From transcript to insights: summarizing safety culture interviews with LLMs.

Wouter Steijn[1*], *Janneke* van de Loo[1], *Dolf* van der Beek[1], and *Jop* Groeneweg[1,2,3]

[1]TNO, Work Health Technology, Sylviusweg 71 2333BE Leiden, The Netherlands
[2]TU Delft, Faculty of Technology, Policy and Management, Jaffalaan 5 2628 BX Delft, The Netherlands
[3]Leiden University, Social and Behavioural Sciences, Wassenaarseweg 52 2333 AK Leiden, The Netherlands

Abstract. The latest developments in AI have the potential to significantly support qualitative analysis of interview transcripts. This study explores the utility of the OpenAI o1-model to assist in efficiently obtaining reliable summarizations of safety culture interviews. Analysis shows that the current approach has the clear potential to significantly improve efficiency of interviewers by providing a concise report that summarizes multiple interviews according to a pre-defined format. However, some hallucinations are present in the generated report. Additional work will aim at reducing their presence, but such hallucinations also emphasizes that LLMs should primarily assist, rather than replace, interviewers in creating a definitive report.

1 Introduction

Since 2022, the rise of applications such as OpenAI ChatGPT and Microsoft Copilot and their underlying AI-models have opened up new venues to explore. One such a venue is how such Large Language Models (LLM) can contribute to scientific research by assisting in the qualitative analysis of interview transcripts. Traditionally, interview-based research has been considered more demanding than survey-based methods because, although it provides deeper insights into the population, it requires significantly more effort to analyse the data. The latest developments in AI have the potential to significantly mitigate this downside by assisting the researcher in the qualitative analysis [1].

Specifically a lot of studies have examined the utility of LLMs to support thematic analysis, for example in the health sector [2-3]. Several studies compared AI powered analysis with that performed by human researchers [1,4-5]. Their research showed that LLMs are significantly faster in generating themes compared to human researchers [1,5] and that the overlap between human-generated and AI-generated themes lies between 71% [1] and 80% [4]. Furthermore, Tai and colleagues demonstrated that LLMs provide consistent output when extracting themes from text [6].

1.1 Current study

In this paper we describe our application of an LLM to summarize semi-structured interviews addressing safety culture in locations of Dutch organizations to create reports for a Dutch governmental regulator. This work differs from previous research as it does not describe a thematic analysis over the transcripts but rather aims at an comprehensive summary of relevant claims ordered by a set of 14 predefined themes related to the safety culture concept as developed by Zwetsloot and colleagues [7]. The interviews are semi-structured as they allow for an open discussion between the interviewer and interviewee during which the themes (e.g. Leadership and Safety communication) are addressed in a fixed order. The LLM needs to identify the portion of the interview that addresses a certain culture related theme and extract relevant claims (supported by practical examples mentioned by the interviewee) in relation to that theme. An added challenge is that as the interviews are semi-structured and overlap exists between the themes. An added challenge is that sometimes relevant information for one theme (e.g., Safety communication) is mentioned while discussing another theme (e.g. Safety Leadership). The aim is for the LLM to generate a comprehensive overview of relevant statements from the interviews to assist (less experienced) interviewers to create their final report.

1.1.1 o1-model

We decided upon using the o1-model of OpenAI, released in September 2024. The o1-model, developed by OpenAI which is known from the chatbot ChatGPT, excels in generating summaries and analysis of text. Furthermore, the o1-model has improved reasoning capacity compared to its predecessors to deal with complex assignments. A leaderboard created by

* Corresponding author: wouter.steijn@tno.nl

Vectora[†] shows that the o1-model has an hallucination rate of 2.4%.

1.1.2 Research aim

Our aim is to explore whether AI, specifically the o1-model, is a valid tool for efficient and reliable support in summarizing interview-transcripts. The summaries should contain as much relevant information as possible, while containing a minimal amount of hallucinations. We base our analysis on a framework that has previously been used to assess the validity of LLM output. The framework used assesses three factors [8], Fluency (i.e., is the output coherent?), Correctness (i.e., is the output objectively correct?) and Citation quality (i.e. is the output based on the provided information?).

2 Method

2.1 Safety culture interviews

The safety culture interviews took place in the fourth quarter of 2024 at two Dutch organisations active in the Geothermal sector. Organisation A will be the focus of our analysis (7 interviews). Organization B has been included in our analysis for comparative purposes (10 interviews).

The interviews lasted between an hour and an hour and half. During the interview, 14 themes related to safety culture were discussed by having a natural conversation. The themes are addressed in a fixed order starting with a fixed main question. The interviews have a semi-structured format so after the main question follow-up questions are based on what the interviewee responds. The themes discussed were in order:

1. Leadership
2. Productivity versus safety
3. Safety communication
4. Employee engagement
5. Management's view on the causes of incidents
6. Incident registration and analysis
7. Learning from incidents process
8. Management of and collaboration with subcontractors
9. Role of the supervisor regarding safety
10. Relationship between process safety and personal safety
11. Maintenance management
12. Handling procedures
13. Execution and follow-up of audits
14. Complexity and resilience

Due to time constraints, some interviews did not discuss all themes. Interviews were held by experienced researchers with on average 14 years of experience with this interview-method and even more experience with the topic of safety culture.

The interviewers took notes during the interview on which to base their own expert analysis. Reports included a summary of findings per theme with claims made during the interviews concerning 'What is going well' and 'What could be improved' in relation to the theme. These claims were supported with quotes and practical examples discussed during the interview. The interviews were also voice-recorded to allow analysis with the o1-model. Informed consent was obtained from the interviewees.

2.2 Analysis with o1-model

Analysis with o1-model was done by a researcher who was not involved in the interviews and had no prior knowledge of the content of the interviews, although the researcher was familiar with the topic safety culture. The o1-model was hosted with Microsoft Azure on a server in Sweden to comply with GDPR requirements.

Audio files containing the interviews were transcribed with WhisperX[‡] [9]. WhisperX is an open source transcription tool. Although transcriptions are not yet perfect, they were sufficiently correct for the o1-model to correctly interpret the text.

Next, the transcripts were processed with the o1-model. The prompts were systematically developed through prompt-engineering. We refer to Schulhoff for an overview concerning prompting [10]. In the first step, the o1-model was given a system-prompt instructing it to summarize any interviews it received in a predefined structure by listing claims on 'what is going well' and 'what could be improved' for each of the 14 safety culture themes. The model was instructed to include concrete examples in the summary and to make no interpretations, only to include claims that were directly discussed in the interviews.

In the second step, the o1-model was given a system-prompt instructing it to make a pre-defined synthesis of the set of summaries it received. Again, the model was instructed to include concrete examples and not to make any interpretations. In addition, the o1-model is instructed to report on which interviews a claim is based. The latter is included to increase transparency in order to detect hallucinations. Claims in this output can be traced back to the interview summary, which in turn can be traced back to the transcripts.

This resulted in a 'model' report with 3 to 6 claims per theme concerning 'what is going well' and 1 to 5 claims concerning 'what could be improved'. The claims in the model report had the following ontology: **Label: content (source)**. Below is an example of a claim for the theme Leadership: *Clear support from top management for safety: In multiple interviews, it is emphasized that the management supports safety in both words and actions. Incidents are systematically discussed in management meetings, and it is explicitly stated that risks are not tolerated if safety cannot be guaranteed. (Interview 1, 2, 5)*

2.3 Validation framework

We decided upon a validation framework based on three factors: Fluency, Correctness and Citation quality [8].

[†] https://github.com/vectara/hallucination-leaderboard

[‡] https://github.com/m-bain/whisperX

Fluency will not be considered here further, as on face validity, the output seemed sufficiently fluent. Instead we focussed on the remaining two factors Correctness (i.e., is the output objectively correct?) and Citation quality (is the output based on the provided context?). Correctness was assessed by determining:

- *Exhaustiveness.* Comparison of the claims in the model generated summaries of the interviews with the notes made by the interviewers. The majority of the claims generated by the o1-model should be similar to what the human researchers identified. Comparison was based on the labels and content;
- *Discriminant capacity.* Comparison of the claims in the model report for two different organisations (i.e., Organisation A and B). Comparison was primarily based on the labels, the content was used to resolve unclarities or doubts;
- *Consistency*. Comparison of the claims in subsequent model reports generated for the same organisation. We will refer to the original model report as version 1, whereas the additional model report generated for this comparison will be referred to as version 2. Comparison was primarily based on the labels, the content was used to resolve unclarities or doubts.

Citation quality was assessed by determining the occurrence of *Hallucinations*; i.e., the number of times that content within the claims in the model report could not be traced back to the original transcripts.

3 Results

3.1 Exhaustiveness

Not all themes were addressed in all interviews. We looked for which themes in each interview the interviewers had made notes (indicating they had discussed the topic) and for which themes in each interview the model generated claims in the summary. Comparison showed that in 68 cases (69.4%) out of 98 possible cases (7 interviews over 14 themes) both claims and notes had been generated and in 8 cases (8.2%) no notes and no claims had been generated. In 16 cases (16.3%) only the model had generated claims, and in 6 cases (6.1%) only the interviewers had made notes.

A comparison of the contents of the notes and claims for a sample of four themes *(Productivity versus safety, Role of the supervisor regarding safety, Maintenance management and Complexity and resilience*) showed that 42 out of 44 statements (95%) made in the notes of the interviewers were also addressed in the claims generated in the summaries by the model. This overlap was reduced to 64.3% (27 statements out 42) when comparing the claims in the model reports generated by the model and the interviewers.

3.2 Discriminant capacity

In the model report for organisation A, a total of 114 claims were made, 62 concerning ‘what is going well’ and 52 concerning ‘what can be improved. Organisation B had 127 statements in total, 66 and 61 respectively. Overlap between the claims was based on whether the claim addressed a similar theme (see Table 1, the comparison between Organisation A (version 1) and Organisation B). In total, 33 claims out of 241 (on average 27.4%) were found to overlap between the output for both organisations. For ‘what goes well’ the overlap concerned was slightly higher with 21 claims out of 128 (32.8%), as opposed to an overlap of 12 claims out of 113 (21.2%) for ‘what could be improved’.

3.3 Consistency

Consistency was determined by comparing the first report with a second report that was generated for the same organisation with the same transcripts. In the model report of version 1, 114 claims were made, 62 concerning ‘wat is going well’ and 52 concerning ‘what can be improved. and in the model report of version 2, 112 claims were made, 58 and 54 respectively. Again, overlap was determined based on whether the claim addressed a similar theme (see Table 1, the comparison between Organisation A (version 1) and Organisation A (version 2)). In total, 171 claims out of 226 (on average 75.7%) were found to overlap between the output for both organisations. For ‘what is going well’ the overlap found was for 92 claims out of 120 total (76,7%) and 79 claims out of 106 (74.5%) for ‘what could be improved’.

Table 1. Number of claims analysed to determine discriminant capacity and consistency, with overlap in brackets.

	Organisation A (version 1)	Organisation B	Organisation A (version 2)
What is going well	62	66 (32.8%)	58 (76.7%)
What can be improved	**52**	61 (21.2%)	54 (74.5%)
Total	*114*	127 (27.4%)	112 (75.7%)

3.4 Hallucinations

The presence of hallucinations in the final report were manually checked for a sample of 4 themes *(Safety communication*, *Incident registration and analysis*, *Relationship between cavern stability and personal safety* and *Complexity and resilience*). These themes contained 34 claims divided over ‘what is going well’ and ‘what can be improved’. 32 of these claims (94.1%) in the model report were traceable to 65 claims in the summaries which were subsequently traced back to passages in the original transcripts. The two remaining claims could not be traced back to claims in the summaries or passages in the transcripts. The

'hallucinations' did not contain falsehoods however, instead they concerned an interpretation or conclusion based on previous claims generated for that theme. One 'hallucination' stated *These reassessments can help continuously improve both the technical condition (e.g., corrosion, subsidence) and the organizational assurance (responsibilities, emergency scenarios).* This is not a wrong conclusion in the context of the interviews, but was also not stated explicitly as such in the interviews.

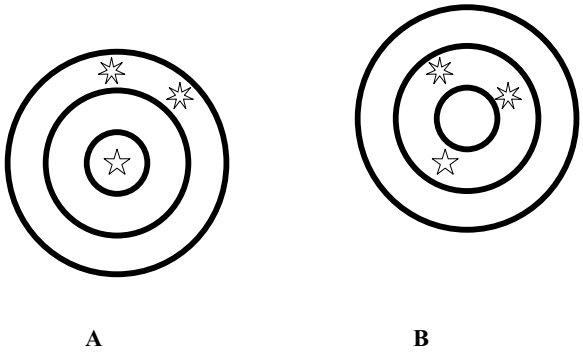
4 Conclusion

The aim in this paper was to determine whether AI, specifically the o1-model, is a valid tool for efficient and reliable support in summarizing interview-transcripts. Our findings lead us to answer this in the affirmative.

Directly supportive to this conclusion were the findings that the generated reports were capable of discriminating between organisations with only an overlap of 27.4% in claims, while also being relatively consistent with 75.7% overlap in claims between two reports generated for the same organisation. It is not surprising that some overlap was found between the model reports of Organisation A and B. It shows that the model reports primarily consisted of organisation specific claims, while some claims are likely to be encountered in any organisation (e.g. *Leadership is both visible and approachable on the work floor*). Similarly, the result in relation to the consistency shows that the o1-model will replicate the majority of the claims when repeating the analysis. It is not surprising that the model reports show deviations. Human researchers would also develop different reports when asked to analyse a set of transcripts twice (with no memory of the other analysis).

We employed a two-stage method to convert the interview transcripts into model reports. The first step was to summarize the individual interviews, and the second step was to create a synthesis of these summaries. For the summaries (step 1 in our approach) we found a near perfect overlap with the notes of the interviewers (95%). However, the overlap of the model report (step 2) with the human-based reports was lower, with 64.3%, compared to previous studies that found 70 to 80% overlap between model and human generated themes [1,4]. This could be the result of the two-staged approach in which consecutive rounds of interpretations by the model may have caused a greater divergence from human interpretation. Another contributing factor was the tendency of interviewers to list specific examples of applications in practice as separate statements, which the model did not.

It should be noted here that an interesting challenge in interpreting these results is that the objective truth (i.e., what is the best qualitative summary of the interviews that represents the true situation at the organisation) is actually unknown. Here we worked with the assumption that the interpretation of the interviewers is the correct one, and deviations of the o1-model from their interpretations would then be problematic. However, in reality this might be far more nuanced. Both the interpretation of the interviewers and by the model are likely not 100% accurate. The 'true' interpretation may lie somewhere in between the differences in interpretation observed here (see figure 1).



A **B**

Fig. 1. Situation A is the comparison between the report based on the expert analysis of the interviewers (5-pointed star) and the generated model reports (7-pointed star) under the assumption that the interviewers are 100% accurate regarding the safety culture. The actual situation might be more like B, with the true interpretation somewhere in the middle.

To leverage the differences between human and AI-driven analysis as a strength rather than a limitation, a mixed-methods approach is recommended in which both sources are treated as equally valid. By systematically examining the overlap and discrepancies between the findings of the o1-model and those of human coders, blind spots can be uncovered as well as complementary insights. This contrastive analysis can serve as a reflective tool, where findings identified by only one method are revisited in the raw data or discussed with subject matter experts (not involved in the project) or the interviewees themselves. That step was not set in this research yet. Such triangulation not only enhances the transparency and validity of interpretations but also promotes a more critical awareness of both human bias (i.e. susceptibility to confirmation bias, filtering and framing of information and socially desirable interpretations of results) and model-dependence (i.e. reproducing bias in dataset, the tendency to overemphasize repetition and may undervalue rare but meaningful statements).

A relatively high percentage of hallucinations was found (2 out 34 claims, 5.9%). So even though the presence of hallucinations is problematic, they are not per se detrimental to the reliability of the generated output. Especially if the interviewers act as a final barrier to avoid such hallucinations from making it into the final report. However, as complacency and over-reliance on the o1-model are realistic risks when employing AI to support a task and given that potential inaccuracies of the output of LLMs was identified by Barak-Corren and colleagues as a challenge that constrains acceptance to use LLMs [2], steps will be taken to reduce the number of hallucinations in the development of our approach. This will be done both by adopting new and better LLMs as they become efficient and through prompt engineering to improve the instructions given to the model.

Taking the above into consideration we emphasize the conclusion, shared with previous authors [6,9], that

LLMs primarily assist and should not replace researchers [6,9]. Keeping researchers in the loop when using an LLM improves the consistency and reliability of the output[1]. The current approach has been developed under the assumption that the model report generated can act as a starting point for the interviewers to write the definitive report. As such, it is recommended to generate the reports in close succession to finalizing the interviews for the interviewers to still have some recollection of what has been addressed. In addition, it will remain good practice for the interviewers to take notes during the interview. However, the interviewer can focus on making notes about important or remarkable statements, rather than taking minutes of the entire interview. This will allow the interviewer to focus more on the conversation.

Based on the results described in this paper we conclude that the use of Generative AI has great potential to increase the efficiency of interview-based studies. It does so by assisting interviewers through the summarization of interview transcripts into a concise overview of relevant claims ordered by the discussed themes. We will continue to fine-tune our current approach through prompt-engineering. In addition, we intend to build on the summarization by LLMs to include more suggestions to assist the interviewer in writing their reports. Next steps will include instructing the model to provide recommendations directly related to found claims concerning 'what can be improved' and to assign the organization a safety culture ladder score from 1 (pathological safety culture) to 5 (generative safety culture) in line with the original approach [12] based on generated reports. These pieces of information are of interest to many organisations to determine where they stand now and what steps they can take to improve their safety culture performance. As such the responsible application of LLMs in support of this work will result in more efficient and thus better use of interview data to support an analysis of the safety culture of organisations, ultimately leading to a more sound foundation for recommendations to help organisations to become safer.

Conceptualization, D.v.d.B. and W.M.P.S.; methodology, J.v.d.L. and W.M.P.S.; validation, D.v.d.B., J.G. and W.M.P.S.; formal analysis, J.v.d.L. and W.M.P.S.; writing - original draft preparation, J.v.d.L. and W.M.P.S.; writing - review and editing, D.v.d.B., J.v.d.L., J.G. and W.M.P.S.; supervision, D.v.d.B.; project administration, D.v.d.B. All authors have read and agreed to the published version of the manuscript.

The study was conducted in accordance with the Declaration of Helsinki and approved by Institutional Review Board (or Ethics Committee) of TNO (8 February 2024; 2024-007).

Informed consent was obtained from all subjects involved in the study.

This research was funded by the State Supervision of the Mines in the Netherlands.

The authors declare no conflict of interest.

References

1. M.R. Prescott, S. Yeager, L. Ham, C.D. Rivera Saldana, V. Serrano, J. Narez, ... & J. Montoya. Comparing the efficacy and efficiency of human and generative AI: Qualitative thematic analyses. JMIR AI. **3**, e54482 (2024)
2. Y. Barak-Corren, R. Wolf, R. Rozenblum, J.K. Creedon, S.C. Lipsett, T.W. Lyons, ... & A.M. Fine. Harnessing the power of generative AI for clinical summaries: perspectives from emergency physicians. Ann. Emerg. Med. **84**(2), 128-138 (2024)
3. S. Qiao, X. Fang, C. Garrett, R. Zhang, X. Li, & Y. Kang. Generative AI for Qualitative Analysis in a Maternal Health Study: Coding In-depth Interviews using Large Language Models (LLMs). medRxiv. 2024-09 (2024)
4. L. Hamilton, D. Elliott, A. Quick, S. Smith, & V. Choplin. Exploring the use of AI in qualitative analysis: A comparative study of guaranteed income data. Int. J. Qual. Methods. **22**, 16094069231201504 (2023)
5. F. Pattyn. The value of generative AI for qualitative research: A pilot study. JDSIS. (2024)
6. R.H. Tai, L.R. Bentley, X. Xia, J.M. Sitt, S.C. Fankhauser, A.M. Chicas-Mosier, & B.G. Monteith. An examination of the use of large language models to aid analysis of textual data. Int. J. Qual. Methods. **23**, 16094069241231168 (2024)
7. G.I. Zwetsloot, J. van Middelaar, & D. Van der Beek. Repeated assessment of process safety culture in major hazard industries in the Rotterdam region (Netherlands). J. Clean. Prod. **257**, 120540 (2020)
8. T. Gao, H. Yen, J. Yu, & D. Chen. Enabling large language models to generate text with citations. arXiv preprint arXiv:2305.14627 (2023)
9. M. Bain, J. Huh, T. Han, & A. Zisserman. Whisperx: Time-accurate speech transcription of long-form audio. arXiv. arXiv:2303.00747 (2023)
10. S. Schulhoff, M. Ilie, N. Balepur, K. Kahadze, A. Liu, C. Si, ... & P. Resnik. The prompt report: A systematic survey of prompting techniques. arXiv. arXiv:2406.06608 (2024)
11. D.L. Morgan. Exploring the use of artificial intelligence for qualitative data analysis: The case of ChatGPT. Int. J. Qual. methods. **22**, 16094069231211248 (2023)
12. D. Parker, M. Lawrie, & P. Hudson. A framework for understanding the development of organisational safety culture. Saf. Sci. **44**(6), 551-562 (2006)