Mirjam Sophia Glessmer, LUND UNIVERSITY, mirjam.glessmer@lth.lu.se
Rachel Forsyth, LUND UNIVERSITY, rachel.forsyth@edusupport.lu.se

SoTL in Process

Superficially Plausible Outputs from a Black Box: Problematising GenAI Tools for Analysing Qualitative SoTL Data

ABSTRACT

Generative AI tools (GenAI) are increasingly used for academic tasks, including qualitative data analysis for the Scholarship of Teaching and Learning (SoTL). In our practice as academic developers, we are frequently asked for advice on whether this use for GenAI is reliable, valid, and ethical. Since this is a new field, we have not been able to answer this confidently based on published literature, which depicts both very positive as well as highly cautionary accounts. To fill this gap, we experiment with the use of chatbot style GenAI (namely ChatGPT 4, ChatGPT 4o, and Microsoft Copilot) to support or conduct qualitative analysis of survey and interview data from a SoTL project, which had previously been analysed by experienced researchers using thematic analysis. At first sight, the output looked plausible, but the results were incomplete and not reproducible. In some instances, interpretations and extrapolations of data happened when it was clearly stated in the prompt that the tool should only analyse a specified dataset based on explicit instructions. Since both algorithm and training data of the GenAI tools are undisclosed, it is impossible to know how the outputs had been arrived at. We conclude that while results may look plausible initially, digging deeper soon reveals serious problems; the lack of transparency about how analyses are conducted and results are generated means that no reproducible method can be described. We therefore warn against an uncritical use of GenAI in qualitative analysis of SoTL data.

KEYWORDS

GenAI, generative AI, qualitative analysis, research methods, academic development

QUALITATIVE ANALYSIS OF SOTL DATA AND GENERATIVE ARTIFICIAL INTELLIGENCE

A lot of SoTL activity relies on analysing qualitative data sets which have been collected in a specific educational context. Traditionally, qualitative analysis means coding of the data based on how people familiar with the context read and interpreted it. This is done as transparently as possible, but it is also labour-intensive and prone to omissions or differences in interpretation, like any complex human judgements. As Generative Artificial Intelligence (GenAI) tools are becoming available, they are increasingly being explored as tools to support SoTL by mitigating human error and speeding up the analysis of large datasets.

GenAI tools are technologies that are able to generate new content, such as text, images, or audio, based on patterns learned from existing data. These technologies are based on Large Language Models, which have been trained using large sets of data to provide outputs, and then use human feedback to validate those outputs. In relation to research, it has been suggested that the tools might

be useful for generating ideas, summarising and synthesising existing content, research and analysis, and asking questions (McCormack 2023). This is despite the issue of "hallucinations," described as "text that is nonsensical, or unfaithful to the provided source input" (Ji et al. 2023, 248).

Hannigan, McCarthy, and Spicer (2024) explain that GenAI tools are "'predicting' responses rather than 'knowing' the meaning of their responses" and coin the term "botshit" for what is produced by uncritically using GenAI output: Something that can happen to be right or wrong, but that is used without regard for its veracity. Lindebaum and Fleming (2023) argue that the use of GenAI tools in qualitative analysis would undermine responsible research and change our understanding of what research should be; they point out that GenAI tools have no stake in the outcome of the research and cannot accommodate the context of the research. This means that GenAI tools could infringe at least one of Felten's five principles for good practice in SoTL, that research should be grounded in context (Felten 2013).

Davison et al. (2024) identified five ethical issues with the use of GenAI for qualitative research: consideration of data privacy, security, interpretive sufficiency, potential biases, and the responsibility and agency of the researchers. They explain that problems in interpretation relate to the absence of context when looking only at one part of the data, the text. The currently available GenAI tools are just analysing a transcript. They do not have information that the interviewer might retain about body language or tone of voice, or historical or organisational context which might influence responses and interpretation. This lack of context is also picked up by Pargman et al. (2024), who point out that the tools are trained on patterns of text: "on what language looks like, not what it means" (75). This challenge of interpretive sufficiency is linked closely to Davison et al. (2024)'s other areas to consider: biases and researcher agency and responsibility. In qualitative research, the researcher usually considers the overall context and their place within it and is able to understand and discuss that this situation can lead to differences in interpretation which may affect the reliability and validity of their results.

However, not all authors are negative about the use of these tools in qualitative analysis. The potential objectivity of only looking at text, without bringing one's own subjective interpretation of tone, voice, gestures, or style of talking, have been welcomed by some, along with the possibility of faster processing (Perkins and Roe 2024). Gamieldien, Case, and Katz (2023), working with a large dataset, found two GTP-3.5-based tools to be quite accurate, compared with human analysis, in identifying themes from 3,800 short student reflections based on short written responses to specific prompts such as, "What did you do differently between Exam 1 and Exam 2?"

As academic developers who regularly support SoTL projects, we are frequently presented with questions of whether GenAI tools can and should be used to analyse SoTL data. Our university pays for certain GenAI tools, which might be interpreted as an invitation to use them for all kinds of purposes, including SoTL. In many disciplines, using GenAI tools in research (developed and trained for a specific purpose) is widely accepted. We wanted to be able to answer the questions concerning GenAI analyses of qualitative SoTL datasets, and contribute to the discussion about how we, as a SoTL community and as SoTL researchers, can learn to work with GenAI in a scholarly way. We set out to explore the potential application of these tools for qualitative analysis in SoTL projects, taking into account the warnings from Lindebaum and Fleming (2023) and from Davison et al. (2024) but also the positive practical example presented by Gamieldien, Case, and Katz (2023).

In considering the use of GenAI in research, we reviewed the "living guidelines" on the use of GenAI in research produced by the European Commission (ERA Forum 2024) which provides advice for researchers, universities, and funders. The guidelines do not address ethical dilemmas directly but suggest that researchers remember that they are responsible for the output of these tools, that they

ought to be transparent in their use, and that they should clearly understand and act on their knowledge of privacy, confidentiality, and intellectual property. This is helpful but puts a lot of responsibility on individual researchers to apply this to their SoTL context and to take the time to understand how the tools work.

Hannigan, McCarthy, and Spicer (2024) have considered the epistemic risks of the use of GenAI tools and suggest the development of "guardrails" for GenAI use: a set of rules, guidelines, limitations for GenAI use, which they suggest should be designed to be technology-, organisation-, and user-oriented. We consider how such guardrails might be developed for the SoTL context.

CASE STUDY OF USING GENAI TO ANALYSE SOTL DATA

For tasks where the veracity of GenAI outputs is difficult, yet crucial, to verify (as we would argue is the case for SoTL activities), Hannigan, McCarthy, and Spicer (2024) suggest an "authenticated" mode of working with GenAI that "requires users to configure extra guardrails with the chatbot and employ critical thinking skills, inductive reasoning, and forecasting to estimate the veracity of a statement produced by a chatbot." Below, we are using an inverse approach, since we already did the full analysis by hand before, to investigate the veracity of GenAI outputs. We took a dataset that we had collected and analysed previously (Glessmer, Persson, and Forsyth 2024). This consisted of 449 students' open text responses to a survey question relating to trust in the classroom, and one of a set of nine interview transcripts exploring students' views about the same topic. Both datasets had been previously consensus coded by at least two out of three SoTL researchers who conducted the interviews.

We tried three commonly available (paid-for) GenAI tools (Microsoft Copilot, Chat GPT4, and ChatGPT4o) to generate a thematic analysis of the data by uploading the relevant files alongside an instruction (or "prompt") which asked the GenAI tools to act on the information in the files. Data was fully anonymized before uploading, according to European Commission living guidelines (ERA Forum 2024).

We designed prompts using suggested approaches from the literature (Lo 2023; Yin et al. 2024). We requested a completely new thematic analysis of the uploaded text files, both using the term "thematic analysis" and also describing the method in a more natural language equivalent. After each question had been processed, we compared the output with our own analysis. We also asked for an analysis of both interview and survey datasets using the codes we had found and applied in our own analysis.

FINDINGS

At first, we were impressed with the outputs we received from these tools as they looked very similar to our own results. We hoped to find that, for future analysis of datasets, it might be possible to compare only a small part of the GenAI tools' output with our own coding, and, if it turned out to be sufficiently similar, trust the GenAI tools to do the rest of the analysis independently. But when looking more carefully, we quickly became concerned with the quality of the outputs and how easily we might have accepted them as good.

We used three different systems: Microsoft Copilot, Chat GPT4, and ChatGPT4o. We have mostly not specified which tools produced which outputs here, as we found similar results with all of them: they respond to prompts as a black box and we have no way of knowing what exactly they are doing. This problem was the same no matter which tool we used, and this is inherent to currently available commercial GenAI (Hannigan, McCarthy, and Spicer 2024). In the following, we give

examples of problems we ran into when trying to analyse qualitative data and probing into outputs to understand where they came from.

Outputs were incomplete

The first questions we put to the GenAI tools asked for general thematic analysis from the two data sets (see Figure 1 for an example prompt). Superficially, the outputs looked similar to the analysis that we had previously done ourselves. But as we compared the outputs with our own analysis, it became apparent that we could not be sure that all the data had been analysed. When the tools were prompted to explain the output by giving all occurrences mapped to a specific code, or counting the number of datapoints linked to a particular theme, only partial results were given.

We do not know if this is because the interface is unable to deal with a dataset of this size (a table with 449 lines of short free-text student responses, 7,500 words in total, and an interview of 7,000 words in a separate file). Despite explicit instructions to go through the data line by line and output a code for each line, the output was missing lines, seemingly randomly, and without flagging that something was missing. By modifying the prompts, we found ways to increase the number of lines which were coded, but not to fully fix it, unless the dataset was broken down into tiny chunks of data that were then each analysed individually. However, breaking down the dataset into smaller parts only helped with this specific issue, the ones below remained.

Outputs were coded differently from our analysis

Whilst there were many similarities with the coding of data to our originally selected themes, there were also differences. After comparing the coding with our own consensus coding, we would not have felt confident with accepting the analysis from the tools. We tried providing more details about what we meant by our codes, but there were still differences. Of course, this happened when we were originally coding the data ourselves, too, but we were able to discuss our decisions and come to consensus. This is not possible with these tools because they cannot provide reliable reasoning about why they have matched certain codes.

One example of such a difference is shown in Figure 2, example A, where we ask for example quotes illustrating how certain themes are discussed in an uploaded interview transcript. We do not agree that the teacher shows “knowledge, skills, and competence in teaching” in the example quote provided by ChatGPT 4o. Considered in the context of the surrounding text in the transcript, the “the teacher knows” in that quote refers to a teacher having invited students to ask questions and thus “knowing” that they should not take it as a reflection on their teaching when students indeed do ask questions. This was not clear from the context of the quote ChatGPT provides.

Outputs wrongly inferred context

An example of GenAI creating context is shown in Figure 2, example B. Here, ChatGPT 4o provides an example quote that is modified from the transcript so it seems that the teacher says that they themselves are a great teacher. Looking at the transcript and listening back to the audio, however, the teacher being a great teacher was an opinion the interviewee expressed as a side thought in between talking about what the teacher did. ChatGPT4o removed the punctuation which made this clear in the transcript.

On another occasion, when we asked a tool to summarise the interview transcript, it did so using gendered pronouns to describe the person being interviewed, despite there being nothing in the body of the document to indicate gender. However, the filename included a female name (which could have been the name of the interviewee or interviewer, the transcriber, an acronym of the project name, or something else).

Figure 1. Screenshot showing one example of a prompt to analyse the data set uploaded in the anonymised file "plain data.docx" in ChatGPT 4o

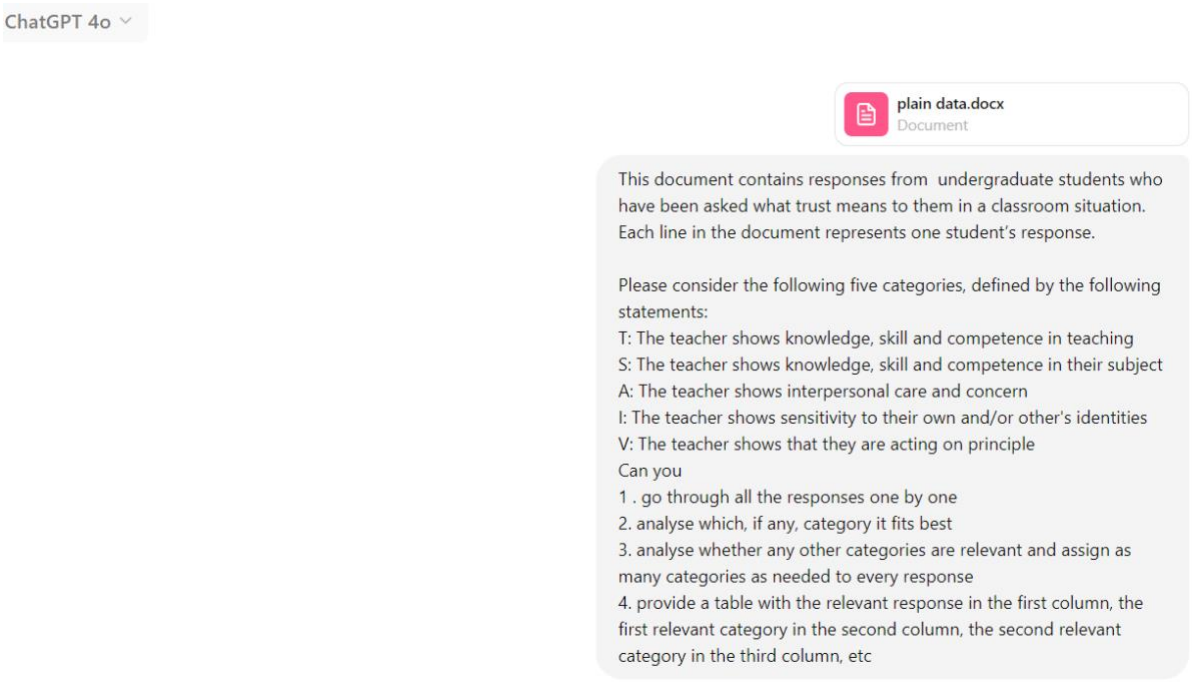
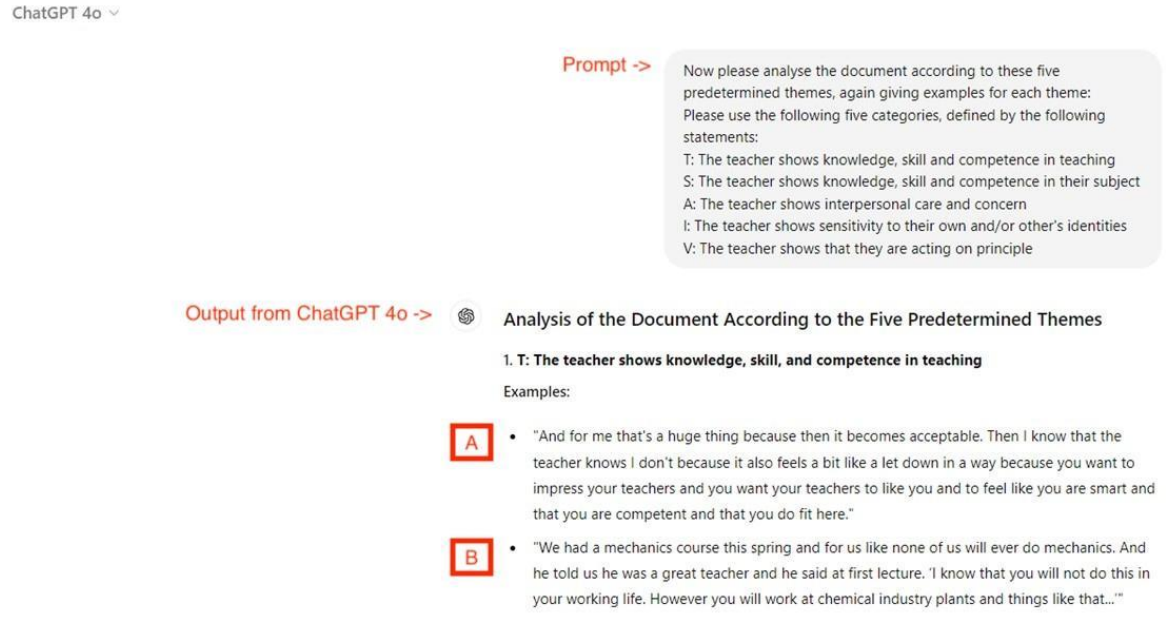


Figure 2. Screenshot of an example prompt asking for the analysis of an interview transcript, and the output by ChatGPT 4o (examples A and B are marked in red and referred to in the text)



Outputs were not reproducible

Continuing with the example above, when we changed the filename to a male name or to not include a name at all, the output used gender neutral pronouns. However, when re-prompting with the female name in the file name, we were unable to reproduce the gendered output, so we have no way of knowing if the inference had come from the filename, the way the interviewee had spoken, or some form of hallucination within the black box.

When we repeated analyses after modifying the prompts, the outputs also sometimes coded statements differently. For example, including the new category "not codable" led to previously skipped lines being coded as one of the previously available categories rather than, as we had expected, being coded with a new category. Other codes were also changed when this additional code had been added. When questioned, the tools often backtracked and changed their response to agree with what we seemed to suggest.

Outputs misinterpreted use of language

It is very common that SoTL research includes student statements in local languages and dialects or referring to school-specific terminology without explanation. One example from our dataset is that many students used the uncommon Swedish verb "att dumförklara," which means condescendingly talking down to someone in a way that makes that person feel stupid. ChatGPT however explains it as "to declare someone as stupid," and example sentences are about making someone look stupid in front of others. As GenAI is trained in English more than in, for example, smaller Scandinavian languages, this is maybe not surprising but is another example of how caution needs to be exercised, since GenAI is trained in a specific context that is likely different from the one in which the SoTL activity is taking place.

Outputs did not reliably reflect the instructions

When asked for a thematic analysis of an uploaded interview transcript, the output looked mostly plausible. However, the output added an explanation to one of its themes that "while not explicitly coded in this example, this is often an underlying theme, in trust relationships." This theme was not present in the interview transcript, so the tool went beyond our instructions to add this context.

DISCUSSION AND CONCLUSIONS

Aware of the danger of producing "botshit" (Hannigan, McCarthy, and Spicer 2024) while trying to explore the opportunities and risks of using GenAI tools to analyse SoTL data, we set out to compare qualitative analysis outputs from GenAI tools with our own analysis of the same data. The initial outputs looked plausible on a surface level, generally similar to our previous analysis, and gave examples which fitted with the themes the tools had generated. However, outputs were incomplete, differed from our analysis, wrongly inferred context, were not reproducible, misinterpreted the use of language, and did not reliably reflect the instructions. There was no way to see how the outputs had been created or what decision processes led to the production of themes. These are serious problems.

Our results show that use of these tools raises many questions about reliability and validity. Whilst Gamieldien et al. (2024) had good results using tools for qualitative analysis of SoTL data, they were working with short answers to relatively closed questions, and we do not know how much checking was required to achieve confidence in the outputs. These GenAI tools put data into an unknown context based on probabilities determined by an undisclosed algorithm and undisclosed training data. Whilst humans also work within contexts and may also not make reliable and valid decisions, we can describe our processes, and they can then be reproduced and critiqued. Since we

don't know what is going on in the analysis we get from GenAI tools, which is inherent in the tools that are currently available (Hannigan, McCarthy, and Spicer 2024), we cannot expect to use them to create real insights into data.

Using Davison et al.'s (2024) five problem areas, we felt confident using the ERA Forum advice to manage data privacy and security. However, we have many concerns that fall under their "interpretative sufficiency," and possibly also "bias" areas. Dealing with those issues is a topic that falls under Davison et al. (2024)'s "researcher responsibility and agency," and we think that a lot more work would be needed to build confidence in outputs. The superficial plausibility of the outputs might disguise these questions if the GenAI tools are used on large datasets. Only probing deeply and reflecting critically surfaced the inconsistencies described above.

We also believe that thought is needed about the use of these tools in relation to the purposes and principles of SoTL. If the first principle of SoTL is to focus inquiry on student learning (Felten 2013), it is important to listen carefully to what is being said by or about students and not just to read an AI generated summary of surveys or interviews. Familiarisation with the full dataset is an important part of context-setting and interpretation.

It is likely that the existing GenAI tools will improve, new ones will be added, and software producers will market qualitative analysis tools to academic researchers. Even though the outputs of future GenAI tools will look more plausible, unless the exact processes are described, the use of such tools in SoTL needs further ethical discussion. In general, one should use analysis of data using GenAI the same way one should use analysis that another researcher produces: If you do not understand the methods, either because they are insufficiently described or because you are not familiar with it, you cannot use it. Davison et al. (2024) suggest that:

> Researchers should engage in critical reflexivity and vigilance to identify, understand and robustly address the ethical issues regarding the use of GenAI in their research practices involving qualitative data analysis. We do not wish to see a situation where we are lulled into thinking that GenAI use is "normal" and that researchers do not need either to pay particular attention to it, or to report their use of it. (6)

Backed by the results from this study, we fully support both the suggestion and the warning on the foundation, and suggest the following guardrails (Hannigan, McCarthy, and Spicer 2024) both for the SoTL community (organisation-oriented) and researcher (user-oriented) to start the discussion.

The SoTL community should:

- Develop a code of conduct, based on Felten's (2013) five principles of SoTL and including ethical considerations about data privacy, security, interpretive sufficiency, potential biases, and the responsibility and agency of the researcher (Davison et al. 2024).
- Provide training for mitigating risk of producing botshit (Hannigan, McCarthy, and Spicer 2024), including how to use GenAI well (e.g. writing good prompts).
- Keep an ongoing dialogue about the use of GenAI across the community to account for future development in GenAI tools, changes in culture, and other unforeseen developments.

SoTL researchers should:

- Keep an open mind towards using GenAI, alongside critical thinking and fact checking
- Think about why they are doing SoTL. To support dialogue with students, we need to actually listen to what they say, or read what they write, not look at something that is

already filtered by GenAI, and thus potentially incomplete, different from careful human analysis, wrongly inferring context, not reproducible, misinterpreting the use of language, and not reliably reflecting the instructions.

We have not addressed here other important areas of discussion about GenAI use, such as the environmental impacts and concerns about training data, as we have focused only on the analysis of SoTL data using these tools. We welcome other SoTL scholars' perspectives and insights on the matter. We hope this article will contribute to the discussion and negotiations within the SoTL community on whether and how GenAI will be used to analyse SoTL data.

DISCLOSURE

We used GenAI tools in the way described in the methods for analysis described there. Anything else is 100% created by the authors.

AUTHOR BIOGRAPHIES

Mirjam Sophia Glessmer (SWE) is an academic developer at Lund University, Sweden, and University of Bergen, Norway. Her main interests are relationships in teaching and learning and how they come into play when teaching for sustainability.

Rachel Forsyth (SWE) is an academic developer at Lund University, Sweden. She is author of Confident Assessment in Higher Education *(Sage 2022) and co-author of the forthcoming* Using Generative AI in Higher Education: Transforming Teaching, Learning, and Student Experience *(Bloomsbury 2025).*

REFERENCES

Davison, Robert M., Hameed Chughtai, Petter Nielsen, Marco Marabelli, Federico Iannacci, Marjolein van Offenbeek, Monideepa Tarafdar, et al. 2024. "The Ethics of Using Generative AI for Qualitative Data Analysis." *Journal of Applied Learning and Teaching*. https://openrepository.aut.ac.nz/server/api/core/bitstreams/7e046f0e-f0ab-4a3f-b2a1-c43f7979fce0/content.

ERA Forum. 2024. *Living Guidelines on the Responsible Use of Generative AI in Research.* European Commission (Brussels). https://research-and-innovation.ec.europa.eu/document/download/2b6cf7e5-36ac-41cb-aab5-0d32050143dc_en?filename=ec_rtd_ai-guidelines.pdf.

Felten, Peter. 2013. "Principles of Good Practice in SoTL." *Teaching & Learning Inquiry* 1 (1): 121–25. https://doi.org/10.20343/teachlearninqu.1.1.121.

Gamieldien, Yasir, Jennifer M. Case, and Andrew Katz. 2023. "Advancing Qualitative Analysis: An Exploration of the Potential of Generative AI and NLP in Thematic Coding." *SSRN*. http://dx.doi.org/10.2139/ssrn.4487768.

Glessmer, Mirjam Sophia, Peter Persson, and Rachel Forsyth. 2024. "Engineering Students Trust Teachers Who Ask, Listen, and Respond." *International Journal for Academic Development*, 1–14. https://doi.org/10.1080/1360144X.2024.2438224.

Hannigan, Timothy R., Ian P. McCarthy, and André Spicer. 2024. "Beware of Botshit: How to Manage the Epistemic Risks of Generative Chatbots." *Business Horizons 67 (5)*. https://doi.org/10.1016/j.bushor.2024.03.001.

Ji, Ziwei, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, et al. 2023. "Survey of Hallucination in Natural Language Generation." *ACM Computing Surveys* 55 (12): 1–38. https://doi.org/10.1145/3571730.

Lindebaum, Dirk, and Peter Fleming. 2023. "ChatGPT Undermines Human Reflexivity, Scientific Responsibility and Responsible Management Research." *British Journal of Management*. https://doi.org/10.1111/1467-8551.12781.

Lo, Leo S. 2023. “The CLEAR Path: A Framework for Enhancing Information Literacy Through Prompt Engineering.” *The Journal of Academic Librarianship* 49 (4): 102720.

McCormack, Mark. 2023. “EDUCAUSE QuickPoll results: Adopting and Adapting to Generative AI in Higher Ed Tech.” *EDUCAUSE*. https://er.educause.edu/articles/2023/4/educause-quickpoll-results-adopting-and-adapting-to-generative-ai-in-higher-ed-tech.

Pargman, Teresa Cerratto, Elin Sporrong, Alexandra Farazouli, and Cormac McGrath. 2024. “Beyond the Hype: Towards a Critical Debate About AI Chatbots in Swedish Higher Education.” *Högre Utbildning* 14 (1): 74–81. https://hogreutbildning.se/index.php/hu/article/view/6243.

Perkins, Mike, and Jasper Roe. 2024. “The Use of Generative AI in Qualitative Analysis: Inductive Thematic Analysis with ChatGPT.” *Journal of Applied Learning and Teaching* 7 (1). https://doi.org/10.37074/jalt.2024.7.1.22.

Yin, Ziqi, Hao Wang, Kaito Horio, Daisuke Kawahara, and Satoshi Sekine. 2024. “Should We Respect LLMs? A Cross-Lingual Study on the Influence of Prompt Politeness on LLM Performance.” *arXiv preprint arXiv*: 2402.14531.