when doing so. This is the link to our OSF project with all files: https://osf.io/h4dux/

This is a late-breaking report to present our current work at the IEEE RO-MAN conference. We are currently also working on a more detailed journal article. Anyone interested in our work is invited to take a look at our pre-print [18] and to monitor our pre-registration [19], [20]. Once a final publication is available, we will link it there.

A. Statements

In 2023, Clarissa Sabrina Arlinghaus started her PhD project about processes of social inclusion and exclusion in human-robot teams within the interdisciplinary research network SAIL [21]. Acknowledging the rise of human-robot interactions at work [22], her focus is on work-related settings where humans and robots cooperate. Considering that work provides not only financial benefits but also fulfills social needs [23], [24], she questions whether these needs are threatened in human-robot teams. In her first PhD study, she investigated two types of social exclusion (i.e., ostracism and rejection) through vignettes depicting human or robot waitstaff in a restaurant [25]. The online experiment included an open-ended question intended for qualitative analysis "Why do you think the robots/waiters behave this way towards you? Could the robots/waiters also behave differently?" and resulted in 296 statements from participants. Initially, these statements were together with Charlotte Wulff manually analyzed, but the substantial effort required led to the idea that a more efficient method should be developed. This idea formed the basis for the new AI method presented in this late-breaking report. We demonstrate the method using these statements, with the understanding that this new method can be applied to a wide range of qualitative data. The context of the collected data is thus secondary in this report. Those interested in the details of how individuals perceive social inclusion and exclusion by robots are encouraged to take a look at the pre-registration from the original study [25], where the final publication will be linked as soon as it is available. Moreover, the data and analyses regarding the PhD study will be made available under the following link: https://osf.io/mnybf/

B. Settings

For our studies, we decided to test the most powerful (gpt-4o-2024-05-03) and the most inexpensive (gpt-3.5-turbo-0125) GPT models that are currently (May 2024) available [26]. For the sake of simplicity, we will refer to them as gpt-3.5-turbo and gpt-4o in this paper. Python 3.12 in Visual Studio Code on macOS Sonoma 14.4.1 was used to access the application programming interface (API). Inspired by [27], we used a low (0.0), medium (0.5), and high (1.0) temperature as well as ten repetitions for every combination of temperature and model. This resulted in a total of 120 runs (60 runs in Study 1 and 60 runs in Study 2).

C. Procedure

In our study, we utilized two GPT models (gpt-3.5 and gpt-4o) to inductively code statements into categories. Potential categories were generated for these statements. This process was repeated multiple times across different models and temperature settings. The generated categories were then reviewed and consolidated to form final sets of categories. In the next step, these categories were assigned back to the original statements, again using gpt-3.5 and gpt-4o with three temperature settings (0, 0.5, 1), to verify the consistency and appropriateness of the categorization. We call this procedure LLM-Assisted Inductive Categorization (LAIC) which we introduce in more detail in our new paper that is currently under review but already available as a pre-print [18]. Usful instruction and explanations on how to use this method can be found in our OSF project.

D. Evaluation

To assess the performance of the two GPT models, we compared their results with categories generated and assigned by two human coders. The evaluation of the outputs was based on quality criteria for qualitative research [28].

Quality criteria of quantitative research (i.e., objectivity, reliability, validity, and generalizability), are not suitable for qualitative research because of their quantitative nature [28]. Instead, the quality of qualitative research should be rated regarding credibility, transferability, dependability, confirmability, and reflexivity [28]. We evaluated outputs from GPT models based on these criteria. Our goal is to assess whether GPT models can be used to code qualitative data and to provide recommendations on what model or temperature should be used for that.

III. Results

In this late-breaking report, the results are presented in a highly condensed form. For a more comprehensive evaluation, please look at our pre-print [18].

A. Study 1

Clarissa Sabrina Arlinghaus and Charlotte Wulff developed inductive categories in advance. Inductive categories were then formed by gpt-4o and gpt-3.5-turbo as part of Study 1.

The categories were very similar, not only between the human coders but also between the different GPT models. In both models, more inductive categories were formed at medium (0.5) or high (1) temperature than at low (0) temperature. According to our estimates, the most important aspects were already included in the low temperature. The new categories that emerged from the medium and high-temperature settings were, in our view, rather unimportant.

It was also noticeable that within the ten repetitions, the categories at temperature 0 were duplicated significantly more often than at temperatures 0.5 or 1. More variants for similar aspects (e.g., "character traits and behaviors", "character and personality", "individual character traits") were named at the medium or high-temperature settings. This creates an additional workload for the people using GPT models for inductive category development because the different label variants concerning similar aspects have to be summarized as one final category to avoid duplication. The AI method aims to reduce the workload. Therefore, we consider lower-temperature settings to be more suitable than medium or high-temperature settings.

B. Study 2

How valid the categories are, was tested in Study 2 by assigning them back to the statements.

Clarissa Sabrina Arlinghaus and Charlotte Wulff did this with both human-coded sets of categories. We then evaluated how often our two human coders agreed. 100 % agreement means that both human coders have chosen the same category for one statement. 0 % agreement describes that they chose different categories for one statement. While agree/disagree is a binary measure, we preferred using percentages to be able to compare it to the 120 runs performed with the two GPT models and three temperature. Agreement was determined for all statements. After that, an average value was calculated for all statements.

For the re-coding process of the GPT models, we consistently used the same settings (model, temperature, and iterations) as applied in Study 1. The GPT models assigned only the categories that were generated under these specific settings, rather than all possible categories. Using Python scripts, we then counted how often categories were assigned to which statements within the ten repetitions to assess the agreement. The most frequent number for an assigned category per statement was divided by the number of all assigned categories for this statement to determine the over-agreement as a percentage. For example, if, within ten runs, the category "programming and technology" was assigned eight times to a certain statement and the category "efficiency and performance" two times, then this was rated as an 80 % agreement for that statement. The agreement was determined for all statements individually for each model and temperature. A mean value was then calculated for each combination of model and temperature.

The quality criteria of dependability and confirmability mean that other researchers agree, that the results can be replicated and that the findings obtained from the data can be confirmed [28]. In our case, this is achieved through a high level of agreement. It turned out that the mean values for GPT models are higher than the mean values for our human coders (see Table I). The level of agreement was particularly high with the temperature setting 0. Therefore, the inductive coding process of GPT models is of higher quality concerning dependability and confirmability than our human codings. Thus, we recommend using a GPT model for inductive category development as it reduces the time efforts and also enhances the quality. For the best results, a low temperature (e.g., 0) should be used.

TABLE I
Levels of Agreement

categories from	mean	standard deviation
human coder 1	70.5 %	45.6 %
human coder 2	**65.2 %**	47.7 %
gpt-3.5-turbo temp 0	92.0 %	14.9 %
gpt-3.5-turbo temp 0.5	87.7 %	18.7 %
gpt-3.5-turbo temp 1	84.2 %	18.7 %
gpt-4o temp 0	**97.1 %**	9.3 %
gpt-4o temp 0.5	92.3%	14.2 %
gpt-4o temp 1	86.7 %	19.6 %

IV. Limitations and Future Work

Our new approach should be further studied. We consider testing different LLMs for future evaluation studies comparing their results. We also think of testing different type of qualitative data as well as different research contexts to demonstrate the broad applicability of LLM-assisted coding.

By doing so, challenges concerning LLMs must be taken into account. For example, this includes the potential generation of inaccurate and biased responses [29]. The outputs of the GPT models were of high quality, but they should still be checked critically. Future studies could categorize data where the risk of biased responses is particularly high to see how different LLMs respond in different settings. Additionally, LLMs raise ethical concerns, including issues related to privacy, the digital divide, and sustainability [30]. The significant energy consumption of LLMs also leads to a substantial carbon footprint [30]. Future research should figure out how to deal with these challenges. Recommendations on how LLMs for qualitative research can be used under particularly privacy-friendly and environmentally-friendly circumstances would be helpful and welcome.

V. Conclusions

Utilizing GPT models for inductive coding of qualitative data is a valid method that has the potential to revolutionize qualitative research in the field of HRI and beyond. It significantly reduces the time required and enables the clustering of big amounts of qualitative statements (e.g., free-text responses in large online studies) into categories. It is way easier to repeat this approach compared to manual coding by multiple researchers. The quality of inductive category formation by GPT models tends to be superior to that performed by humans. The choice of the GPT model is less critical than the selection of the temperature setting. For high consistency of outputs, the temperature should be set to a low value (e.g., 0).

References

[1] L. Veling, and C. McGinn, Qualitative research in HRI: A review and taxonomy, International Journal of Social Robotics, vol. 13, pp. 1689-1709, February 2021. https://doi.org/10.1007/s12369-020-00723-z

[2] P. Mayring, Qualitative content analysis, Forum: Qualitative Social Research, vol. 1, no. 2, article 20, June 2000. https://doi.org/10.17169/fqs-1.2.1089

[3] N. Vollstedt, and S. Rezat, An introduction to Grounded Theory with a special focus on axial coding and the coding paradigm, in G. Kaiser and N. Presmeg (Eds.), Compendium for Early Career Researchers in Mathematics Education, ICME-13 Monographs, April 2019. https://doi.org/10.1007/978-3-030-15636-7 _4

[4] H.-F. Hsieh, and S. E. Shannon, Three approaches to qualitative content analysis, Qualitative Health Research, vol. 15 no. 9, pp. 1277-1288, November 2005. https://doi.org/10.1177/1049732305276687

[5] I. G. Raskind, R. C. Shelton, D. L. Comenau, H. L. Cooper, D. M. Griffith, and M. C. Kegler, A review of qualitative data analysis practices in health education and health behavior research, Health Education and Behavior, vol. 46, no. 1, pp. 32-39, February 2019. https://doi.org/10.1177/1090198118795019

[6] S. P. Church, M. Dunn, and L. S. Prokopy, Benefits to qualitative data quality with multiple coders: Two case studies in multi-coder data analysis, Journal of Rural Social Sciences, vol. 34, no. 1, article 2, August 2019. https://egrove.olemiss.edu/jrss/vol34/iss1/2/

[7] P. Baumgartner, A. Smith, M. Olmsted, and D. Ohse, A framework for using machine learning to support qualitative data coding, Open Science Framework, November 2021. https://doi.org/10.31219/osf.io/fueyj

[8] L. Kunold. Seeing is not feeling the touch from a robot, in Proceedings of the 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), pp. 1562-1569, September 2022. https://doi.org/10.1109/RO-MAN53752.2022.9900788

[9] C. L. Gittens, and D. Garnes, Zenobo on Zoom: Evaluating the human-robot interaction user experience in a video-conferencing session, in 2022 IEEE International Conference on Consumer Electtonics (ICEE), pp. 1-6, March 2022. https://doi.org/10.1109/ICCE53296.2022.9730259

[10] M. T. Latkovikj, and M. B. Popovska, Online research about online research: advantages and disadvantages, E-methodology, vol. 6, no. 6, pp. 44-56, May 2020. http://dx.doi.org/10.15503/emet2019.44.56

[11] E. Roesler, S. Rudolph, and F. W. Siebert, Exploring the role of sociability, ownership, and affinity for technology in shaping acceptance and intention to use personal assistance robots, International Journal of Social Robotics, February 2024. https://doi.org/10.1007/s12369-024-01098-1

[12] M. Söderlund, Who is who in the age of service robots: The impact of robots' demand for user identification in human-to-robot interactions, Computers in Human Behavior: Artificial Humans, vol. 1, article 100013, September 2023. https://doi.org/10.1016/j.chbah.2023.100013

[13] C. S. Arlinghaus, C. Straßmann, and A. Dix, Increased morality through social communication or decision situation worsens the acceptance or robo-advisors, May 2024. http://dx.doi.org/10.31219/osf.io/bufjh

[14] C. Straßmann, and N. C. Krämer, A two-study approach to explore the effect of user characteristics on user's perception and evaluation of virtual asisstant's appearance, Multimodal Technologies and Interaction, vol. 2, no. 4, article 66, October 2018. https://doi.org/10.3390/mti2040066

[15] K. I. Roumeliotis, and N. D. Tselikas, ChatGPT and Open-AI models: A preliminary review, Future Internet, vol. 15, no. 6, article 192, May 2023. https://doi.org/10.3390/fi15060192

[16] R. Chew, J. Bollenbacher, M. Wenger, J. Speer, and A. Kim, LLM-assisted content analysis: Using large language models to support deductive coding, ArXiv, June 2023. https://doi.org/10.48550/arXiv.2306.14924

[17] R. H. Tai, L. R. Bentley, X. Xia, J. M. Sitt, S. C. Fankhauser, A. M. Chicas-Mosier, and B. G. Monteith, B. G., An examination of the use of large language models to aid analysis of textual data, International Journal of Qualitative Methods, vol. 23, pp. 1-14, January 2024. https://doi.org/10.1177/16094069241231168

[18] C. S. Arlinghaus, C. Wulff, and G. W. Maier, Inductive CodingwithChatGPT - An Evaluation of Different GPT Models Clustering Qualitative Data into Categories, OSF Preprints, July 2024. https://doi.org/10.31219/osf.io/gpnye

[19] C. S. Arlinghaus, and G. W. Maier, Clustering statements with ChatGPT – Study 1, Open Science Framework, May 2024. https://doi.org/10.17605/OSF.IO/TQNFK

[20] C. S. Arlinghaus, and G. W. Maier, Clustering statements with ChatGPT – Study 2, Open Science Framework, May 2024. https://doi.org/10.17605/OSF.IO/TP4BH

[21] SAIL, Sustainable life-cycle of intelligent socio-technical systems, Project website. https://www.sail.nrw

[22] S. K. Ötting, L. Masjutin, J. J. Steil, and G. W. Maier, Let's work together: A meta-analysis on robot design features that enable successful human-robot interaction at work, Human Factors, vol. 64, no. 6, pp. 1027–1050, September 2022. https://doi.org/10.1177/0018720820966433

[23] K. I. Paul, H. Scholl, K. Moser, A. Zechmann, and B. Batinic, Employment status, psychological needs, and mental health: Meta-analytic findings concerning the latent deprivation model, Frontiers in Psychology, vol. 14, article 1017358, March 2023. https://doi.org/10.3389/fpsyg.2023.1017358

[24] K. Isaksson, Unemployment, mental health and the psychological functions of work in male welfare clients in Stockholm, Scandinavian Journal of Social Medicine, vol. 17, no. 2, pp. 165–169, June 1989. https://doi.org/10.1177/140349488901700207

[25] C. S. Arlinghaus, and G. W. Maier, Different forms of social exclusion in a robo-restaurant, Open Science Framework, January 2024. https://doi.org/10.17605/OSF.IO/ZAM24

[26] OpenAI, Pricing, OpenAI, 2024. https://openai.com/api/pricing/

[27] J. Davis, L. Van Bulck, B. N. Durieux, and C. Lindvall, The temperature feature of ChatGPT: Modifying creativity for clinical research, JMIR Human Factors, vol. 11, article e53559, March 2024. http://dx.doi.org/10.2196/53559

[28] I. Korstjens, and A. Moser, Series: Practical guidance to qualitative research. Part 4: Trustworthiness and publishing, European Journal of General Practice, vol. 24, no. 1, pp. 120-124, 2018. https://doi.org/10.1080/13814788.2017.1375092

[29] J. G. Meyer, R. J. Urbanowicz, P. C. N. Martin, K. O'Conner, R. Li, P.-C. Peng, T. J. Bright, N. Tatonetti, K. Jae Won, G. Gonzales-Hernandez, and J. H. Moore, ChatGPT and large language models in academia: Opportunities and challenges, BioData Mining, vol. 16, article 20, 203. https://doi.org/10.1186/s13040-023-00339-9

[30] S. A. Khowaja, P. Khuwaja, K. Dev, W. Wang, W., and L. Nkenyereye, ChatGPT Needs SPADE (Sustainability, PrivAcy, Digital divide, and Ethics) evaluation: A review, Cognitive Computing, 2024. https://doi.org/10.1007/s12559-024-10285-1