RESEARCH ARTICLE

BON VIEW PUBLISHING

The Value of Generative AI for Qualitative Research: A Pilot Study

Frédéric Pattyn[1],*

[1]*Department of Business Informatics, Ghent University, Belgium*

Abstract: This mixed-methods approach study investigates the potential of introducing generative AI (ChatGPT 4 and Bard) as part of a deductive qualitative research design that requires coding, focusing on possible gains in cost-effectiveness, coding throughput time, and inter-coder reliability (Cohen's Kappa). This study involved semi-structured interviews with five domain experts and analyzed a dataset of 122 respondents that required categorization into six predefined categories. The results from using generative AI coders were compared with those from a previous study where human coders carried out the same task. In this comparison, we evaluated the performance of AI-based coders against two groups of human coders, comprising three experts and three non-experts. Our findings support the replacement of human coders with generative AI ones, specifically ChatGPT for deductive qualitative research methods of limited scope. The experimental group, consisting of three independent generative AI coders, outperformed both control groups in coding effort, with a fourfold (4x) efficiency and throughput time (15x) advantage. The latter could be explained by leveraging parallel processing. Concerning expert vs. non-expert coders, minimal evidence suggests a preference for experts. Although experts code slightly faster (17%), their inter-coder reliability showed no substantial advantage. A hybrid approach, combining ChatGPT and domain experts, shows the most promise. This approach reduces costs, shortens project timelines, and enhances inter-coder reliability, as indicated by higher Cohen's Kappa values. In conclusion, generative AI, exemplified by ChatGPT, offers a viable alternative to human coders, in combination with human research involvement, delivering cost savings and faster research completion without sacrificing notable reliability. These insights, while limited in scope, show potential for further studies with larger datasets, more inductive qualitative research designs, and other research domains.

Keywords: qualitative research, Cohen's Kappa, generative AI, ChatGPT, BARD

1. Introduction

Generative AI (GAI) has fundamentally transformed various industry practices, enabling businesses to innovate more efficiently and sustain revenue amid budget constraints [1]. The adoption of GAI reflects a growing imperative for productivity enhancements, as companies seek to achieve more with fewer [2]. While GAI has propelled innovation and delivered tangible outcomes across industries [3], academia has lagged in embracing these technologies, potentially risking long-term research productivity and growth [4].

The potential impact of GAI on research productivity is undeniable. In domains like quantum technologies, GAI has facilitated the development of new experimental methods and protocols, allowing researchers to explore uncharted territories with unprecedented efficiency [5]. Similarly, GAI applications in bioinformatics have significantly reduced the time required to determine complex protein structures, advancing scientific discoveries in a cost-effective manner [6]. Beyond these examples, GAI's capacity to accelerate progress spans multiple scientific fields, offering breakthroughs in hydrogen fusion, matrix manipulation, and antibody generation [2].

However, academia's slow adoption of GAI could lead to a widening gap in research productivity compared to industry [7]. Studies suggest that declining research productivity may threaten the long-term viability of academic research, especially as funding pressures intensify. The traditional strategy of increasing the number of researchers without enhancing productivity is unsustainable, pointing to the need for innovative approaches to research design and methodology.

The integration of GAI into academic research could help bridge this gap and improve productivity, enabling academia to catch up with industry and drive further innovation. GAI, particularly through models like ChatGPT, offers promising pathways to improve research efficiency while possibly maintaining or enhancing reliability. This could be crucial in supporting scientific advancements and addressing global challenges, such as climate change and emerging health threats [8]. By embracing GAI and integrating it into research methodologies, academia can ensure it remains a key contributor to economic and societal development, fostering a brighter future through increased research productivity.

GAI has demonstrated substantial promise as a potent tool [9, 10] to increase cost-effectiveness through scalability [11], parallelization [11], speed [9, 12], and increased efficiency [13]. Its application within research domains is predominantly observed in facilitating literature reviews, where it excels in distilling key points, overarching themes

*Corresponding author:** Frédéric Pattyn, Department of Business Informatics, Ghent University, Belgium. Email: Frederic.Pattyn@ugent.be

from extensive texts [14] and identifying patterns [9]. However, concerns regarding the generation of fake data [15] or fabricated [10, 13, 14] and misleading [16] information have been noted. Such issues rightfully raise concerns about the implications for academic integrity when integrating GAI into research methodologies [17].

Qualitative research emerges as a particularly apt domain for the application of GAI, offering substantial support to researchers in their analytical endeavors [16]. This suitability stems from several key factors: firstly, the data involved are generated by humans, thereby mitigating concerns related to data quality; secondly, as cutting-edge large language models, these AI systems exhibit unparalleled proficiency in interpreting human language, reflecting their revolutionary potential in understanding complex linguistic constructs [11, 14, 18]; and thirdly, their role is confined to aiding language analysis rather than contributing to the study's composition [14]. These attributes collectively position GAI as a valuable tool in deductive qualitative research, enhancing the depth and breadth of linguistic analysis without compromising the integrity of the research process.

The objective of this study is to examine how GAI can improve deductive qualitative research methodologies, aiming to provide researchers with enhanced tools and insights. Our research focuses on assessing whether integrating GAI can uphold, or even surpass, the current standards of reliability in deductive qualitative analysis. Through a mixed-methods approach, we investigated practices, challenges, and gathered expert recommendations on integrating GAI in deductive qualitative research within the field of Computer Science. This involved semi-structured interviews with five domain experts, which allowed us to delve into the opportunities, risks, and limitations of GAI in qualitative research. Furthermore, to enrich our analysis, we compared GAI-generated replication results with the results from an ongoing study that utilized only human-driven coding. These collective insights inform the potential of GAI to contribute meaningfully to deductive qualitative research, setting the stage for further exploration in this evolving field.

2. Interviews

We conducted five interviews using a convenience sampling method [19]. These interviews provided insights into how GAI supports coding in qualitative research within the Computer Science domain. Semi-structured interviews were carried out via digital platforms, each lasting between 10 and 25 min. We followed the general interview guidelines [20], to ensure consistency and reliability in our data collection. In accordance with our ethics policy, we ensured that all participants and their shared information would remain confidential. We performed a thematic analysis [21] on the verbatim transcripts of the interviews.

Table 1 provides details about the participants' demographics (i.e., P1–P5), including their roles, years of experience, and perspectives on utilizing GAI to support coding in qualitative research within the Computer Science domain.

The interviews indicate a general interest in employing GAI for coding tasks in qualitative Computer Science research, but with close supervision from human researchers. Although GAI cannot fully replace human involvement, it can effectively assist in deductive qualitative research, particularly when the coding options are predefined, leading to increased productivity. The feedback underscores the importance of empirical studies to confirm GAI's effectiveness. In a deductive context, GAI can be valuable for coding, while in inductive, more creative, or less analytical challenging studies, it can be used to double-check results, a rare

Table 1
Demographic details of interview participants (age, gender, experience, expertise)

ID	Role	Years of experience	Point of view
P1	PhD researcher (PhD candidate)	6 years (Computer Science)	Can't fully replace humans with Generative AI, but could support coding. AI could be better suited when the options are already predefined. Productivity in terms of measures of time, AI could be better in that sense. Results need to be in conjunction with an empirical user study.
P2	Researcher (PhD)	19 years (Computer Science)	Must start exploring the possibility, but with caution. Qualitative research is complex and will require a human researcher to take a much more involved role. When doing deductive studies, genAI could play a coding role, leveraging the technology's strengths. For inductive ones, genAI could be used for double-checking the findings, which is not happening in qualitative research work often, due to the highly manual coding process.
P3	Senior Requirements Engineer & Project Manager (PhD)	10 years (Computer Science)	GAI has the potential of supporting researchers a lot in their recoding/reconciliation tasks. We must take our initial steps in leveraging GAI in research where it really could bank on its strengths. I confirm that ChatGPT (and other DL/AI tools) show to be a lot more consistent and systematic than humans.
P4	Professor	18 years (Computer Science)	AI is taking a huge space in SE confs and ethical aspects of research/er either in qualitative or quantitative studies, are being addressed. I have never used ChatGPT, so it would not be prudent for me to express a proper opinion, but I'm looking into the possibilities of starting research on Generative AI for SE.
P5	Professor	19 years (Computer Science)	Currently don't use AI tools for as part of their research or as part of any qualitative analysis. Has no argument against the use of it, although highly enjoy the personal in-depth aspect of coding. Does believe it holds potential for deductive, repetitive tasks that require less creativity or analysis.

step due to the manual nature of the process. GAI's consistency and systematic approach are noted, indicating its potential to support researchers with recording and reconciliation tasks.

The interviewees also touched on ethical considerations in qualitative and quantitative studies involving GAI, emphasizing the growing focus on these issues in software engineering conferences.

3. Pilot Study Design

In confronting the complexities inherent in methodological processes, delineating the procedural steps [22] offers an opportunity to identify potential enhancements where the trade-off of a tolerable risk to quality is deemed acceptable [16]. To concentrate the breadth of this investigation, attention was exclusively directed towards the qualitative research domain [23], with a particular emphasis on scenarios necessitating data coding [24, 25] of a study [26] consisting of 122 responses, or lines of data.

In this study, the authors relied entirely on human input for the coding process. This involved mapping tasks—such as advertising budget, business case analysis, cost estimation, and requirements prioritization—into one of six predefined domains: executive leadership, product strategy & planning, engineering & development, marketing, sales, and customer support & success. To assist the coders, each task was accompanied by a brief description to ensure a consistent understanding and interpretation across the team.

We prepared the data and recruited coders. The original study selected three experts and three non-experts to carry out the coding. As illustrated in Figure 1 (created by the authors), the coders completed the task without intermediate reassessment or data sharing with researchers. After the coding was completed, the researchers processed the data, calculating metrics like Cohen's Kappa to measure inter-rater reliability, and used ChatGPT to double-check for inconsistencies across the various domains.

The researchers disclosed the original dataset along with information regarding the time taken by the coders to complete the task, their costs, and their respective throughput times.

Research methodologies that incorporate data coding often require navigating a delicate balance among budget constraints, project timelines, and research rigor. This balance stems from the fact that more stringent methodologies often need more individuals to independently code data, not to mention the variability in coder expertise (non-experts versus experts). Therefore, researchers aiming for high data fidelity must allocate a significant part of their project's budget to recruiting, supervising, and compensating coders, as Figure 2 (created by the authors) illustrates.

The primary research question (RQ1) of this study is: "Can coding driven by GAI accomplish designated tasks with greater financial and temporal efficiency while preserving, or even enhancing, inter-coding reliability (measured by Cohen's Kappa) compared to traditional methods reliant on human effort?" This question stems from the hypothesis that the major cost factors in qualitative research might be effectively reduced by employing one or several GAI coders. This approach seeks to maintain academic rigor and quality while decreasing both the costs and time required for scholarly inquiries, as demonstrated in Figure 3 (created by the authors).

The secondary research question (RQ2) explores the potential advantages of employing domain experts for coding tasks over non-experts or GAI alternatives. Additionally, this study investigates (RQ3) whether a hybrid approach, integrating both human expertise and GAI capabilities, could yield superior outcomes compared to any singular coding methodology. The research design configuration is shown in Figure 4 (created by the authors).

This study replicates the original research using the original data but replaces the human coding effort with three GAI coders. We extend the effort, cost, and throughput time data from the original

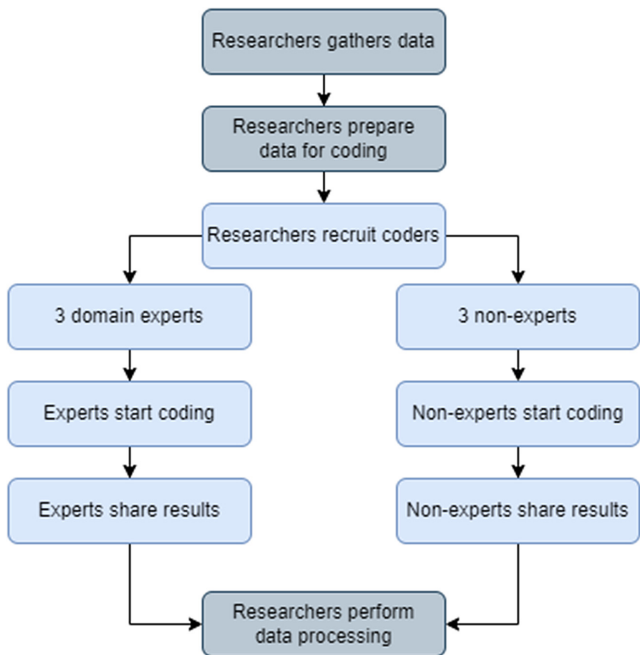**Figure 1**
Human coding process workflow

Figure 2
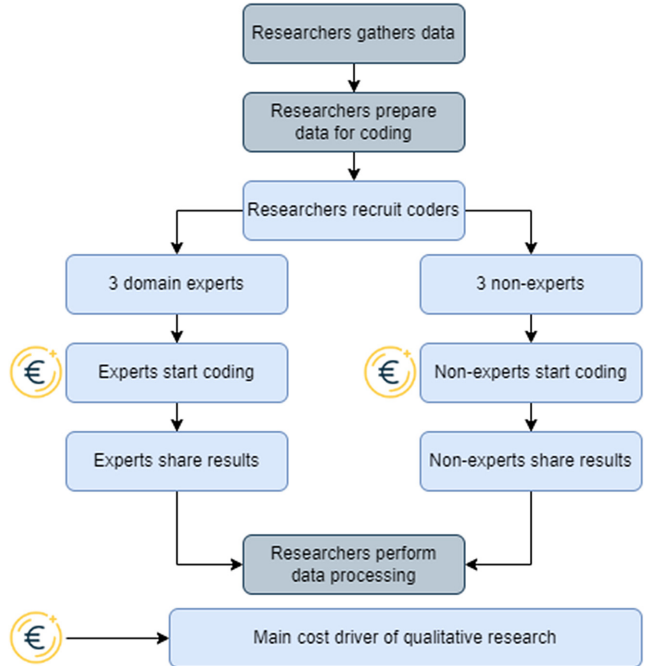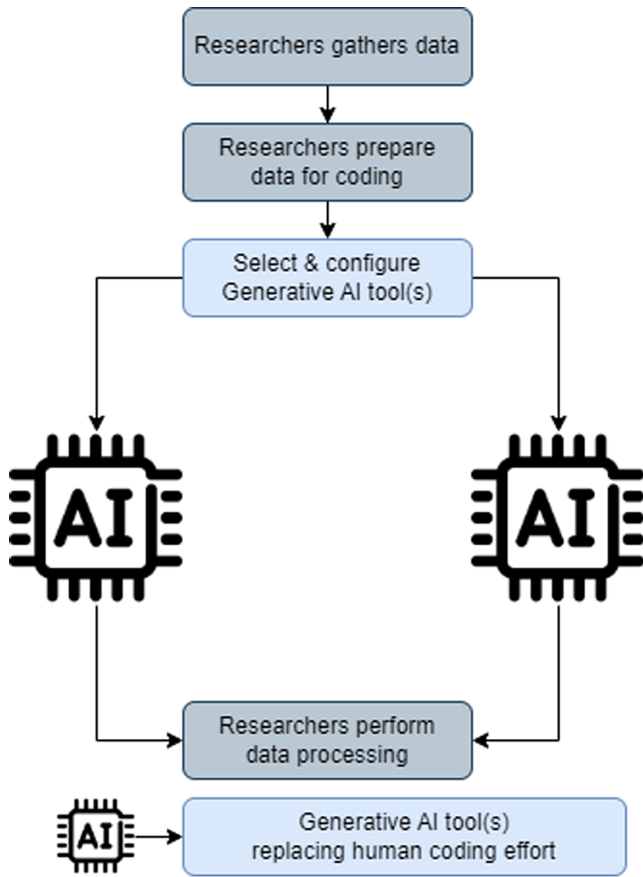**Cost analysis of human-coded process workflow**

Figure 3
Efficiency improvements through GAI-driven coding process workflow

Figure 4
Pilot study design and methodology layout

study with the data generated from this study. These combined data are used to calculate inter-coding reliability.

3.1. Control group 1: Non-experts

The non-expert control group consists of three individuals [27] sourced through the website Upwork. Another option for recruiting non-experts for coding purposes could be Google's Mechanical Turk. All data for this group have been gathered from the authors involved with the original research [26].
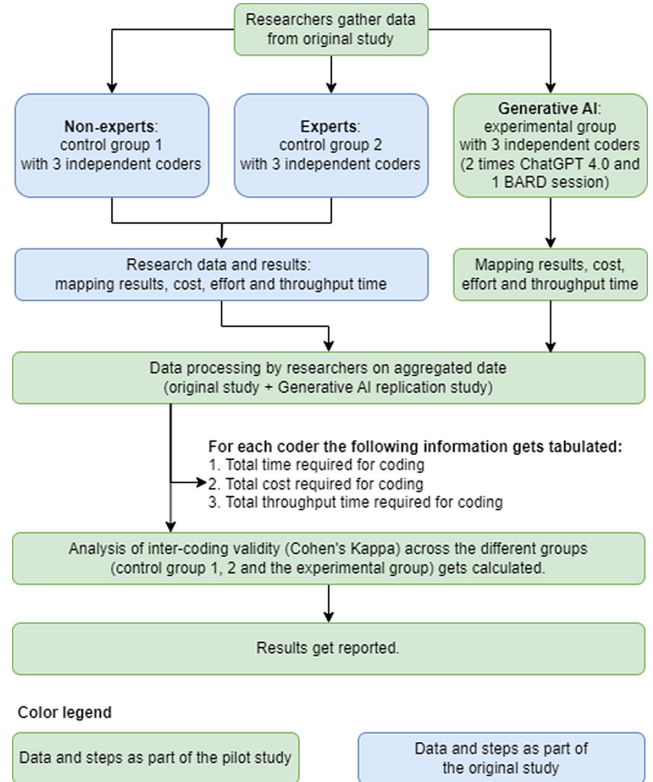
3.2. Control group 2: Experts

The expert control group comprises three individuals selected through a rigorous process based on their domain-specific expertise and possession of a doctoral degree. Recruiting these experts presented significant challenges and was achieved through targeted networking. All data for this group were also collected from the authors involved with the original research [26].

3.3. Experimental group 3: Generative AI

For the experimental group, we chose to use ChatGPT 4, a paid version that allows for the opening of two independent coding sessions, thus preventing contamination through learning effects [17, 18]. This lack of contamination was a concern with earlier versions like ChatGPT 3 [28]. ChatGPT 4 is particularly suited to

complex language interpretations due to its advanced capabilities with over 175 billion parameters [9]. Additionally, to enhance the autonomy of the GAI coders, we selected Bard [18, 29] from Google as an additional tool. To ensure consistency across different platforms, we meticulously defined the context and the expected output format (e.g., table design) for all coders using an identical prompt:

> Considering software product management and business context, provide me a table with 2 columns. The first column is all the activities, and the second column is one of the predefined domains that you must choose maps the most closely considering the provided context. List of activities: ... List of domains ...

3.4. Preparatory activities: Not accounted for in final analysis

It is important to note that the time spent standardizing the data preparation phase for all groups to a consistent duration of one h, as well as the time required to recruit and orient the coders, will not be considered in the final analysis. These preparatory activities, exclusive to the control groups comprising human coders, often demand substantial time investment and can significantly influence the overall throughput time of the research effort.

4. Results

4.1. Total human effort (hours)

Analyzing the outcomes from different setups, as Table 2 illustrates, a significant variation in coding effort emerges. GAI coders completed the task in just 8 min, a stark contrast to the 4 h

Table 2
Comparative analysis for human and GAI coders (effort, cost, and throughput time)

Scenario	Total effort (hours)	Cost (in €)	Throughput time (days)
Control: Non-experts			
Person 1 (2)	2 h 15 min	€35.00	8 days
Person 2 (4)	1 h 35 min	€20.00	2 days
Person 3 (5)	1 h 5 min	€15.00	3 days
Control: Experts			
Person 4 (1)	1 h 7 min	€35.00	2 day
Person 5 (3)	2 h 12 min	€65.00	4 days
Person 6 (6)	45 min	€25.00	6 days
Generative AI			
ChatGPT 4 (1)	5 min	€0.72	<1 day
ChatGPT 4 (2)	5 min	€0.00	<1 day
BARD (1)	3 min	€0.00	<1 day

Figure 5
Efficiency comparison of throughput times by coding group

and 55 min required by non-experts and 4 h and 4 min by experts. This demonstrates a 17% reduction in human effort when using experts compared to non-experts. Overall, GAI coders achieved an average efficiency gain four times greater than that of human coders: 3.59 times more efficient than experts and 4.34 times more than non-experts.

4.2. Total human effort (cost savings)

The time required to complete tasks significantly influences the coding process's overall cost. As detailed in Table 2, expert coders were nearly twice as costly as non-experts. Employing BARD, which is free, resulted in a 100% cost saving compared to experts, equating to savings of 125 euros, and 70 euros compared to non-experts. The cost analysis for ChatGPT 4.0 is more complex, with an annual license fee of approximately 240 euros. Considering the daily usage and multiple uses of a single license, attributing a day's worth of the license fee at approximately 0.72 euros is reasonable. This cost, slightly higher than BARD's, still achieves notable savings despite the annual fee.

4.3. Total throughput time (days)

Table 2 also highlights a clear performance gain in throughput time when comparing the experimental group with the control groups. The GAI coders averaged less than a day per coder, significantly faster than the 4.33 days for non-experts and 4 days for experts. This comparison underscores the substantial efficiency of GAI coders, who can operate in parallel, unlike their human counterparts, whose schedules depend on sequential availability.

Figure 5 (created by the authors) illustrates the promising results of utilizing AI coders in parallel, showing that the throughput time for the experimental group can be approximately half a day. This represents an efficiency gain of 18 times faster than the non-expert group and 12 times faster than the expert group.

4.4. Results on inter-coding reliability

We chose Cohen's Kappa to assess the quality of inter-coding reliability [30]. As per Table 3, the control groups generally showed minimal levels of agreement, with exceptions at both ends of the spectrum: 0.41 and 0.11. The experimental group's Cohen's
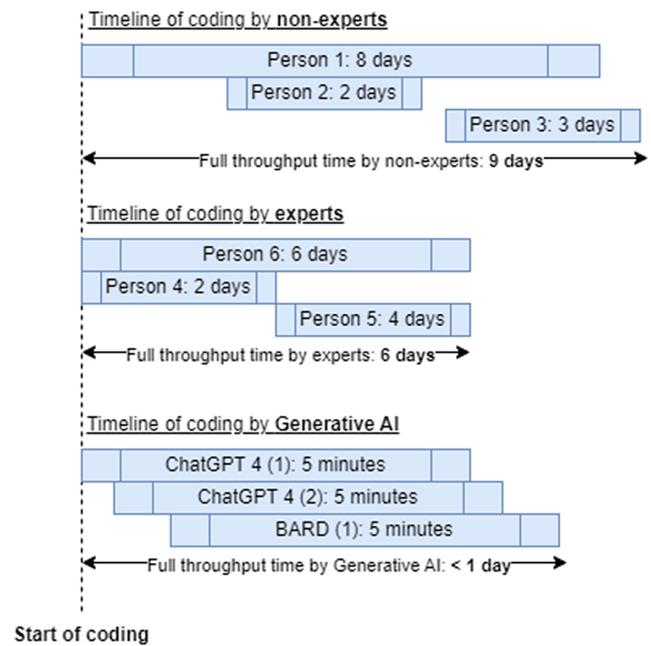
Table 3
Cohen's Kappa inter-coder reliability measures across groups

Group	Inter-coding reliability (Cohen's Kappa)
Control: Non-experts	Person 1 × Person 2 = **0.41**
	Person 1 × Person 3 = 0.32
	Person 2 × Person 3 = 0.35
	Person 1 × Person 2 × Person 3 = 0.18
Control: Experts	Person 4 × Person 5 = 0.37
	Person 4 × Person 6 = 0.26
	Person 5 × Person 6 = 0.24
	Person 4 × Person 5 × Person 6 = 0.11
Generative AI	ChatGPT 4 (1) × ChatGPT 4 (2) = **0.77**
	ChatGPT 4 (1) × BARD (1) = **0.52**
	ChatGPT 4 (2) × BARD (1) = **0.65**
	ChatGPT 4 (1) × ChatGPT 4 (2) × BARD (1) = **0.48**
Hybrids	ChatGPT 4 (1) × Person 1 = 0.37
	ChatGPT 4 (1) × Person 2 = **0.48**
	ChatGPT 4 (1) × Person 3 = 0.37
	ChatGPT 4 (1) × all non-experts (Person 1,2,3) = 0.16
	ChatGPT 4 (1) × Person 4 = **0.54**
	ChatGPT 4 (1) × Person 5 = **0.45**
	ChatGPT 4 (1) × Person 6 = **0.30**
	ChatGPT 4 (1) × all experts (Person 4,5,6) = 0.09

Kappa values were consistently higher, with the lowest value (0.48) surpassing the best control group value, and the highest (0.77) approaching strong agreement levels.

Examining hybrid scenarios revealed that all individual pairing values improved, moving from minimal to weak agreement levels. Cohen's Kappa values were on average higher when combining

GAI with experts (0.43) than with non-experts (0.40), although the differences were slight. However, it's important to note that BARD (1) experienced instances of "hallucination" [29], where responses did not fully align with the structured prompt, which likely contributed to the lower Cohen's Kappa values observed in datasets processed with BARD, compared to those analyzed with ChatGPT.

5. Limitations

The principal limitation of this pilot study lies in its scope. Although the results are promising, the volume and complexity of the data evaluated were relatively modest. Over time, we expect the development of new GAI frameworks that could offer enhanced support for research endeavors [31], potentially addressing some of the current shortcomings. Notably, we detected instances of 'hallucination' within this study—where the GAI generated superficial, inaccurate, or erroneous content [32]. This issue, highlighted in previous studies, could significantly impact the reliability of coding in qualitative research.

Ethical considerations also play a crucial role, especially with the increased incidence of reported issues since 2012 [2]. The use of GAI in scholarly research raises concerns about fair use [13], limited knowledge base [11, 16], and a lack of context [11, 16], which could lead to biased outputs [11, 13]. Such biases, even trivial ones like a preference for certain numbers [33], underscore the indispensable need for human oversight in critical decision-making processes [28].

The broader scientific community faces a 'reproducibility crisis' [34] that could be exacerbated by the inherent limitations of GAI-driven language models [28]. This underscores the necessity for the developers and providers of such technologies to implement robust measures, including comprehensive pre-release evaluations by independent third parties [35].

6. Research Agenda

Considering the limited size of the pilot study, further investigation and research are required to confirm the following coding tactics when refining the power of GAI coding for larger datasets:

1. Force the GAI to up his certainty level during the mapping process by adding the following to the prompt: "leave pairs empty if than 40% certain." You can make this as strict as you think is appropriate for your research.
2. Introduce internal intermediate validation (per 100 pairs for instance) and leverage the self-learning capabilities of the GAI systems to improve internal consistency. This can be done by using the following intermediate prompt: "Consider all combinations that you've validated so far, if you take them all into account, are there any combinations of which you would like to change the output value of the third column to further improve internal consistency and validity? If so, also provide the reason why in the fourth column." This last column gives you the opportunity to validate and overrule the suggested changes.
3. Introduce external intermediate validation (per 100 pairs for instance) and leverage the self-learning capabilities of the GAI systems to improve overall internal consistency. Here creates a new independent chat (ChatGPT 4.0 by OpenAI) or prompt (Bart by Google) where for the prompt the output of the coders (human and/or GAI) is used as the input for this one (column 1: category; column 2: domain by coder 1; column 3: domain by coder 2). This can be done by using the following

intermediate prompt: Consider all combinations provided below, if you take them all into account, for those combinations coder 1 and coder 2 don't agree on the defined mapping, what domain would you domain would you suggest, add it to a fourth column, and the reason why, in the fifth column. All of this is to improve the general overall internal consistency and validity.

Having used the following tactics [26], a Cohen's Kappa value of 0.94 was achieved, and all inconsistencies found by ChatGPT were approved.

7. Summary

This study sought to determine whether GAI coders could replace human coders to enhance the efficiency of scholarly work in terms of cost and time, without compromising inter-coding reliability. The dataset analyzed comprised 122 items, each needing classification into one of six predefined options. The experimental group, consisting of three independent GAI coders, significantly outperformed both control groups in terms of coding effort (four times less) and throughput time (15 times faster), achieved by leveraging parallel processing. These findings support the first part of Research Question 1 (RQ1).

Regarding inter-coding reliability, the experimental group also excelled, with Cohen's Kappa values ranging from 0.48 to 0.77, indicating a moderate level of agreement. In contrast, both control groups showed significantly lower agreement levels, validating RQ1 in full.

Research Question 2 (RQ2) assessed the effectiveness of expert versus non-expert coders. The results showed minimal advantage in favor of experts, which does not justify their higher cost, especially with larger datasets. This slight benefit in coding speed and throughput time does not offset the higher rates and recruitment efforts required for experts, leading to the rejection of RQ2.

For Research Question 3 (RQ3), the evidence suggests that a hybrid coding approach, integrating human expertise and GAI capabilities, is superior. Not only does it reduce costs and average throughput time, but it also improves inter-coder reliability, particularly when combining GAI with expert human coders, thus affirming RQ3.

8. Conclusion

The findings of this study indicate that replacing human coding labor, both novices and experts, with GAI solutions like ChatGPT is an effective first step toward integrating GAI in qualitative Computer Science research. This approach, particularly suitable for deductive methodologies where potential coding responses are predefined, demonstrates GAI's viability and cost-effectiveness. This transition could significantly reduce research costs and expedite timelines without sacrificing the quality of inter-coding reliability.

Although experts generally deliver faster and more efficient results than novices, the cost-benefit analysis does not favor them due to negligible differences in Cohen's Kappa values. A hybrid approach, combining the efficiency of GAI with the nuanced judgment of human experts, emerges as the most promising strategy [9]. This method not only optimizes cost-effectiveness but also enhances the overall integrity and value of the research.

8.1. Implications for stakeholders

For academic institutions and research organizations, this finding implies a pathway to streamline research processes while

maintaining quality. By integrating GAI, research teams can optimize resources and potentially reallocate funds to other critical areas, enhancing overall productivity. For individual researchers, this hybrid approach may offer new opportunities for skill development, allowing them to focus on higher-level analytical tasks, or coding tasks that can't be done reliably yet by GAI while GAI handles more repetitive, and easier coding work.

8.2. Limitations

It is important to acknowledge that the study was conducted on a relatively modest coding task, which may not reflect the complexities of larger datasets, different research contexts, or methods. Therefore, the scalability of GAI in handling more significant challenges requires further research, both in terms of increased coding-task volume and complexity. Additionally, while GAI shows promise in coding, it may not fully replace human expertise, especially in tasks that require deeper domain knowledge or subjective interpretation.

8.3. Future scope

Future research should delve into the broader applications of GAI in coding, with a focus on its effects on larger datasets and complex research designs, including inductive methodologies. This exploration should include innovative methods to use AI while ensuring that human oversight remains intact for accuracy and validity. Investigating the ethical implications of AI, particularly around bias and data security, is vital to ensure responsible use. Further research should also examine the potential challenges or roadblocks in replacing human coders with AI, emphasizing that a shift of this magnitude requires more than a single study. There should be a focus on refining the hybrid approach, achieving a balanced blend of AI and human expertise, and applying it across various research domains.

Ethical Statement

This study does not contain any studies with human or animal subjects performed by any of the authors.

Conflicts of Interest

The author declares that he has no conflicts of interest to this work.

Data Availability Statement

The data that support the findings of this study are not publicly available due to privacy concerns. However, anonymous data are available on request from the corresponding author upon reasonable request.

References

[1] Pattyn, F. (2023). The hidden costs of ignoring cash flow: A call for strategic requirements prioritization at startups during an era of rising interest rates. In *2023 IEEE 31st International Requirements Engineering Conference Workshops,* 375–378.

[2] Maslej, N., Fattorini, L., Brynjolfsson, E., Etchemendy, J., Ligett, K., Lyons, T., . . ., & Perrault, R. (2023). Artificial intelligence index report 2023. *arXiv Preprint:2310.03715.*

[3] Bilgram, V., & Laarmann, F. (2023). Accelerating innovation with generative AI: AI-augmented digital prototyping and innovation methods. *IEEE Engineering Management Review*, *51*, 18–25.

[4] Estermann, T., Bennetot Pruvot, E., Kupriyanova, V., & Stoyanova, H. (2020). *The impact of Covid-19 crisis on university funding in Europe*. Belgium: European University Association ASBL.

[5] Krenn, M., Landgraf, J., Foesel, T., & Marquardt, F. (2023). Artificial intelligence and machine learning for quantum technologies. *Physical Review A*, *107*(1), 010101.

[6] Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., . . ., & Potapenko, A. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583–589.

[7] Bloom, N., Jones, C. I., Van Reenen, J., & Webb, M. (2020). Are ideas getting harder to find? *American Economic Review*, *110*(4), 1104–1144.

[8] Nolan, A. (2024). *Artificial intelligence in science: Challenges, opportunities and the future of research*. France: Fondation Ipsen.

[9] Mijwil, M. M., Hiran, K. K., Doshi, R., Dadhich, M., Al-Mistarehi, A. H., & Bala, I. (2023). ChatGPT and the future of academic integrity in the artificial intelligence era: A new frontier. *Al-Salam Journal for Engineering and Technology*, *2*(2), 116–127.

[10] Rahman, M. M., Terano, H. J., Rahman, M. N., Salamzadeh, A., & Rahaman, M. S. (2023). ChatGPT and academic research: A review and recommendations based on practical examples. *Journal of Education, Management and Development Studies*, *3*(1), 1–12.

[11] Kalla, D., & Smith, N. (2023). Study and analysis of ChatGPT and its impact on different fields of study. *International Journal of Innovative Science and Research Technology*, *8*(3), 827–833.

[12] Wang, Y., Pan, Y., Yan, M., Su, Z., & Luan, T. H. (2023). A survey on ChatGPT: AI-generated contents, challenges, and solutions. *IEEE Open Journal of the Computer Society*, *4*, 280–302.

[13] Meyer, J. G., Urbanowicz, R. J., Martin, P. C., O'Connor, K., Li, R., Peng, P. C., . . ., & Moore, J. H. (2023). ChatGPT and large language models in academia: Opportunities and challenges. *BioData Mining*, *16*(1), 20.

[14] Zhu, J. J., Jiang, J., Yang, M., & Ren, Z. J. (2023). ChatGPT and environmental research. *Environmental Science & Technology*, *57*(46), 17667–17670.

[15] Haman, M., & Školník, M. (2023). Using ChatGPT to conduct a literature review. *Accountability in Research*, 1–3.

[16] Khlaif, Z. N., Mousa, A., Hattab, M. K., Itmazi, J., Hassan, A. A., Sanmugam, M., & Ayyoub, A. (2023). The potential and concerns of using AI in scientific research: ChatGPT performance evaluation. *JMIR Medical Education*, *9*, e47049.

[17] Wu, T., He, S., Liu, J., Sun, S., Liu, K., Han, Q. L., & Tang, Y. (2023). A brief overview of ChatGPT: The history, status quo and potential future development. *IEEE/CAA Journal of Automatica Sinica*, *10*(5), 1122–1136.

[18] Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*, *3*, 121–154.

[19] Baltes, S., & Ralph, P. (2022). Sampling in software engineering research: A critical review and guidelines. *Empirical Software Engineering*, *27*(4), 94.

[20] Patton, M. Q. (2002). *Qualitative research & evaluation methods*. USA: Sage.

[21] Kiger, M. E., & Varpio, L. (2020). Thematic analysis of qualitative data: AMEE guide no. 131. *Medical Teacher*, *42*(8), 846–854.

[22] Guleria, A., Krishan, K., Sharma, V., & Kanchan, T. (2023). ChatGPT: Ethical concerns and challenges in academics and research. *The Journal of Infection in Developing Countries*, *17*(09), 1292–1299.

[23] Basit, T. (2003). Manual or electronic? The role of coding in qualitative data analysis. *Educational Research*, *45*(2), 143–154.

[24] Elliott, V. (2018). Thinking about the coding process in qualitative data analysis. *Qualitative Report*, *23*(11), 2850–2861.

[25] Williams, M., & Moser, T. (2019). The art of coding and thematic exploration in qualitative research. *International Management Review*, *15*(1), 45–55.

[26] Pattyn, F. (2023). Preliminary structured literature review results using ChatGPT: Towards a pragmatic framework for product managers at software startups. In *2023 IEEE 31st International Requirements Engineering Conference Workshops*, 367–370.

[27] Church, S. P., Dunn, M., & Prokopy, L. S. (2019). Benefits to qualitative data quality with multiple coders: Two case studies in multi-coder data analysis. *Journal of Rural Social Sciences*, *34*(1), 2.

[28] Lund, B. D., Wang, T., Mannuru, N. R., Nie, B., Shimray, S., & Wang, Z. (2023). ChatGPT and a new academic reality: Artificial intelligence-written research papers and the ethics of the large language models in scholarly publishing. *Journal of the Association for Information Science and Technology*, *74*(5), 570–581.

[29] Gupta, M., Akiri, C., Aryal, K., Parker, E., & Praharaj, L. (2023). From ChatGPT to threatGPT: Impact of generative AI in cybersecurity and privacy. *IEEE Access*, *11*, 80218–80245.

[30] McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica*, *22*(3), 276–282.

[31] Koubaa, A., Boulila, W., Ghouti, L., Alzahem, A., & Latif, S. (2023). Exploring ChatGPT capabilities and limitations: A survey. *IEEE Access*, *11*, 118698–118721.

[32] Sallam, M. (2023). The utility of ChatGPT as an example of large language models in healthcare education, research and practice: Systematic review on the future perspectives and potential limitations. *MedRxiv Preprint: 2023.02.19.23286155*.

[33] Azaria, A. (2022). ChatGPT usage and limitations. *HAL Open Science: 03913837*.

[34] Gundersen, O. E. (2023). Improving reproducibility of artificial intelligence research to increase trust and productivity. *Artificial Intelligence in Science*, 262.

[35] Soice, E. H., Rocha, R., Cordova, K., Specter, M., & Esvelt, K. M. (2023). Can large language models democratize access to dual-use biotechnology? *arXiv Preprint:2306.03809*.