Proceedings of 2024 IEEE 12th International Conference on Intelligent Systems (IS)

From Data to Documentation: Exploring the Use of ChatGPT's Custom Instructions for Report Generation

Janne Harjamäki, Pekka Sillberg, Mika Saari, Petri Rantanen, Jari Soini and Pekka Abrahamsson
Tampere University
Pori, Finland

***Abstract*—Generative artificial intelligence has attracted global attention and interest in software engineering research, practical applications, and in business, especially in the past two years. Tools like Gemini, Copilot, and ChatGPT have been widely studied in various professional contexts for their exceptional ability to produce human-like content. Furthermore, the number of these artificial intelligence tools is increasing explosively. Within this context we have picked a very specific use case: namely, converting notes into a full, uniform report using ChatGPT's custom instructions. In this paper, we present the process of creating custom instructions for a certain task, how the instructions were tested, and the results of our study. The study shows that even though using custom instructions alone is perhaps not quite sufficient yet for fully automatic report generation, the approach can still save a considerable amount of time and resources. This study also highlights the pitfalls and obstacles faced, and should help those who are planning to embark on a similar quest.**

***Index Terms*—Generative AI; ChatGPT; Custom Instructions; Reports**

I. Introduction

Generative Artificial Intelligence (GenAI) has been the focus of global attention and interest in software engineering (SE) research, practical applications, and business, especially over the past couple of years. Tools like Gemini[1], Copilot[2], and ChatGPT[3] have been widely studied in various professional contexts because of their exceptional ability to produce human-like content. Furthermore, the number of these artificial intelligence tools is increasing dramatically, and there are many websites that compile lists of AI tools[4] [5].

In our laboratory at Tampere University, Finland, we have also had students take a look at the various AI tools available on the market, try them out, and write notes about their work [1]. Based on these notes we created a report that reviews 85 AI tools [2]. While doing this work we faced a very common use case: how to convert annotations, bullet lists, and various technical notes into a full textual report? Doing the work manually can be time-consuming, tedious, and error-prone. Thus, we set ourselves the task of finding out how well ChatGPT's Custom Instructions (CI) could be leveraged to automate the conversion process, what steps would be required, and what issues might arise.

Within the context of this paper, the analysis is focused on ChatGPT's CI feature, but a similar approach could be taken with OpenAI's API[6], or with custom GPTs[7]. Also, even though the use case is on transforming tool notes into a tool review report, the approach presented in this paper could be applied to other cases as well.

The structure of this paper is as follows. In Section II, we present the background and related research. In Section III, we describe how we proceeded in creating the CI required for generating the tool review report. Section IV presents our test results and preliminary metrics. And finally, Section V summarizes this research.

II. Background

When discussing AI-generated text and its research, search results often yield outcomes like "how to identify AI-generated text?" As a response to this, tools for detecting AI-generated text, such as Outfox [3], have been developed. Additionally, there are objectives to develop a watermark for AI-generated text [4].

AI-generated content (AIGC) utilizes generative large-scale AI algorithms to assist people or replace them in creating massive, high-quality, and human-like content more quickly and affordably, based on user-provided instructions. Ref. [5] present an in-depth study on the working principles of AIGC, security and privacy threats, current solutions, and future challenges. In addition, these GenAI tools also create numerous opportunities for research [6].

In one study [7], GenAI was used to enrich the data story with predetermined relevant context. The study showed how GenAI can improve the depth and relevance of data narratives.

When discussing content produced by artificial intelligence, it is also essential to consider the ethics of AI [8]. A significant

[1]https://gemini.google.com/app
[2]https://copilot.microsoft.com/
[3]https://openai.com/chatgpt
[4]https://aitoolsdirectory.com/
[5]https://aitoolreport.beehiiv.com/
[6]https://platform.openai.com
[7]https://openai.com/blog/introducing-gpts

979-8-3503-5098-2/24/$31.00 ©2024 IEEE

challenge for the reliability and sustainability of AI technology is the transparency of its sources and results. This is crucial due to the tendency of AI technology to "hallucinate" information. The issue of hallucination has been highlighted in the context of ChatGPT in research by [9], [10], among others.

Ref. [10] also discusses of utilizing genAI in systematic literature reviews among the concerns which may affect the practical usefulness and reliability. The process described in our study is an attempt to utilize automated genAI systematically for carrying out a specific task at acceptable level.

III. Method

We developed and improved an artifact (ChatGPT's custom instructions) using iterative design cycles. As part of our research development, we leveraged previously collected material [1] as samples from a wide range of available AI tools. We utilized the paid version of ChatGPT plans (ChatGPT Plus with GPT-4 [11]) in our research endeavors.

A. Custom Instructions Definition

ChatGPT's Custom Instructions Definition (CID) is a two-part form that allows users to define the rules and needs behind the prompts for the response to be created. The first section, CIDa, describes the user (background) and the second section, CIDb, defines the response to be provided (task).

The concept was to develop the CID through two different levels of design cycle iterations. In major updates, the CID description would be changed significantly, requiring previously tested tools to be reprocessed. In minor updates, the CID content would be fine-tuned, and the changes might only affect individual tools. In practice, the design cycles here were conducted as major updates. All updates were recorded as versions, listed and summarized below.

Version 1.0: CIDa covered the user's and document's background, target audience, and guided the general format of the document. CIDb provided instructions for producing a tool review text, allowing integration of external content through discussions in Finnish. The evaluation was influenced by user experiences, and culminated in a summary.

Version 2.0: The content of the sections was separated further. CIDa briefly outlined the user's background and the objectives of the work. CIDb emphasized English presentation style, the length of the summary, and specified structural elements such as the business model, technical characteristics, and the nature of the tool. Lists were eliminated, and full sentences were required. Unnecessary mentions of tool handling through separate sessions were removed.

Version 3.0: CIDa included a more concise description of the user's background. CIDb's description was revised for better repeatability, requiring specific structures and simplified phrasing for clarity. Testing of text formatting options explored alternatives for presenting pricing information.

Version 4.0: CIDb excluded AI-generated opinions from the content and refined the specification of titles and content more precisely. The language in user discussions was standardized so as not to affect the content's final presentation language.

Version 5.0: CIDa refined the user's background to be university-specific, and CIDb's content was structured into blocks. The role of the AI shifted from providing content suggestions to generating content directly from the source. The description length was capped at A4 size, and the testing results showed a decline in quality.

Version 5.1: CIDa and CIDb were formatted into a continuous script without line breaks or lists. A new task was added in CIDb to report sections in the source text that lacked handling instructions. The AI was then able to generate the required sections directly from the source text.

Version 6.0: Discussions with users were removed from CIDa, focusing solely on source text based content. CIDb was clarified with named structures and conditional content. An increase to 1.5 x A4 was proposed as the upper limit for the text size. The process began testing multiple AI tool materials within each iteration round, which was feasible due to the stabilized content structure. The need to assess whether the number of different sections had increased was also identified.

At the beginning of testing, several iterations were conducted using the same AI tools. Later, starting from version 6.0, the approach shifted to testing multiple AI tool materials within each iteration round. This method became feasible due to the stabilization of the content structure. There was also a need to determine whether the number of different sections had increased.

Version 7.0: CIDb specified the maximum text size as $1.5\times$A4, with stronger directives for omitting prices. The "User Review" section received updated guidance on content priority and extent based on available space. Testing indicated adherence to instructions but revealed difficulties in identifying tool names and preventing list formats.

Version 8.0: Refinements in CIDb included specifying titles for sections and introducing a new "Use Cases" section suggested by the analysis phase. However, the integration of APIs and connections to other tools was not considered initially and needed updates.

Version 9.0: Formatting rules were emphasized to exclude lines or bullets, and "Use Cases" was added conditionally. The structure and content wording in "Key Offerings" were restructured to improve clarity and relevance.

Version 10.0: CIDb underwent updating in each section, with revised text formatting guidance and an emphasis on analysis as a separate action following text generation. Challenges remained in adhering to text formatting guidelines.

Version 11.0: The structure of CIDb was defined more rigorously with titles and order, separating the analysis task into its own phase. Although the structure did not always follow the desired format, the sections' texts were generated appropriately.

Version 12.0: The focus shifted towards creating a more automated process. Mergers and renaming improved the structure, e.g., combining "Business Model" and "Key Offerings" into "Business Models" and renaming "User Experience" to "Getting Started." The addition of "Analytical Commentary" provided insights for refining sections and definitions.

Version 13.0: Final updates included shortening sentence forms to fit CIDb's character limits and refining sections like "Business Models" and "Getting Started" to enhance clarity and functionality. Challenges with text production adhering to non-list formats persisted, leading to the final design cycle, which marked a significant moment in refining CID effectiveness and clarity.

B. Results of CID Development

As a result of the design cycle iterations, a CID able to transform notes into a report was created. This can be found in Appendix B in [2]. Also, a partially automated process was developed where ChatGPT was able to perform individual tasks under user supervision. The elements of partial automation included:

1) Content reformatting
2) Setting content in LaTeX format
3) Content analysis

The functions for Element 1 were produced through CIDs. For Elements 2 and 3, pre-created prompt scripts were used. In the process, it was possible to chain the elements in a series (1-2-3). Fig. 1 illustrates the results of the development work conducted through the design cycle iterations.

Normal Process Execution: The process begins in *1. CI Setup* by loading the *CID* into the ChatGPT Custom Instructions form. A new prompt session is opened. *Source Text material* is copied into the prompt in *2. Content Creation*. No additional explanation or stimulus is required. ChatGPT processes the content according to the instructions in the CI and produces a result. The user reviews the generated description against the acceptance criteria. A valid result progresses to the next stage, where the user copies and pastes the *LaTeX Format Prompt Definition* into the terminal and initiates execution in *3. LaTeX Formalization*. As before, ChatGPT processes the content and the user validates the outcome. A valid result progresses to the next stage, *4. Tool Review*, where the generated evaluation (*Review Text*) is copied into the desired target material. If selected, the process moves to stage *5. Self Analyze*, where ChatGPT is tasked (using *Analyze Prompt Definition*) with analyzing the result using the provided instructions (*CID*) and the *Source Text* material. Successful analysis outcomes (*Analyze Text)*) are recorded for future development actions in *6. Save Analysis*. Corrective actions for CID are done via *7. CID Update* and corrective actions for *LaTeX Format Prompt Definition* are done in *8. LaTeX Format Prompt Update*. Corrective actions for *Analyze Prompt Definition* are done in *9. Analyze Prompt Update*, where also a trigger for *CID Update* can be initiated.

Process Exceptions: If the content in *2. Content Creation* deviates significantly from the required standards, the regenerate function is used. If problems related to the *CID* are identified in the results, a trigger for corrective actions to the *7. CID Update* is initiated. If the result in *3. LaTeX Formalization* is incomplete or incorrect, it is corrected by regenerating. If issues are noticed in the *LaTeX Format Prompt Definition*, a trigger for corrective actions to the *8. LaTeX Format Prompt Update* is initiated. The regenerate function can be used in case of issues in *5. Self Analyze*. If problems or deficiencies of *Analyze Prompt Definition* are detected during the evaluation process, a trigger for corrective actions to the *9. Analyze Prompt Update* is initiated.

IV. Tests

It is important to point out that, as a rule, we did not expect the results (reviews) produced by the AI to be completely flawless. There are two reasons for this. First, it is challenging to make the AI systematically produce coherent results that fit the instructions because of the inherit nature of genAI to produce slightly varying results. Second, in most cases the issues turned out to be minor and easily fixable by humans. In other words, at least in our use case it was faster for the human to make the minor corrections to the reviews produced by the AI rather than to spend an extensive amount of time modifying the prompts, CI, or regenerating the results. Running an indefinite amount of regenerations *might* eventually produce flawless results, but doing so would not be practical. Furthermore, at the time of making this study, ChatGPT [11] had a quota of 40 messages per three hours, which in practice places limitations on how many regenerations can be made with our set of 85 tools without the process becoming tediously time-consuming. Thus, we kept the number of regenerations to a minimum to simulate a process that the user could actually follow in practice. The process for accepting the generated result was as follows:

1) If all (six) required sections are present in the generated output, there are no bulleted lists, or other formatting issues especially required to be excluded, the results are accepted.
   - The extra "conclusions" section is accepted. This does not interfere with the required six sections and is simple to remove (by a human). Attempting to remove the "conclusions" section also runs the risk of ruining the other six, otherwise correct, sections.
   - The inclusion of monetary values is accepted, even though it was not desired.
2) If the required sections are not present, up to four regenerations (five attempts in total) are permitted to achieve the acceptable result (as described above). There is some variance on how quickly ChatGPT produces results, but running four regenerations is generally quick[8], i.e., in

[8]There was some variance on how quickly ChatGPT responded, and the service often began to slow down around 13:00-15:00 UTC time, i.e., when the workday starts in the United States (morning).

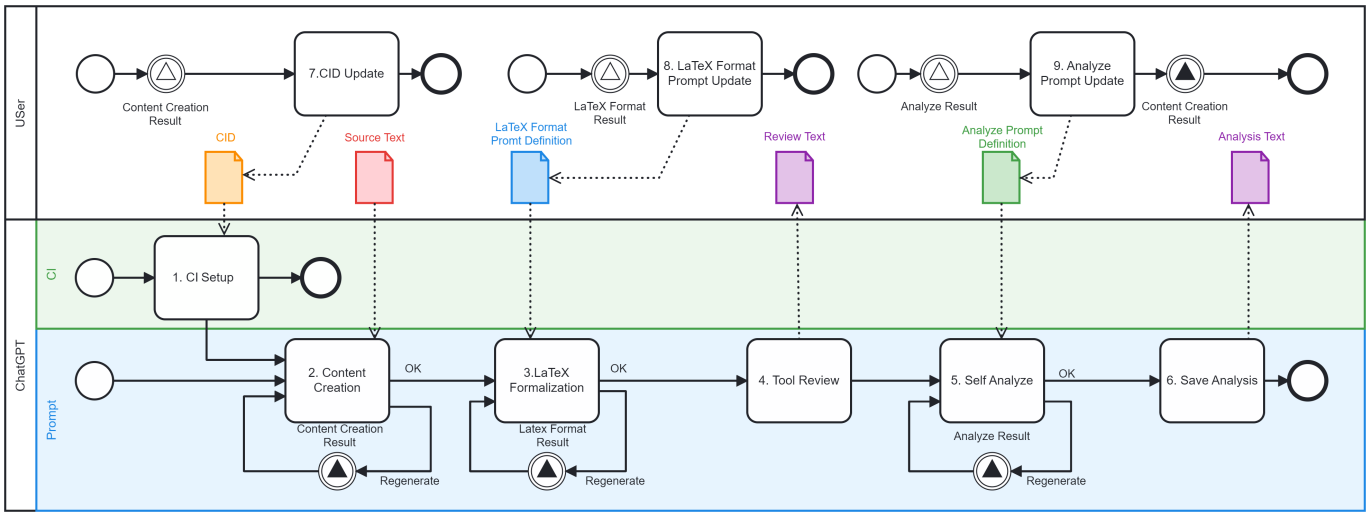

Fig. 1. Illustration for content transformation process using GenAI.

our use case within a reasonable timeframe that the user could be expected to wait for a better answer. The regenerations were performed using the *Regenerate* button in the web user interface of ChatGPT.

- If, after three regenerations, the output still has serious errors, such as several unwanted sections, entirely incorrect content, or major formatting issues (e.g., bullet lists), the regeneration button is ignored and a new (separate) prompt is given to the AI. The reason for this is that, if serious issues arise, regenerations might slightly vary the seriousness of the generated issues, but do not seem to entirely remove them, requiring changes to the CI or prompts.

The source material to be transformed was copy-pasted into the ChatGPT's chat prompt (input) as is, without any additional instructions. All instructions were provided solely using the ChatGPT's CI configuration. Each review was processed in its own session, opening a new ChatGPT prompt to prevent content leaking from previous sessions (or generated reviews) into new (following) reviews. In each case, the number of regenerations before an acceptable result was achieved and documented, as well as any issues detected. These are further discussed in the following subsections (IV-A and IV-B). All tool reviews were processed between February 1st and April 2nd, 2024, using ChatGPT Plus (GPT-4) [11].

A. Regenerations

Fig. 2 illustrates how many regenerations were required by following the process described above to produce acceptable results. The X-axis shows the number of regenerations and the Y-axis (and the number on top of the bars) shows how many reviews were produced by the regeneration count concerned. The cases (regenerations) where flawless results (no formatting issues, no monetary values, no extra or incorrect sections) were produced are included separately in the diagram with the *(OK)* text. The asterisks highlight the regenerations that include cases where separate prompts were required to produce acceptable results. More specifically, in *3*, there were three cases requiring re-prompts; in *4*, one re-prompt; and in *10 (OK)*, one re-prompt. The re-prompts are discussed further in the next section (Section V).

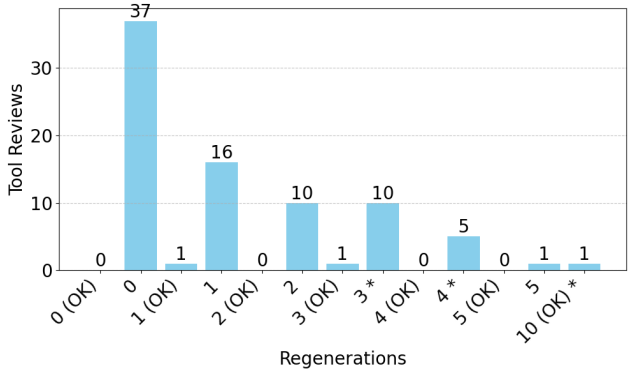

Fig. 2. The number of regenerations required to produce an acceptable tool review. The regeneration counts marked with an asterisk (*) contain results that required additional prompts. Counts with (OK) show the number of flawless results.[9] [10]

The figure shows that, in general, no more than three regenerations were required, making the limit of maximum of four regenerations a reasonable number. The diagram also shows five and ten regenerations. The higher number of regenerations were mistakenly run for two reviews, and could also be removed from the diagram as invalid results, but are left for purposed of illustration. Interestingly, in one case, the

[9]In the review report generated by the AI [2], in Section 2.6.7., the section *Use Cases* is missing. The issue was detected after the report was published. The numbers listed in this figure do not include this malformed review.

[10]A total of three reviews that had problems in their generation have been excluded. Thus, the total sum of reviews in the figure is 83, not 85.

AI produced a review with an extra conclusion section, but when converting[11] the review with ChatGPT to LaTeX format for the publication of the review report, this extra section was removed by ChatGPT in the conversion. Still, in Fig. 2, this result is counted in the category *0*, as the initial result was not entirely flawless.

In two cases, ChatGPT produced multiple result options. In both cases, two result options were given, with each result containing incorrect sections. These two cases have been excluded from the figure. Once, the ChatGPT timed out (issue with the service), failing to produce any result. This was not counted as a "regeneration" attempt, and is not included in the numbers shown in Fig. 2. Also, the amounts include only the results achieved using the final version of the CI. Trial runs with other versions and regenerations run purely for pre-testing to check whether the setup works at all, have been omitted.

B. Detected Issues

There were some issues in the majority of the generated reviews. Only 12 reviews can be considered reasonably flawless. In five out of these 12, there were 1–2 monetary values mentioned in the generated text even though the AI was specifically instructed not to include them, with the rest (seven) having no detected issues of any kind. The issues are listed in more detail in Table I.

TABLE I
NUMBER OF INCORRECT OR MISSING SECTIONS

General Issues	
Unwanted Sections	87
Monetary Values	34
Format and Style	21
Missing Sections	
Description	2
Use Cases	10
Business Models	0
Getting Started	0
Limitations	0
User Reviews	0

By far the most common issue detected was the generation of unwanted sections. Out of the 87 unwanted sections, 44 were some variations of a "conclusions" section, in which the AI represented insights or summary of the review's content. In 14 cases the AI included an untitled conclusion section at the end of a requested section. Two extra sections (named "Other" and "Features" (with Microsoft Account) appeared in two separate reviews. The remaining 41 unwanted sections can be traced as content "leaking" from the original source material. The source material [1] had separate sections *UX* and *Key offering*. Even though not instructed to do so, the AI included various versions (e.g., *User Experience (UX)*, *UX (User Experience)*, *Key offerings*, ...) of these two sections in the final reviews.

[11]Two times the LaTeX conversion on ChatGPT produced *sections* even when requested to only use *subsections*, but in general, at least in our use case, converting to LaTeX worked exceptionally well.

The monetary values were only included in the *Business Models* section, regardless of where they appeared in the source material. The values themselves seemed to be copied as is – there were no modifications to the numeric values or currencies. Also, ChatGPT does not understand the concept of units as such, which may result in invalid units being copied from the source material and ending up mixed with valid units[12].

The majority of the formatting and style issues were closely related to the case of *UX* and *Key offering* described above – certain elements were copied from the original content even though the AI was especially forbidden to use them. Out of the 21 *Format and Style* issues, 19 were cases where the format used in the source material (bulleted or numbered lists) was copied to the generated (new) material. The remainder included a case in which the requested section *Business Models* appeared in the singular, and one strange case of the colon character (:) appearing after each section title.

Based on our results, it also seems that it is more likely for ChatGPT to add unwanted sections than to omit required sections. It should also be noted that in the cases of the missing *Use Cases* sections, the source material was also lacking, having only one or two sentences that could perhaps in our opinion be interpreted as discussing potential use cases. Thus, it is likely that a human would also have left out the sections, or added a notification that no information was available.

V. CONCLUSIONS

Additional artifacts were developed during the actual CID Development. A partially automated content transformation process was defined to manage large amounts of source and target materials. LaTex Format Prompt Definition was needed to manage the review text format and Analyze Prompt Definition was needed to obtain ChatGPT's view of its decisions on handling sections.

In our case, the most serious issue was ChatGPT's tendency to copy formatting or content from the source material, even though it was specifically instructed not to. Often, it was very challenging to eliminate these additions. In general, naming the sections in the CI that should not be added does seem to work, but depending on the source material, new sections may keep popping up, making updating the instructions a tedious process. When using the ChatGPT's CI, an additional problem is the maximum length of the instructions (1500 characters), which may become an issue if a higher number of exclusion instructions are required.

A lesser issue was the generation of a conclusion-style section in many reviews. In principle these could also be considered major issues as generating these sections was not what was requested, but on the other hand, they are generally easy to spot visually or could be searched for automatically.

[12]For example, $/month, $/mon and $/m may end up being used in the same sentence, always meaning *month*, when discussing online service fees. This is especially problematic with units that generally have other meanings, such as m more commonly being associated with the SI unit *meter*

The inclusion of an extra conclusion section does not seem to affect the other sections, making simply removing the extra section an easy task. Also, it could be possible to set the AI to be more deterministic (e.g., modifying temperature values), or ask it to be more deterministic or nor not to offer any insights of its own. How well this works would depend on the use case (and source material).

As mentioned in Section IV, in cases where the results generated by the AI had major issues that were not resolved by running regenerations, we used separate extra prompts to help the AI to produce desired content. Interestingly, providing "new" instructions was not required. Prompts like "use the sections defined in CI, do not use bullet lists", "keep the content and text the same, but use the sections given in CI", "could you write only sections mentioned in instructions and each section without bullets or lists", or simply "use the sections defined in CI" were enough to make ChatGPT produce better results. All of these instructions were already part of the CI. Thus, it was generally enough to ask ChatGPT to follow the instructions it had already been given, without giving any new, more elaborate instructions.

As discussed in Section III, we also used ChatGPT to convert the produced reviews into LaTeX format for the final report. For this purpose, ChatGPT seemed to work more or less flawlessly with both GPT 3.5 Turbo and GPT 4 models. Also, there was no need to provide any detailed instructions or prompts, simply asking to "convert the following text into LaTeX" was enough. As a conclusion, it can be said that simple format conversions are quite simple to implement, but in cases where some variety in the generated texts is required while still conforming to some basic template, fine-tuning the AI responses becomes more challenging.

Considering our original research goal, presented in Section I, i.e., how well ChatGPT's CI could be leveraged for systematically transforming source material (in this case technical notes) into a uniform report, the results are somewhat mixed. On the one hand, it cannot be ruled out that CI could be constructed in a way that would provide "flawless" results. On the other hand, creating good instructions can be challenging, and if making the instructions takes more time than achieving the result (making the report) without them, this will greatly diminish any potential gains. Fortunately, based on our results, even though it seems that human intervention is still required, the results are still reasonably good. It is also worth noting that, even though we did a rough check that the AI generated content (reviews) did not contain hallucinations or serious mistakes, in the context of this study, we did not make an extensive analysis on the quality of the produced text.

Considering the potential gains, one way to look at this is the amount of time (and resources) saved. For a human, converting notes into a properly formatted textual format can be tedious and time-consuming. It can easily take 15 to 30 minutes to read the source material, create new text, and format it in the desired style. An AI does the same task in seconds or in a couple of minutes, and even including a quick quality check by a human to fix generally quite minor issues, time savings are bound to be involved. At the very least, creating custom instructions for repetitive text processing, conversion, and formatting tasks would hand over many of the more tedious tasks from human to the AI.

ACKNOWLEDGMENT

This work was co-funded by the European Union and the Regional Council of Satakunta.

REFERENCES

[1] D. Lahtinen, *Review of 85 AI Tools*. Tampere University, 2023. [Online]. Available: https://www.avoinsatakunta.fi/wp-content/uploads/2024/03/Review_of_85_AI_Tools.pdf

[2] J. Harjamäki, P. Rantanen, D. Lahtinen, P. Sillberg, M. Saari, J. Grönman, Z. Rasheed, A. M. Sami, and P. Abrahamsson, *The Report of 85 AI Tools, GenAI Content Production: Enhancing Repeatability and Automation with ChatGPT*. Tampere University, 2024. [Online]. Available: https://www.avoinsatakunta.fi/wp-content/uploads/2024/04/The_Report_of_85_AI_Tools.pdf

[3] R. Koike, M. Kaneko, and N. Okazaki, "Outfox: Llm-generated essay detection through in-context learning with adversarially generated examples," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, pp. 21 258–21 266, Mar. 2024. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/30120

[4] J. Kirchenbauer, J. Geiping, Y. Wen, J. Katz, I. Miers, and T. Goldstein, "A watermark for large language models," in *Proceedings of the 40th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., vol. 202. PMLR, Jul. 2023, pp. 17 061–17 084. [Online]. Available: https://proceedings.mlr.press/v202/kirchenbauer23a.html

[5] Y. Wang, Y. Pan, M. Yan, Z. Su, and T. H. Luan, "A survey on ChatGPT: Ai-generated contents, challenges, and solutions," *IEEE Open Journal of the Computer Society*, vol. 4, pp. 280–302, 2023. [Online]. Available: https://ieeexplore.ieee.org/document/10221755/

[6] A. Nguyen-Duc, B. Cabrero-Daniel, A. Przybylek, C. Arora, D. Khanna, T. Herda, U. Rafiq, J. Melegati, E. Guerra, K.-K. Kemell, M. Saari, Z. Zhang, H. Le, T. Quan, and P. Abrahamsson, "Generative artificial intelligence for software engineering – a research agenda," Oct. 2023. [Online]. Available: https://dx.doi.org/10.2139/ssrn.4622517

[7] A. Lo Duca, "Using retrieval augmented generation to build the context for data-driven stories," in *Proceedings of the 19th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - IVAPP*. SciTePress, Mar. 2024, pp. 690–696.

[8] B. C. Stahl and D. Eke, "The ethics of ChatGPT – exploring the ethical issues of an emerging technology," *International Journal of Information Management*, vol. 74, p. 102700, Feb. 2024. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0268401223000816

[9] C. S. Smith, "Hallucinations could blunt ChatGPT's success," Mar. 2023. [Online]. Available: https://spectrum.ieee.org/ai-hallucination

[10] M. M. Hossain, "Using ChatGPT and other forms of generative AI in systematic reviews: Challenges and opportunities," *Journal of Medical Imaging and Radiation Sciences*, vol. 55, no. 1, pp. 11–12, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1939865423019021

[11] OpenAI. How can I access GPT-4? [Online]. Available: https://help.openai.com/en/articles/7102672-how-can-i-access-gpt-4