Open Peer Review on Qeios

Can ChatGPT code the technical part of a Bachelor's Thesis in Informatics?

Solin Amin[1], Johan Hellström[1], Arend Hintze[1]

1 Dalarna University College

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.

Abstract

This study delves into the potential of Large Language Models (LLMs), specifically ChatGPT by OpenAI, in assisting with computer code writing for Bachelor's theses in the field of Information Systems at Dalarna University, Sweden. While LLMs have shown promise in various applications, their efficacy in academic coding tasks remains underexplored. Through a pilot study, we investigated the extent to which ChatGPT can support or even take over the coding aspect of a Bachelor's thesis. Our findings indicate that ChatGPT can expedite the coding process and empower students to undertake technical analyses. However, the success of this collaboration is contingent on the student's ability to engage in a "critical dialog" with the model, assessing its outputs and seeking refinements as needed. While our results are context-specific and tied to the competencies of the involved students, they underscore the potential of LLMs in academic coding tasks and highlight the need for broader investigations in this domain.

Solin Amin[†], **Johan Hellström**[†], and **Arend Hintze**[*]

Dalarna University, Högskolegatan 2, Falun, 791 88, Sweden

[*]Corresponding Author. Email: ahz@du.se

[†]These Authors contributed equally

1. Introduction

While large language models (LLMs) (Radford, Wu, Child, Luan, Amodei, Sutskever et al., 2019) were already recognized and utilized, ChatGPT, introduced by OpenAI, has prominently showcased their practicality to the broader audience (Okan, 2023; Leiter, Zhang, Chen, Belouadi, Larionov, Fresen and Eger, 2023). Beyond the myriad of potential applications and the associated ethical and societal implications, our focus is on the efficacy of ChatGPT (serving as a representative for other LLMs) in assisting with computer code writing for a Bachelor's thesis. In particular, we present the findings of a pilot study conducted at Dalarna University, Sweden, in *systemvetenskap* -- best translated as Informatics or Information Systems Studies. Our primary inquiry was whether ChatGPT could assist or fully handle the coding aspect of a Bachelor's thesis in this discipline. While the succinct answer is "Yes, it expedited the process and empowered students to undertake a highly technical analysis," the methodologies underpinning these conclusions are equally compelling to explore.

With every new phenomenon comes new scientific challenges. Obviously, the developers of ChatGPT trained their LLM using deep learning (Goodfellow, Bengio and Courville, 2016), which implies that the differences between expected answers and answers given by the LLM were computed during training. These discrepancies, termed "loss," form the foundation for the backpropagation algorithm (Zhang, Bengio, Hardt, Recht and Vinyals, 2021). Moreover, juxtaposing accuracies from test and training datasets offers insights into the model's generalization capabilities. Nonetheless, while these metrics are definitive within the machine learning realm, they offer limited insights into the broader applicability of an LLM, and they scarcely address the question we have posed in this context.

Evaluating the effectiveness of a tool, particularly one that offers cognitive assistance, can be likened to gauging the ease of delegating a task to another individual. If the task cannot be fully delegated, it is akin to determining the extent of assistance from a collaborator, such as a teacher or mentor. Notably, the success of this collaboration varies based on the nature of the task and the individual's proficiency. For instance, an expert coder can swiftly address minor issues, whereas a beginner might struggle to articulate the problem accurately. While challenges in collaboration have been extensively explored in the context of human-human interactions, ChatGPT, being a relatively new tool, lacks established research methodologies for assessing human-LLM interactions. The chat interface of human-LLM interaction might misleadingly suggest that the LLM is analogous to another human. While this could be a reasonable assumption, it is not backed by concrete data. Given the ongoing debates about whether LLMs possess cognitive capabilities comparable to humans, it is premature to assume that their interactions with humans can be evaluated using the same methods as human-human interactions. Given this data gap, our approach must be treated as a preliminary study aimed at data collection, with the understanding that our exploration of this subject is in its early stages[1].

As educators, we are acutely aware of the educational (Abd-Alrazaq, AlSaad, Alhuwail, Ahmed, Healy, Latifi, Aziz, Damseh, Alrazak, Sheikh et al., 2023) and ethical challenges (Kasneci, Seßler, Küchemann, Bannert, Dementieva, Fischer, Gasser, Groh, Günnemann, Hüllermeier et al., 2023) associated with the use of artificial intelligence (AI) tools. In Sweden, the material submitted for student evaluation and certification must be the student's work, excluding contributions from AI tools[2]. While this seems like a sound regulation on the surface, the absence of a clear definition of AI casts doubt on the enforceability and relevance of such a rule. For instance, where do we draw the line between the legitimacy of a spell checker's red underlining in writing software and a paragraph refined by ChatGPT? Similarly, why is seeking a mentor's assistance to identify an acceptable coding error while soliciting the same help from ChatGPT is considered a breach of regulations? Given that these questions remain unresolved and our intention to steer clear of this debate, our focus will be on the potential outcomes of AI assistance or even the entire delegation of specific tasks or sub-tasks to AI.

Determining the extent to which a Large Language Model (LLM) like ChatGPT can assist in coding for every Bachelor's thesis is not feasible. A one-size-fits-all answer is elusive, given the vast diversity in degrees and potential thesis topics. Therefore, our focus narrows to a context we are intimately familiar with the field of "Information Systems." In this domain, students typically delve into computer systems' functionality, integration, reliability, creation, service, and societal impact.

Since we are already curious about ChatGPT, we asked a hypothetical question: to what degree the answers of ChatGTP are gender biased? While this query --or its variants-- aligns with typical "Information Systems" thesis topics, it is essential to note that our primary interest is not the gender bias question itself. Instead, it represents any topic necessitating data collection, analysis, statistical evaluation, and coding. Our core inquiry revolves around gauging the coding support ChatGPT can offer for such thesis projects.

To address this question, two students from the Summer 2023 cohort of their Bachelor's Thesis (Amin and Hell- ström, 2023) sought to leverage ChatGPT in designing an experimental methodology encompassing data collection, analysis, and result evaluation. They also relied on ChatGPT to provide the requisite code for this methodology. However, for a rigorous experiment, it is essential to establish a benchmark or expectation against which outcomes can be measured and conclusions drawn. A significant challenge is that most of the available accounts of such experiences are from something other than peer-reviewed sources. Our prior experiences and these non-peer-reviewed accounts suggest that wholly outsourcing a problem to ChatGPT is not feasible. Expecting a perfectly tailored analysis script from a mere problem statement is unrealistic.

Consequently, the students adopted a "critical dialog" approach. They would present a method and code request to ChatGPT, test the provided code, and, based on its efficacy, either move forward or re-engage with ChatGPT for refinements. Once results were generated, the students critically assessed them. Any inconsistencies triggered another round of dialog with ChatGPT to seek resolutions. A significant caveat here is that the efficacy of this dialog is contingent on the student's ability to assess ChatGPT's outputs critically. Therefore, the outcomes of our study are intrinsically tied to the specific students' competencies and cannot be universally applied. Furthermore, ChatGPT's proficiency will likely fluctuate across diverse thesis topics, adding another layer of complexity to generalizing our findings. Nevertheless, given the scarcity of peer-reviewed literature on this subject, our study offers valuable preliminary insight. It underscores the need for broader investigations across varied topics and emphasizes the correlation between ChatGPT's success and a student's capacity for "critical dialog."

In the following, we will give a report about the "critical dialog", highlight pitfalls and recurrent motifs, and show that ChatGPT, even though ChatGPT had significant shortcomings, could provide meaningful support that improved the quality and speed of the coding part of a fictional Bachelor Thesis project.

2. Methods

2.1. Data Collection

ChatGPT offers an interface with a text box for user input. While this interface can accommodate a wide range of queries, we primarily used it in our study to define problems, seek clarifications, request code, or report code-related errors. Every interaction encompassing our queries and ChatGPT's responses was documented. The entire "critical dialog" has been archived for future reference (Hellström, 2023).

2.2. Code Evaluation

We transferred the code snippets, provided by ChatGPT and written in Python 3, into a Jupyter notebook for execution. All requisite software libraries and datasets were pre-installed to ensure a seamless testing environment. If a code snippet encountered an execution error, we adopted one of two approaches: if the error was evident or within our capacity to rectify, we addressed it directly; otherwise, we relayed the issue to ChatGPT for further guidance.

3. Results

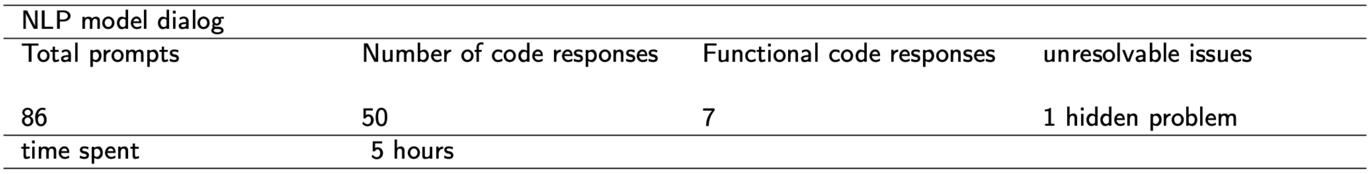
3.1. Critical Dialog

The critical dialog with ChatGPT can be segmented into four distinct phases. Initially, we engaged in a discussion with ChatGPT to outline a potential analysis approach. The primary objective was to identify suitable computational tools and methodologies that could be employed to gauge the extent of potential gender bias in ChatGPT's prior responses. We did not delve deeply into a comprehensive discourse with ChatGPT on the broader nuances of gender analysis (Doughman, Khreich, El Gharib, Wiss and Berjawi, 2021) or the multifaceted definitions of gender and identity (Dev, Monajatipoor, Ovalle, Subramonian, Phillips and Chang, 2021). Instead, we adopted ChatGPT's suggested male-neutral-female spectrum for our analysis.

It is crucial to emphasize that the quality of the generated code is not inherently tied to the relevance or quality of the data analysis it is designed to undertake. For instance, quantifying flies in images captured at varied locations might not have scientific value. However, it poses a coding challenge analogous to many encountered in computational quantitative data analysis projects. This is not to downplay the importance or intrigue of studying gender bias in ChatGPT's responses (or any chatbot, for that matter). However, our focus here is on something other than the scientific merit of the study but instead on the coding required to facilitate the analysis.

ChatGPT suggested various methodologies to discern gender bias in textual data. From the proposed options, we opted for natural language processing using RoBERTa (Liu, Ott, Goyal, Du, Joshi, Chen, Levy, Lewis, Zettlemoyer and Stoyanov, 2019) (a refined variant of BERT (Devlin, Chang, Lee and Toutanova, 2018; Touvron, Lavril, Izacard, Martinet, Lachaux, Lacroix, Rozière, Goyal, Hambro, Azhar et al., 2023)) and word frequency and sentiment analysis via NLTK (Bird, Klein and Loper, 2009). It is worth noting that the specific analytical method chosen is not the primary concern; our main interest was in soliciting code from ChatGPT to execute the selected analyses. We gravitated towards these methods as they align with our assessment of appropriate techniques to tackle the issue at hand.

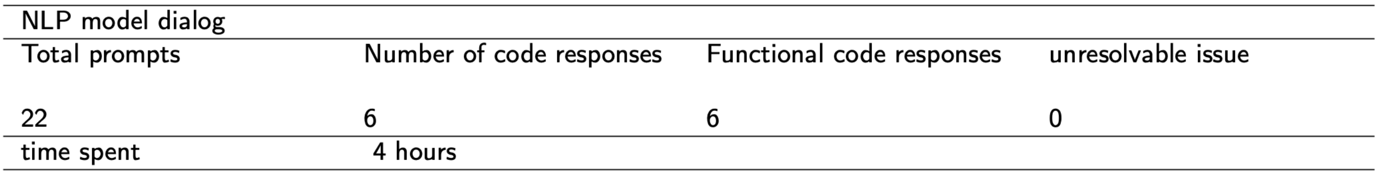
Subsequently, our next move was to request ChatGPT to furnish code that facilitates the utilization of RoBERTa, leveraging a gender bias training set (Dinan, Fan, Wu, Weston, Kiela and Williams, 2020). BERT, akin to ChatGPT, is an advanced deep-learning large language model. While BERT boasts a myriad of inherent capabilities, it is often recommended (ChatGPT concurred) to fine-tune a RoBERTa model derived from BERT for specific tasks. As such, this fine-tuning process was integrated into our workflow. The code snippets provided by ChatGPT frequently exhibited errors, necessitating iterative refinements (refer to Table 1). While ChatGPT managed to rectify all coding-related issues, a conceptual challenge persisted, which went unnoticed. We address this in step 4.



NLP model dialog			
Total prompts	Number of code responses	Functional code responses	unresolvable issues
86	50	7	1 hidden problem
time spent	5 hours		

Table 1. Code accuracy and estimated time

In the third phase of our analysis, we heeded ChatGPT's recommendation to undertake a sentiment and word frequency analysis. Initially, ChatGPT proposed using the Natural Language Toolkit (NLTK) (Bird et al., 2009) to pinpoint potential biases. A sentiment analysis complemented this to discern whether the identified bias leaned towards being positive or negative. While the code generated by ChatGPT executed without hitches, some adjustments were necessary to fine-tune the output and ensure it met our desired specifications.



NLP model dialog			
Total prompts	Number of code responses	Functional code responses	unresolvable issue
22	6	6	0
time spent	4 hours		

Table 2. Code accuracy and estimated time

In the final phase of our analysis, we inquired whether the previously employed analytical methods could be integrated. The underlying rationale was to cross-validate the results from each method, ensuring consistency and reliability. Observe that ChatGPT did not suggest this. Discrepancies between the outcomes would indicate either a flawed methodology or errors during the analysis. When prompted about comparing the methods, ChatGPT introduced an alternative deep learning tool for detecting gender bias named "AI4EU". While we explored this recommendation, we discovered that the tool's API functioned differently than ChatGPT had indicated. This discrepancy is likely attributable to ChatGPT 3.5's training data, which encompasses knowledge up to 2021 but does not extend beyond that year. Consequently, if there were alterations to the "AI4EU" API post-2021, ChatGPT would be unaware of them. Given this limitation, we decided to forgo this approach. Instead, we sought a comparative analysis between the sentiment and natural language processing results and those derived from our fine-tuned RoBERTa model.

Our integrated approach swiftly uncovered a significant oversight in the second phase of our analysis. While fine-tuning RoBERTa, ChatGPT mistakenly presumed that the training data was bifurcated into two categories: male and female. In reality, the training data encompassed three distinct classes: male, female, and neutral. This misjudgment led ChatGPT to generate code tailored for a two-class fine-tuning, resulting in substantial code errors during the fourth phase. A resolution remained elusive despite our attempts to rectify these errors through ChatGPT. Only after an extensive team discussion did we pinpoint the root of the issue. Once we provided ChatGPT with the correct context, the project resumed smoothly, and the results achieved consistency.

Comparison of approaches			
Total prompts	Number of code responses	Functional code responses	unresolvable issue
194	37	17	one API outdated
time spent	7.5 hours		

Table 3. Code accuracy and estimated time

In summary, generating code through a "critical dialog" with ChatGPT spanned approximately 16.5 hours. While this might seem laborious or even daunting, it is essential to juxtapose this duration against the potential time and effort required without the assistance of ChatGPT. The students, nearing the completion of their Bachelor's degree in "Information Systems" and only awaiting their thesis, estimated that in the absence of ChatGPT, they would have required at least double the time to accomplish the task. This estimation stems primarily from their lack of experience and expertise in deep learning and neural network applications. Although RoBERTa is well-documented, it undeniably stands as an advanced tool. In our academic setting, we introduce deep learning at the Master's level. Consequently, the students' time projection also accounts for the need to explore and grasp an entirely novel technology. In summation, not only did ChatGPT expedite the coding process, but it also furnished a solution that surpassed the students' initial coding proficiency.

4. Discussion

As Large Language Models (LLMs) continue to advance in capability, they usher in a plethora of questions, particularly within the realm of education. The balance between the potential benefits and inherent risks remains ambiguous. While LLMs can be seamlessly integrated into pedagogical practices, enhancing learning outcomes, they simultaneously present students with tempting avenues to circumvent genuine effort. However, before delving into these complexities, it is imperative to first assess the proficiency of LLMs, with our focus being on ChatGPT. Such an evaluation, though, introduces its own set of challenges, encompassing considerations like the domain of application, educational level, examination context, and more. Given our limited experience with LLMs and the scarcity of peer-reviewed literature offering established methodologies for such inquiries, we opted to embark on this pilot study. We aimed to gauge ChatGPT's efficacy in aiding code development for a Bachelor's Thesis in "Information Systems".

Our study yielded several noteworthy findings. Firstly, ChatGPT's capabilities, while impressive, are not yet at a level where one can fully delegate the tasks of programming and computational analysis. Instead, a mode of interaction we term as "critical dialog" emerged as the most effective approach. In this method, ChatGPT is iteratively prompted to address the problem at hand, and the generated code is subsequently tested. If issues arise, the process is reiterated. The efficacy of this approach is closely tied to the student's proficiency. While our resources did not permit testing across varied student skill sets, the frequent instances where students rectified the code indicate that a foundational programming knowledge is essential, and advanced skills could prove beneficial.

Secondly, we encountered a significant technical oversight on ChatGPT's part: it incorrectly assumed two classes instead of three while fine-tuning the RoBERTa model. Despite our considerable efforts, this error remained unresolved by ChatGPT. This is different from saying that, given more time, ChatGPT would not have rectified the issue. Additionally, potential knowledge gaps stemming from ChatGPT's last training data update in 2021 might introduce other limitations, as evidenced by the outdated API information.

Thirdly, when we juxtaposed the time taken to complete the project using ChatGPT against an estimate of the duration our students might have required independently, a significant difference emerged. While it is challenging to pinpoint exact durations, our assessment suggests that the students would likely have taken double the time, factoring in the learning curve associated with deep learning. This observation underscores the advanced technical solutions that ChatGPT brought to the table, surpassing the typical expectations for a Bachelor's thesis in our context.

Lastly, a crucial distinction emerged between the ability to produce technically sound code and the capability to conceptualize a robust scientific experiment. ChatGPT's proposed binary gender classification serves as a case in point. While textual data, including ChatGPT's responses, can exhibit male/female biases, the realm of gender identity and associated biases is far more nuanced than a mere male/female dichotomy. Consequently, while the initial inquiry into ChatGPT's potential gender bias might have been rooted in a valid concern, ChatGPT's approach was limited in its consideration of gender identities. As previously emphasized, even if the analytical approach is flawed, it does not necessarily detract from the technical precision of the code. However, it accentuates the paramount importance of vigilant oversight when directing ChatGPT's analytical endeavors. Our initiative to cross-validate further highlighted this, as it unearthed a significant error during the RoBERTa fine-tuning phase. This incident reinforces the necessity of critical oversight, especially given that ChatGPT neither proposed this validation step nor successfully rectified the identified issue within our allocated exploration time.

5. Conclusion

Based on this pilot study, we identified two critical dimensions that need to be explored in the future. Firstly, our investigation was confined to ChatGPT's coding ability in data analytics. A broader spectrum of academic programming challenges should be probed, given the potential variability in ChatGPT's proficiency across different domains. Secondly, the skill level of the users could significantly influence the outcomes, necessitating its examination in subsequent studies.

Given that the final code was functionally sound, which resulted in considerable time savings, and the analytical approach surpassed our anticipations, we deduce that, in this context, ChatGPT can serve as a remarkably efficient tool. One could contend that exposure to novel technologies (in this case, deep learning) via ChatGPT offers students a valuable learning experience. Conversely, had students embarked on a thesis exploring gender bias in ChatGPT and relied on ChatGPT for the analysis, they would have contravened the University's ethical guidelines, as the AI's contributions cannot be claimed as personal efforts in evaluative contexts. Intriguingly, with proper disclosure—ensuring no misrepresentation of code creation as their own endeavor—the evaluation would pivot to their genuine contributions, encompassing aspects like research formulation, literature review, critical analysis, writing proficiency, and presentation. Assuming a critical discussion on the choice of binary gender bias, the students' work would likely meet approval standards, potentially achieving a higher research quality. To truly gauge the advantages and limitations of integrating ChatGPT's contributions into theses (or other scholarly pursuits), it is imperative to juxtapose these outcomes with projects devoid of ChatGPT's assistance, considering the aforementioned dimensions.

Footnotes

[1] That is, not grounded in established scientific evidence or methodologies.

[2] While we find no specific legal text, many Universities or Schools include similar statements in their guidelines (Jackalin; Olsson, 2023)

References

- Abd-Alrazaq, A., AlSaad, R., Alhuwail, D., Ahmed, A., Healy, P.M., Latifi, S., Aziz, S., Damseh, R., Alrazak, S.A., Sheikh, J., et al., 2023. Large language models in medical education: Opportunities, challenges, and future directions. JMIR Medical Education 9, e48291.
- Amin, S., Hellström, J., 2023. Can chatgpt generate code to support a system sciences bachelor's thesis?
- Bird, S., Klein, E., Loper, E., 2009. Natural language processing with Python: analyzing text with the natural language toolkit. " O'Reilly Media, Inc.".
- Dev, S., Monajatipoor, M., Ovalle, A., Subramonian, A., Phillips, J.M., Chang, K.W., 2021. Harms of gender exclusivity and challenges in non-binary representation in language technologies. arXiv preprint arXiv:2108.12084 .
- Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 .
- Dinan, E., Fan, A., Wu, L., Weston, J., Kiela, D., Williams, A., 2020. Multi-dimensional gender bias classification, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online. pp. 314–331. URL: https://www.aclweb.org/anthology/2020.emnlp-main.23, doi:10.18653/v1/2020.emnlp-main.23.
- Doughman, J., Khreich, W., El Gharib, M., Wiss, M., Berjawi, Z., 2021. Gender bias in text: Origin, taxonomy, and implications, in: Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing, pp. 34–44.
- Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep learning. MIT press.
- Hellström, J., 2023. Chatgpt dialogues generated during thesis work in systems science. figshare

URL: https://doi.org/10.6084/m9. figshare.22822091.v1, doi:10.6084/m9.figshare.22822091.v1.

- Jackalin,M.,. Teaching. URL: https://www.su.se/staff/services/teaching/guidelines-on-using-ai-powered-chatbots-in-education-and-research-1.649009.
- Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., et al., 2023. Chatgpt for good? on opportunities and challenges of large language models for education. Learning and individual differences 103, 102274.
- Leiter, C., Zhang, R., Chen, Y., Belouadi, J., Larionov, D., Fresen, V., Eger, S., 2023. Chatgpt: A meta-analysis after 2.5 months. arXiv preprint arXiv:2302.13795 .
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V., 2019. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 .
- Okan, Ç., 2023. Ai and psychiatry: The chatgpt perspective. Alpha Psychiatry 24, 41.
- Olsson, E., 2023. Lärarens rädsla: Att ai-boten förbjuds i skolan.
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al., 2019. Language models are unsupervised multitask learners. OpenAI blog 1, 9.
- Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al., 2023. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 .
- Zhang, C., Bengio, S., Hardt, M., Recht, B., Vinyals, O., 2021. Understanding deep learning (still) requires rethinking generalization. Communications of the ACM 64, 107–115.