AI Ethics Journal

Volume 3, Issue 1
Fall 2023
*Special Issue Edition

Risks and Impacts of Generative AI

AI and Human Reasoning: Qualitative Research in the Age of Large Language Models

Muneera Bano[1], Didar Zowghi[1], Jon Whittle[1]
1 CSIRO's Data61, Australia.

DOI: https://doi.org/10.47289/AIEJ20240122

Abstract

Context: The advent of AI-driven large language models (LLMs), such as Bard, ChatGPT 3.5 and GPT- 4, have stirred discussions about their role in qualitative research. Some view these as tools to enrich human understanding, while others perceive them as threats to the core values of the discipline. **Problem:** A significant concern revolves around the disparity between AI-generated classifications and human comprehension, prompting questions about the reliability of AI-derived insights. A minimal overlap between AI and human interpretations amplifies concerns about the fading human element in research. **Objective:** This research is exploratory and aims to compare the comprehension capabilities of humans and LLMs, specifically Google's Bard and OpenAI's ChatGPT 3.5 and GPT-4. **Methodology:** We conducted an experiment with a sample of Alexa app reviews, initially classified by two human analysts against Schwartz's human values. Bard, ChatGPT 3.5 and GPT-4 were then asked to classify these reviews and provide the reasoning behind each classification. We compared the results of LLMs with human classifications and reasonings. **Results:** The results revealed varied levels of agreement between AI models and human analysts concerning their interpretation of Schwartz's human values. ChatGPT showed a closer alignment with certain human perspectives, though overall comparisons displayed more disagreements than agreements. **Conclusion:** Our results highlight the potential for effective human-LLM collaboration, suggesting a synergistic rather than competitive relationship. Researchers must continuously evaluate LLMs' role in their work, thereby fostering a future where AI and humans jointly enrich qualitative research.

History

Received 22 June 2023

Accepted 31 October 2023

Published 13 December 2023

Keywords

Large Language Models, Qualitative Research, ChatGPT, GPT4

Contact

Muneera Bano

Senior Research Scientist

CSIRO's Data61

Research Way, Clayton VIC 3168

Email: muneera.bano@csiro.au

Acknowledgements

None

Disclosure of Funding

None

Introduction

Generative AI models, particularly large language models (LLMs) such as Google's Bard, OpenAI's ChatGPT 3.5 and GPT-4, are becoming increasingly sophisticated, offering potential applications in a variety of fields (Van Dis et al., 2023). These advanced AI applications have been meticulously designed and trained on vast datasets, allowing them to generate human-like text to answer questions, write essays, summarise text, and even engage in conversations (Dergaa et al., 2023). The promise they offer is not just in their ability to process information but also in their potential to mimic human-like comprehension and generation of text (Byun, Vasicek, and Seppi, 2023).

The transformative influence of these LLMs is being felt across a variety of fields, but perhaps one of the most intriguing applications lies in the domain of qualitative research. Qualitative research is an exploratory approach used to gain a deeper understanding of underlying reasons, opinions, and motivations. It involves collecting non-numerical data, often through methods like interviews, focus groups, or observations, to explore concepts, phenomena, or experiences in depth (Hennink, Hutter, and Bailey, 2020). This form of research provides rich descriptive insights that help in grasping the nuances and complexities of human behaviour and social contexts (Aspers and Corte, 2019).

Traditionally, qualitative research has always hinged upon the unique human ability to interpret nuances and discern underlying meanings from complex, often ambiguous data. However, the advent of LLMs, with their ability to handle large volumes of data, identify intricate patterns, and generate contextually appropriate responses, has sparked curiosity about their possible roles in qualitative research. The confluence of LLMs and qualitative research offers tantalizing possibilities, but it also raises profound questions. How reliable and valid are AI-generated interpretations compared to those derived from human understanding? What are the implications if the two do not align?

Since the debut of OpenAI's ChatGPT in November 2022, there has been a surge of academic interest in analyzing its potentials across various fields of study. A survey of Scopus[1] revealed that 587 papers reference ChatGPT in their titles or abstracts, with a distribution that spans diverse domains: 247 from medicine, 147 from social sciences, 116 from computer science, and 83 from engineering. Google Scholar[2] lists 7200 articles with ChatGPT mentioned in their titles, with dominant entries coming from the health and education sectors. This indicates an unexpectedly swift adoption of this AI tool by researchers in the health and social science disciplines. A systematic review (Sallam, 2023) of health education using ChatGPT shows 85% of the 60 included records praised the merits of ChatGPT, underlining its effectiveness in improving scientific writing, research versatility, conducting efficient data analysis, generating code, assisting in literature reviews, optimizing workflows, enhancing personalized learning, and bolstering critical thinking skills in problem-based learning, to name a few.

However, employing LLMs in specialized research could potentially introduce issues such as inaccuracies, bias, and plagiarism. Upon tasking ChatGPT with a set of medical research queries related to depression and anxiety disorders, it was observed that the model often produced inaccurate, overblown, or misleading as reported in (Van Dis et al., 2023). These errors could arise from an inadequate representation of relevant articles in ChatGPT's training data, an inability to extract pertinent information, or a failure to distinguish between credible and less credible sources. Evidently, LLMs may not only mirror but potentially amplify human cognitive biases (James Manyika, 2019) such as availability, selection, and confirmation biases (Kliegr, Bahník, and Fürnkranz, 2021; Bertrand et al., 2022).

[1] Search was conducted on 14th June 2023 with just one keyword "ChatGPT" to appear in title, abstract of keywords of the published papers.

[2] Search was conducted on 14th June 2023 with just one keyword ChatGPT to appear in title of the published papers.

The discourse concerning the potential replacement of humans by machines, and the capacities in which this may occur, has already gained significant momentum (Chui, Manyika, and Miremadi, 2016; Michel, 2020; Prahl and Van Swol, 2021). A parallel debate addresses the degree to which machines, particularly artificial intelligence, embody elements of human traits or personhood. Scholarly literature includes investigations into the theoretical creative autonomy attributed to AI poets (Amerika, Kim, and Gallagher, 2020), the experiences ostensibly undergone by our smart devices (Akmal and Coulton, 2020), and provocative inquiries into the existence of souls within voice assistants (Seymour and Van Kleek, 2020). Adding a dramatic dimension to this discourse, a former Google engineer postulated that Google's language models, specifically LaMDA (Thoppilan et al., 2022), possess sentience and are therefore entitled to rights typically reserved for humans (Griffiths, 2022).

In response to the doomsday hype of 'LLMs replacing the Human Researcher' (Cuthbertson, 2023), our research aims to explore and examine the alignment between human and AI comprehension. We designed an experiment using Schwartz's human values framework (Schwartz, 2012). Specifically, we delve into the comparison of LLMs-driven and human classifications of Alexa voice assistant app reviews, as they provide rich and diverse qualitative data. Our goal is to understand the extent to which LLMs can replicate or align with human understanding and the implications of any misalignment.

The contribution of our research lies in providing much-needed insights, derived primarily from an experiment, into the intersection of AI and qualitative research, a rapidly evolving area with significant implications for the future of the field. By exploring the capabilities and limitations of LLMs in understanding and interpreting qualitative data, we offer a valuable contribution to the ongoing discourse around AI's role in qualitative research.

The organization of this article adheres to the following structure: Section 2 contextualizes the research through an exploration of background information and a review of relevant literature. The subsequent Section 3 describes the design of our exploratory experiment. The results of the investigation are presented in Section 4, which is followed by a discussion in Section 5 that contemplates the repercussions of these results and offers critical insights applicable to qualitative research methodologies. Finally, Section 6 outlines the limitations of our research and Section 7 draws the study to a close by presenting a conclusion and delineating potential avenues for future research.

Background and Related Work

Large Language Models such as ChatGPT, have generated extensive interest and research across various academic fields in less than a year of its launch. The existing literature on the topic covers several domains, including AI's application in research and academia, its role in education, its performance in specific tasks, and its use in particular sectors like library information centers and medical education.

Numerous studies have explored the capabilities and limitations of AI in research and academia. Tafferner et al. (2023) analyzed the use of ChatGPT in the field of electronics research and development, specifically in applied sensors in embedded electronic systems. Their findings showed that the AI could make appropriate recommendations but also cautioned against occasional errors and fabricated citations.

Kooli (2023) delved into the ethical aspects of AI and chatbots in academia, highlighting the need for adaptation to their evolving landscape. Echoing this sentiment, Qasem (2023) explored the potential risk of plagiarism that could stem from the misuse of AI tools like ChatGPT. To balance the benefits and potential misuse, Burger et al. (2023) developed guidelines for employing AI in scientific research processes, emphasizing both the advantages of objectivity and repeatability and the limitations rooted in the architecture of general-purpose models.

The role of AI in education is another pivotal theme in the literature. Wardat et al. (2023) investigated stakeholder perspectives on using ChatGPT in teaching mathematics, identifying potential benefits and limitations. Similarly, Yan (2023) explored the use of ChatGPT in language instruction, pointing to its potential but also raising concerns about academic honesty and educational equity. Jeon and Lee (2023) examined the relationship between teachers and AI, identifying several roles for both and emphasizing the continued importance of teacher's pedagogical expertise. In a broader study of public discourse and user experiences, Tlili et al. (2023) highlighted a generally positive perception of AI in education but also raised several ethical concerns.

A strand of research has also evaluated AI's performance in specific tasks traditionally conducted by humans. Byun, Vasicek, and Seppi (2023) showed that AI can conduct qualitative analysis and generate nuanced results comparable to those of human researchers. In another task-specific study, Gilson et al. (2023) demonstrated that ChatGPT could answer medical examination questions at a level similar to a third-year medical student, underscoring its potential as an educational tool.

Research has explored the use of ChatGPT in specific sectors. Panda and Kaur (2023) investigated the viability of deploying ChatGPT-based chatbot systems in libraries and information centers, concluding that the AI could provide more personalized responses and improve user experience. Similarly, Gilson et al. (2023) indicated the potential of ChatGPT as an interactive medical education tool, further expanding the potential application areas of AI in different sectors.

Although existing literature has extensively covered AI's impact in various domains, several gaps remain. Notably, a lack of qualitative research comparing human reasoning against LLMs is evident. Burger et al. (2023), and Byun, Vasicek, & Seppi (2023) have made an initial foray into this area, demonstrating that ChatGPT can perform certain research tasks traditionally undertaken by human researchers, producing complex and nuanced analyses of qualitative data with results arguably comparable to human-generated outputs. Despite these promising findings, these studies do not investigate AI and human reasoning within the qualitative research context.

Experiment Design

The aim of our research was to compare human comprehension with that LLMs, specifically within the context of qualitative research. We sought to understand the depth to which these LLMs could analyze and provide reasoning for their judgment. Our exploratory research was guided by the following research question:

RQ: How do the analytical reasoning abilities of LLMs compare to human comprehension in the context of qualitative research?

To design and conduct our experiment (see Figure 1), we leveraged the framework of Schwartz's theory of human values, a well-regarded model that encapsulates ten basic universal values present across cultures (Schwartz, 2012). These include power, achievement, hedonism, stimulation, self- direction, universalism, benevolence, tradition, conformity, and security. This conceptual schema enabled us to perform a comparative analysis between human and AI reasoning within a structured and widely accepted paradigm of human values.

We designed an experiment to explore through a case of the Amazon Alexa voice assistant app's reviews. These reviews provide a rich source of qualitative data, with users expressing their opinions, perceptions, and values implicitly or explicitly in their feedback (Shams et al., 2021). We randomly selected a sample of 50 Alexa app reviews from a set that had previously been classified by a human analyst according to Schwartz's human values (Shams et al., 2023). This study created a benchmark for our comparison with the classifications generated by the

LLMs. Shams et al. (2013) in their study were aiming to conduct an empirical analysis of user feedback for Amazon's Alexa app to identify a set of essential human values and validate them as requirements for AI systems within distinct usage contexts, a technique that could potentially be extrapolated to other AI platforms.

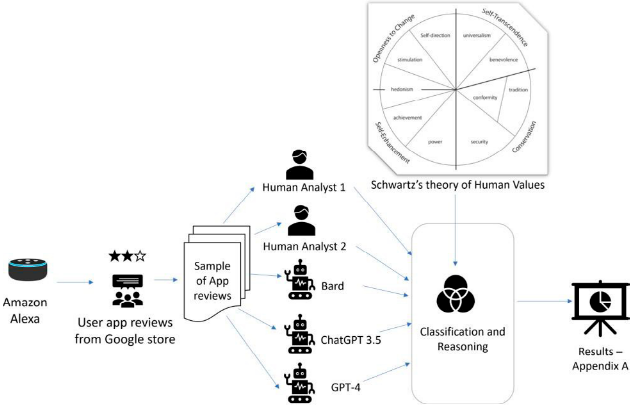


Figure 1. Experiment Design

Though a randomized selection of 50 reviews represents a limited sample, our primary objective was not to focus on the sample size but to explore the variation in responses between the human analyst and LLMs. Our priority was mainly on the 'why' aspect of all classifications.

Our experiment involved prompting Google's Bard and OpenAI's ChatGPT 3.5 and GPT-4 to generate their classifications for the same reviews. We were interested not only in their classification outcomes but also in their rationale for each categorization. This design aimed to gauge the LLMs' depth and their capability to reason within the context of Schwartz's human values framework as compared to human comprehension.

Designing appropriate prompts for LLMs is a critical process (White et al., 2023) and it can significantly influence the outcome of any LLMs' analysis. The composition and specificity of prompts can guide the models' analysis and processing of the task, thus affecting the results. A well-structured, clear, and contextually rich prompt helps the LLMs focus on the essential aspects of the task, reducing the likelihood of errors or misinterpretations. For every individual app review, we used the same prompt as below for all three LLMs:

Following is an app review from a user of Amazon Alexa. Analyse the review text and classify it against Schwartz's theory for Human Values, both main and sub values. Provide your reason on why you classified it against that value.

Our prompt design considered these three elements (structure, clarity, and context). Firstly, the prompt is clear as it explicitly outlines the task at hand, namely the analysis of an Amazon Alexa user review. Secondly, it displays structure, sequentially detailing each step to be undertaken, beginning with the review analysis, followed by classification against Schwartz's theory of Human Values, and concluding with an explanation for the chosen classification. Lastly, the prompt provides the context; it not only specifies the source of the review (Amazon Alexa) but also guides the model to employ a particular theoretical framework (Schwartz's theory of Human Values). Such specificity allows the model to tune its responses based on the understanding of the context provided, including both main and sub-values, thereby facilitating a more nuanced analysis.

The first author conducted the entire experiment with LLMs. To triangulate the results obtained from the comparisons and further discuss the findings and insights, the second and third authors then conducted an independent review of the results obtained from the human analysts and LLMs to form an opinion about the reasonability of the classifications.

Results

The detailed results of the 50 app review classifications are provided as an online artefact[3]. Looking at the agreements and disagreements on the classification of main values of Schwartz's human values (see Figure 2), several intriguing observations were noted.

[3] https://docs.google.com/spreadsheets/d/1iy5Rl0BvsH4DukEcuI2YQlOroYzX3LGf/edit#gid=1473701098

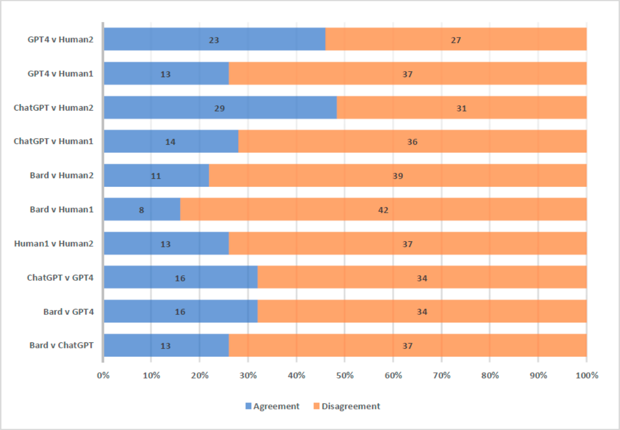


Figure 2. Agreements vs Disagreement Chart for main values

Among the AI models, ChatGPT seems to be closer in its interpretations to both other AI models (Bard and GPT4) and humans, especially Human2. While AIs show varied levels of agreement with humans, it is noteworthy that ChatGPT has a significant agreement with Human2, suggesting that certain AI models might align more closely with certain human perspectives. All combinations show more disagreements than agreements, indicating the inherent diversity in interpretation among both humans and AIs.

AI vs AI Comparisons: The highest agreement among the AI models is seen between Bard v GPT4 and ChatGPT v GPT4 both at 16 out of 50. Bard v ChatGPT has a slightly lower agreement at 13 out of 50. The AI models generally have more disagreements than agreements, with disagreements ranging from 34 to 37 out of 50.

Human vs Human Comparisons: Human1 v Human2 have an agreement of 13 out of 50, which is similar to some of the AI vs AI comparisons. This suggests that human interpretations can be as varied as the discrepancies between AI models. It is noteworthy that Human1 has more extensive knowledge and expertise with Schwartz's human values theory compared to Human2.

AI vs Human Comparisons: Bard has the lowest agreement with Human1 at only 8 out of 50, whereas its agreement with Human2 is slightly higher at 11 out of 50. ChatGPT shows a marked difference in its agreement with the two humans. It has a higher agreement with Human2 at 29 out of 50 compared to only 14 with Human1. This suggests that Human2's interpretations might be more in line with ChatGPT compared to Human1. GPT4 has a moderate level of agreement with both humans: 13 with Human1 and 23 with Human2.

The divergences found among the classifications trigger compelling questions about the reliability of the results generated by LLMs. The inconsistencies among the insights derived from LLMs and human interpretations lead to speculation about the capability of LLMs in fully appreciating and navigating the intricacies of human language and contextual nuances. This view is especially prevalent among qualitative researchers who consider these discrepancies as a warning that LLMs might not be adequately equipped.

Another facet that emerged from our analysis was that in some instances classifications made by ChatGPT 3.5 and GPT-4 appeared to be more logical and reasonable. This was determined by the triangulation conducted by the second and third authors comparing LLMs classifications to the humans. For example, in one of the instances in our review analysis, the human analyst classifies the review: "I'd enjoy and find this app very useful if it did WHAT it was supposed to WHEN it was supposed to" as "Benevolence" and "Loyalty". While ChatGPT 3.5 classifies it as "Achievement" and "Competence", and GPT4 mentions "Achievement" and "Capability". The two peer reviewers considered answers from LLMs to be more logical and reasonable than that of humans. This could suggest that LLMs can offer a fresh, alternative perspective that might not have been identified by human researchers.

In the analysis of app reviews, an intriguing observation emerged when two reviews, labelled 20 and 22, were purposefully repeated as 41 and 42. Human analysts recognized the repetition, classified the duplicated reviews identically both times, demonstrating a degree of consistency in interpretation. Contrarily, the LLMs, both Google's Bard and ChatGPT 3.5, treated the repeats as unique instances and displayed variations in their classifications and reasoning between the duplicates. Such discrepancies reveal a limitation in the LLMs' consistency within the same context of research. This could have implications, especially where consistency and recall of previous interpretations are paramount.

Table 1 showcases three examples where human analysts and LLMs, diverge significantly in their respective reasoning. The table reveals an intriguing variation in categorization and reasoning methodologies across two human analysts and three LLMs when analyzing app reviews. For Example 1, there's a notable divergence in interpretations: while humans perceived themes of helpfulness and achievement, the LLMs explored diverse values ranging from hedonism to security. In contrast, Example 2 presents a conceptual convergence on the app's challenges, though the reasoning differs subtly in terms of the values associated, indicating a consistent underlying sentiment yet varied nuances in its interpretation. Meanwhile, Example 3 exemplifies a consistent agreement among all analysts, highlighting the positive, hedonic sentiments expressed in the review. The findings underscore the complex nature of sentiment analysis, with humans and LLMs occasionally converging on shared interpretations or veering in distinct directions based on their unique inferential frameworks.

	Example 1 Divergent Categorizations	**Example 2 Conceptual Agreement**	**Example 3 Consistent Categorizations**
App Review	This is super easy to navigate and makes setting Amazon's Echo DOT super easy too... as well as other smart plugs, bulbs, etc. I did have some trouble earlier today where the app just suddenly didn't want to work even after clearing cache, force- stopping, and uninstalling. But that went away and I haven't had any problems...and hopefully, I won't have any in the future.	I only use Alexa for listening to Kindle books while I'm working on other tasks. That's it. But the process is buggy. Sometimes it repeats the same section of the book 5 times before moving on. Sometimes it just stops playing, then I need to close the app and open it again to start all over. The audio player doesn't always stay visible or accessible. Often times I'll just have to stop trying to use Alexa and give up on the book I'm listening to until I can deal with it again.	I have several devices. I have had them for years now. I love the music and I love the sound quality. Also, the unbelievable amount of music available. I seriously have not been able to ask for something that is not available. If so, I just ask a different way and it was played. 100% the best music selection and sound quality available. That is my opinion, and I know music. Firestick HDMI streams the ultimate HD sound waves. You select your system and you will not be disappointed.
Human 1 Categorization	Benevolence and Helpful Reason: It is helpful as it is easy to navigate and easy to set up.	Universalism and Achievement Reason: It helps the users to enrich wisdom, as it plays books. However, it is not always intelligent to move on with book chapters, rather playing the same sections again and again.	Hedonism and Enjoying Life Reason: The user is satisfied with the music, sounds like he/she enjoyed the experience with Alexa regarding the recreation. Additionally, Alexa is intelligent enough to play music even if it is asked in a different way.

	Example 1 Divergent Categorizations	**Example 2 Conceptual Agreement**	**Example 3 Consistent Categorizations**
Human 2 Categorization	Achievement and Capability Reason: Addressed: Reviewer finds it easy to use the app and navigate its features.	Achievement and Capability Reason: Ignored: Reviewer faces issues using Alexa for a specific task, diminishing their productivity.	Hedonism and Pleasure Reason: Addressed: Reviewer enjoys the music quality and vast selection offered by Alexa.
Bard Categorization	Hedonism and Enjoyment of Life Reason: The reviewer mentions that Alexa is easy to use and that it's cool to use. This shows that the reviewer values enjoyment of life.	Hedonism and Stimulation Reason: The reviewer is expressing frustration with Alexa's unreliability and the company's failure to provide a smooth and enjoyable user experience. This suggests that they value novelty and excitement.	Hedonism and Stimulation Reason: The reviewer is expressing their love of music and Alexa's ability to provide a high-quality listening experience. This suggests that they value novelty and excitement.
ChatGPT Categorization	Universalism and Equality Reason: The review praises the app's ease of navigation and its role in setting up devices, with a mention of a temporary issue, aligning with the value of universalism and equality.	Hedonism and Pleasure Reason: The review discusses the frustration of encountering bugs while using the app for listening to Kindle books, related to the value of pleasure.	Hedonism and Pleasure Reason: The review highlights the pleasure derived from the app's music selection and sound quality, related to the value of hedonismo and pleasure.
GPT-4 Categorization	Security and Safety Reason: The reviewer appreciates the app's ease of use and the security it provides in controlling other smart devices, reflecting a value for safety in their digital and physical environment. A smooth, problem-free experience contributes to their sense of safety and stability.	Stimulation and Excitement Reason: The user seeks stimulation and excitement through multitasking but is met with disappointment due to the app's buggy nature, which interrupts their experience.	Hedonism and Pleasure Reason: The reviewer finds pleasure in the music selection and sound quality provided by the app and devices, indicating a value for experiences that bring personal enjoyment and satisfaction.
Insights	There is quiet a discrepancy between the reasoning provided by the human analysts and the LLMs. While both human analysts focused on the ease of use and capability, relating it to 'Benevolence', 'Helpful', and 'Achievement', the LLMs varied significantly in their interpretations. Bard leaned toward 'Hedonism' and the enjoyment aspect of the app, highlighting the coolness factor. In contrast, ChatGPT focused on 'Universalism and Equality', emphasizing the app's wide-reaching	Both Human 1 and ChatGPT associate the review with 'Hedonism', relating the app's problematic functionality to a lack of pleasure derived from its use. However, where Human 1 also brings in 'Universalism and Achievement' due to the app's educational utility, Human 2 and GPT-4 touch on 'Achievement and Capability' and 'Stimulation and Excitement', respectively. Both interpretations hint at the disappointment faced by the user	There is an agreement across both human analysts and LLMs around the theme of 'Hedonism' and pleasure derived from the product. Both human analysts clearly identified the reviewer's satisfaction with Alexa's music capabilities and sound quality, associating it with 'Hedonism and Enjoying Life' or 'Hedonism and Pleasure'. This sentiment was echoed by both ChatGPT and GPT-4, who similarly classified the review under 'Hedonism and Pleasure'. Bard's

	Example 1 Divergent Categorizations	Example 2 Conceptual Agreement	Example 3 Consistent Categorizations
	capability and inclusivity. GPT-4's interpretation was rather unique, associating the review with 'Security and Safety' - a perspective neither the human analysts nor the other LLMs touched upon. This divergence suggests a broader interpretative range among the LLMs, especially when dealing with reviews that may contain multiple themes or sentiments.	when their expectations were not met. Bard's analysis, however, took a slightly different angle, focusing on the 'Stimulation' aspect but associating it with the company's failure to provide novelty and excitement. While the underlying sentiment is consistent across interpretations, the nuances captured by each entity offer a multifaceted understanding of the review.	interpretation, while still centered around 'Hedonism', added an element of 'Stimulation', hinting at the excitement derived from the product. This example underscores instances where clear, positive sentiments in reviews lead to consistent categorizations across different evaluative entities.

Table 1. Comparison of three scenarios of Human vs AI agrément.

The variations between human interpretations underscore the subjectivity inherent in understanding and classifying feedback. While humans bring in personal biases, they also capture a depth and holistic understanding that's unique to human cognition. While AI models demonstrate proficiency in interpreting and classifying feedback, their understanding tends to be more structured and might miss out on the nuanced or emotional aspects that humans naturally grasp. However, the consistency of AI models can be valuable, especially when dealing with large datasets. On the other hand, human reviewers bring depth, context, and a broader perspective.

Some researchers assert that LLMs, given their current technological stature, are incapable of completely comprehending the profound complexities of human emotions and experiences (Bender et al., 2021; Alkaissi and McFarlane, 2023; Rudolph, Tan, and Tan, 2023). Consequently, their use in qualitative analysis should be treated with caution. The argument furthers that the LLMs missing context sensitivity and focus on functional aspects could lead to flawed or incomplete conclusions. But the prompts developed by humans need to provide a rich context in order to address this issue.

Contrastingly, advocates of AI-assisted qualitative analysis propose that LLMs can furnish invaluable insights and complementary viewpoints, aiding researchers in achieving a more all-encompassing understanding of the data (Dwivedi et al., 2023). The researchers in favour of LLMs further posit that with the consistent evolution and enhancement of AI, a synergistic approach combining human acumen and AI capabilities can lead to more robust analysis.

This ongoing discussion brings forth crucial questions for qualitative researchers concerning the degree of their reliance on LLMs in their work. While LLMs hold the potential to transform qualitative research by delivering additional perspectives and insights, it is imperative for researchers to also acknowledge their limitations and maintain a keen awareness of the humanistic elements inherent to qualitative research.

To answer our research question: *How do the analytical reasoning abilities of LLMs compare to human comprehension in the context of qualitative research?*

Answer: LLMs exhibit varied analytical reasoning abilities compared to human comprehension in the context of qualitative research. While some AI models (as in our experiment ChatGPT) may align more closely with certain human perspectives, there is inherent diversity in interpretation among both humans and LLMs. Notably, even human-to-human comparisons show discrepancies, suggesting that both LLMs and humans possess subjective interpretation capabilities in qualitative analysis.

Discussion

In this section, we move deeper into the broader implications of our findings. By situating our results within a wider context and comparison with existing research ideas, we aim to shed light on the overarching significance and potential impact these insights might have on the evolution of qualitative research.

AI and Humans

Despite the considerable potential of LLMs in qualitative research, the indispensable role of the human researcher for verifying the validity and reliability of the results remains critical. LLMs, while robust and efficient, exhibit limitations in their understanding of complex human experiences, contexts, and semantics, occasionally leading to the generation of inaccurate or invented information, a phenomenon known as 'hallucinations' (Rudolph, Tan, and Tan, 2023; Alkaissi and McFarlane, 2023). These hallucinations can misdirect the interpretation of research results, compromise validity, and introduce unintentional bias or error. Therefore, the human researchers' involvement becomes vital in scrutinizing, verifying, and interpreting the results generated by LLMs, ensuring that the outcomes are consistent with the actual context and preserving the integrity of the research. Furthermore, the human researcher's expertise and critical thinking are required to continually improve their comprehension over time, helping in enhancing their capabilities while minimizing potential drawbacks.

LLMs are poised to redefine the interplay between AI and human involvement in the research process. When it comes to inductive reasoning and open-ended data collection, LLMs are capable of deriving insights from unstructured data without predetermined hypotheses and continuously collecting and analyzing massive amounts of data from diverse sources. However, while these capabilities can expedite the research process, the question remains whether LLMs can truly replicate the intuitive reasoning processes and interpretive nuances inherent to human researchers. Similarly, while LLMs can process large amounts of qualitative data collected in naturalistic settings, the nuanced understanding, cultural sensitivity, and context-awareness that human researchers bring to these settings are unlikely replicable by LLMs in their current state.

Polanyi's concept of 'tacit knowledge' (Collins, 2005) which is also the 'implicit' component of Nonaka's SECI model (Li and Gao, 2003) underscores the unique human ability to perform certain tasks in unexpected and inexplicable ways. This inherent capability, however, may not be explicitly replicated or comprehended by LLMs, due to the unpredictable nature of such knowledge that is often grounded in personal experience and intuition.

A further manifestation of Human-LLM collaborative research could involve delineating distinct roles for each entity to optimize the research process. Here, LLMs could function as 'inter-rater reliability testers' (Armstrong et al., 1997), contributing to the research conducted by human analysts, while the human participants would be responsible for the verification of the information and analytical results generated by the LLMs. This iterative process, involving reciprocal roles, has the potential to yield more robust and efficient research outcomes, underscoring the mutual enrichment of human insight and machine efficiency.

Stochastic Parrots for Qualitative Research

LLMs have demonstrated remarkable capacity to generate human-like text, understand context, and interact dynamically with users. Their potential, however, should not overshadow the challenges they pose, especially concerning the interpretation of meanings in qualitative research. A significant advantage of these models lies in their ability to process and analyze vast amounts of data quickly and relatively accurately, providing a broad view of patterns and trends that could otherwise be missed.

Nonetheless, Bender et al. (2021) present cogent arguments about the risks associated with these models, primarily centered around their training on massive and diverse text datasets. This training can result in the replication and amplification of biases present in the data, leading to potentially harmful outputs. Additionally, the text generation process of LLMs remains fundamentally opaque, raising questions about transparency and interpretability.

Despite LLMs' adeptness at generating linguistically coherent responses, they do not genuinely comprehend the meanings, nuances, and deeper implications of words and phrases. While humans possess a holistic understanding of language, encompassing cultural, emotional, historical, and symbolic dimensions, LLMs can only provide approximations based on learned patterns. They may miss out on the rich tapestry of meanings a human researcher could decipher.

To mitigate the risks associated with LLMs, (Bender et al., 2021) propose several steps, including (a) reducing model size, (b) increasing transparency, and (c) establishing ethical guidelines for their use. Smaller, more controlled models could potentially minimize harm, while greater transparency could facilitate a better understanding of the mechanisms behind their text generation. Ethical guidelines would also establish a framework for responsible and equitable use of these models.

The determination of an appropriate size for LLMs, a balance between the model's complexity and its predictive accuracy, is best achieved through a collaboration of machine learning experts, ethicists, and domain-specific experts. As for the selection of ethical guidelines governing LLMs use should be context-dependent and reflective of the values and perspectives of a diverse range of stakeholders (Zowghi and da Rimini, 2023). This selection process necessitates an inclusive approach, possibly involving a blend of established ethical frameworks tailored to the specifics of the AI system and its deployment (Sanderson et al., 2023). The ethical guidelines established by a diverse and inclusive committee of stakeholders need to be periodically reviewed and updated to align with evolving societal norms and technological advancements.

The use of LLMs in qualitative research also introduces a new set of ethical considerations. Concerns around privacy, data misuse, and the risk of perpetuating existing biases in the data they are trained on are prevalent. Additionally, the advent of LLMs in the academic sphere raises questions about intellectual property rights and authorship. In this changing landscape, the role of the human researcher may shift towards orchestrating the research process, ensuring ethical compliance, and interpreting and contextualizing the findings generated by LLMs. As technology continues to advance, the importance of critical reflection on these shifts and their implications will grow.

Evolution of Qualitative Research

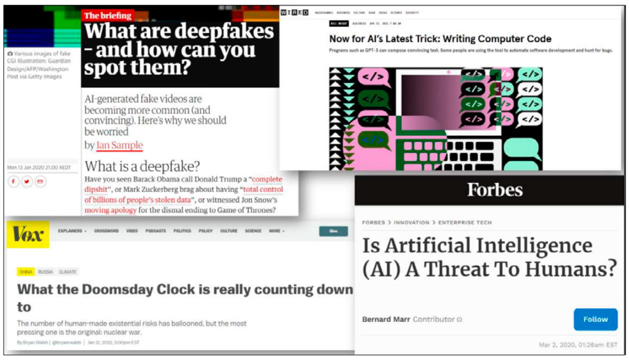
LLMs have the potential to significantly impact data analysis in qualitative research, as they can speed up the process and handle larger datasets than humans can feasibly manage. For example, Byun, Vasicek, and Seppi (2023) demonstrated that AI is capable of conducting qualitative analysis and generating nuanced results. However, such studies often do not delve into the reasoning behind AI vs. human interpretation, which could significantly impact the findings. In addition, there is the question of whether LLMs can truly understand and articulate the symbolic and cultural nuances that underpin human behavior, elements that are paramount to the work of prominent anthropologists and sociologists, such as Malinowski's participatory observation (Malinowski, 1929), Weber's concept of verstehen, or empathetic understanding (Weber, 1949), and Geertz's interpretation of culture (Geertz, 1973).

The application of LLMs could potentially enhance the efficiency of established qualitative methodologies such as Grounded Theory (Charmaz, 2014; Glaser, Strauss, and Strutzel, 1968), Interpretive Interactionism (Denzin 2001), and Narrative Analysis (Franzosi, 1998), particularly in terms of initial data analysis. However, these methodologies were developed with the understanding that the researcher's empathy, interpretation, and contextual understanding are integral to the process. As such, it is unlikely that the essential humanistic aspects of these approaches can be fully replaced by LLMs, indicating a shift rather than an absolute transformation in these methodologies (Dwivedi et al., 2023).

AI Doomsday

The escalating discourse on the potential risks of AI and LLMs, amplified by recent media reports (Figure 3), is leading to a growing unease among various professional communities. They are coming to terms with the stark reality that AI might soon eclipse their roles and replace them in their jobs. This existential dread has been underscored by developments such as the AI Doomsday Clock[4] inching closer to midnight, symbolizing the perceived imminent danger of a catastrophic AI disaster.

Findings from our exploratory experiment, coupled with an overview of existing research and an understanding of capabilities of LLMs, do not support a doomsday scenario for qualitative researchers. Contrary to pervasive fears, the reality we've discerned suggests a future where human researchers and LLMs can coexist and contribute complementarily to the field of qualitative research.



Various images of fake CGI Illustration: Guardian Design/AFP/Washington Post via Getty Images

The briefing

What are deepfakes – and how can you spot them?

AI-generated fake videos are becoming more common (and convincing). Here's why we should be worried

by Ian Sample

Mon 13 Jan 2020 21.00 AEDT

What is a deepfake?

Have you seen Barack Obama call Donald Trump a "complete dipshit", or Mark Zuckerberg brag about having "total control of billions of people's stolen data", or witnessed Jon Snow's moving apology for the dismal ending to Game of Thrones?

WIRED

Now for AI's Latest Trick: Writing Computer Code

Programs such as GPT-3 can compose convincing text. Some people are using the tool to automate software development and hunt for bugs.

Vox

EXPLAINERS CROSSWORD VIDEO PODCASTS POLITICS POLICY CULTURE SCIENCE MORE Give

CHINA RUSSIA CLIMATE

What the Doomsday Clock is really counting down to

The number of human-made existential risks has ballooned, but the most pressing one is the original: nuclear war.

By Bryan Walsh | @bryanrwalsh | Jan 21, 2022, 3:00pm EST

Forbes

FORBES > INNOVATION > ENTERPRISE TECH

Is Artificial Intelligence (AI) A Threat To Humans?

Bernard Marr Contributor

Follow

Mar 2, 2020, 01:26am EST

Figure 3. Media amplification of AI Doomsday fears

Limitations

While our study provides valuable insights into the utilization of LLMs in qualitative research, these findings are inevitably influenced by our own areas of expertise and the specific experimental design we employed. The research is also constrained by two primary limitations. Firstly, the sample size we chose for the study, although increasing the sample size could have altered the statistical outcomes. However, our primary interest lay not in large-scale data analysis, but in exploring the reasoning patterns of human analysts and LLMs during the classification of app reviews.

Conclusion and Future Work

The insights obtained from our experiment underscore the significance of careful considerations regarding the use of AI models play in qualitative research. The modest alignment between human and AI classifications, coupled with the comparatively higher concordance between the AI models, illuminates the complex dynamics at play when incorporating AI into qualitative analysis. Our findings accentuate that, despite the promise of AI for augmenting analysis, the unique human touch—an element intrinsic to qualitative research—cannot be disregarded. This

[4] https://www.vox.com/22893594/doomsday-clock-nuclear-war-climate-change-risk

essential human element, embedded in understanding and interpreting context, remains a critical factor (for now) in maintaining the richness and depth of qualitative investigations.

The considerable variations highlighted between human and AI comprehension in this study encourage further exploration in the field of AI integration into qualitative research. Future work could delve deeper into understanding the basis for such disparities, thereby refining the synergistic interplay between AI and human analysis. Furthermore, investigating how to leverage the different perspectives offered by AI, while keeping the human touch intact, could lead to more comprehensive and nuanced insights.

Lastly, addressing the ethical implications of AI usage in qualitative research, especially considering AI's limitations and potential for biases, will form a critical part of future studies. As we venture further into this new era of AI-assisted research, it is imperative to navigate these challenges to harness the full potential of this technological advancement in a responsible and ethical manner.

References

Akmal, Haider, and Paul Coulton. (2020). "The divination of things by things." In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 1-12.

Alkaissi, Hussam, and Samy I McFarlane. (2023). 'Artificial hallucinations in ChatGPT: implications in scientific writing', *Cureus*, 15.

Amerika, Mark, Laura Hyunjhee Kim, and Brad Gallagher. (2020). "Fatal error: artificial creative intelligence (ACI)." In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 1-10.

Armstrong, David, Ann Gosling, John Weinman, and Theresa Marteau. (1997). 'The place of inter-rater reliability in qualitative research: An empirical study', *Sociology*, 31: 597-606.

Aspers, Patrik, and Ugo Corte. (2019). 'What is qualitative in qualitative research', *Qualitative sociology*, 42: 139-60.

Bender, Emily M, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. (2021). "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big??" In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 610-23.

Bertrand, Astrid, Rafik Belloum, James R Eagan, and Winston Maxwell. (2022). "How cognitive biases affect XAI-assisted decision-making: A systematic review." In *Proceedings of the 2022 AAAI/ACM conference on AI, ethics, and society*, 78-91.

Burger, Bastian, Dominik K Kanbach, Sascha Kraus, Matthias Breier, and Vincenzo Corvello. (2023). 'On the use of AI-based tools like ChatGPT to support management research', *European Journal of Innovation Management*, 26: 233-41.

Byun, Courtni, Piper Vasicek, and Kevin Seppi. (2023). "Dispensing with Humans in Human-Computer Interaction Research." In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, 1-26.

Charmaz, Kathy. (2014). *Constructing grounded theory* (sage).

Chui, Michael, James Manyika, and Mehdi Miremadi. (2016). 'Where machines could replace humans-and where they can't (yet)'.

Collins, Harry M. (2005). 'What is tacit knowledge?' In *The practice turn in contemporary theory* (Routledge).

Cuthbertson, Anthony. (2023). 'Why tech bosses are doomsday prepping'. https://www.independent.co.uk/tech/chatgpt-ai-chatbot-microsoft-altman-b2274639.html.

Denzin, Norman K. (2001). *Interpretive interactionism* (Sage).

Dergaa, Ismail, Karim Chamari, Piotr Zmijewski, and Helmi Ben Saad. (2023). 'From human writing to artificial intelligence generated text: examining the prospects and potential threats of ChatGPT in academic writing', *Biology of Sport*, 40: 615-22.

Dwivedi, Yogesh K, Nir Kshetri, Laurie Hughes, Emma

Louise Slade, Anand Jeyaraj, Arpan Kumar Kar, Abdullah M Baabdullah, Alex Koohang, Vishnupriya Raghavan, and Manju Ahuja. (2023). '"So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy', *International Journal of Information Management*, 71: 102642.

Franzosi, Roberto. (1998). 'Narrative analysis—or why (and how) sociologists should be interested in narrative', *Annual review of sociology*, 24: 517-54.

Geertz, Clifford. (1973). 'Thick Description: Toward an interpretive theory of culture', *The interpretation of cultures: Selected essays*: 3-30.

Gilson, Aidan, Conrad W Safranek, Thomas Huang, Vimig Socrates, Ling Chi, Richard Andrew Taylor, and David Chartash. (2023). 'How does CHATGPT perform on the United States Medical Licensing Examination? the implications of large language models for medical education and knowledge assessment', *JMIR Medical Education*, 9: e45312.

Glaser, Barney G, Anselm L Strauss, and Elizabeth Strutzel. (1968). 'The discovery of grounded theory; strategies for qualitative research', *Nursing research*, 17: 364.

Griffiths, Max. (2022). 'Is LaMDA sentient?', *AI & SOCIETY*: 1-2.

Hennink, Monique, Inge Hutter, and Ajay Bailey. (2020). *Qualitative research methods* (Sage).

James Manyika, Jake Silberg, Brittany Presten. (2019). 'What Do We Do About the Biases in AI?'. https://hbr.org/2019/10/what-do-we-do-about-the-biases-in-ai.

Jeon, Jaeho, and Seongyong Lee. (2023). 'Large language models in education: A focus on the complementary relationship between human teachers and ChatGPT', *Education and Information Technologies*: 1-20.

Kliegr, Tomáš, Štěpán Bahník, and Johannes Fürnkranz. (2021). 'A review of possible effects of cognitive biases on interpretation of rule-based machine learning models', *Artificial Intelligence*, 295: 103458.

Kooli, Chokri. (2023). 'Chatbots in education and research: a critical examination of ethical implications and solutions', *Sustainability*, 15: 5614.

Li, Meng, and Fei Gao. (2003). 'Why Nonaka highlights tacit knowledge: a critical review', *Journal of knowledge management*.

Malinowski, Bronislaw. (1929). 'Practical anthropology', *Africa*, 2: 22-38.

Michel, Jan G. (2020). 'Could Machines Replace Human Scientists?: Digitalization and Scientific Discoveries.' in, *Artificial Intelligence* (Brill mentis).

Panda, Subhajit, and Navkiran Kaur. (2023). 'Exploring the viability of ChatGPT as an alternative to traditional chatbot systems in library and information centers', *Library Hi Tech News*, 40: 22- 25.

Prahl, Andrew, and Lyn M Van Swol. (2021). 'Out with the humans, in with the machines?: Investigating the behavioral and psychological effects of replacing human advisors with a machine', *Human- Machine Communication*, 2: 209-34.

Qasem, Fawaz. (2023). 'ChatGPT in scientific and academic research: future fears and reassurances', *Library Hi Tech News*, 40: 30-32.

Rudolph, Jürgen, Samson Tan, and Shannon Tan. (2023). 'ChatGPT: Bullshit spewer or the end of traditional assessments in higher education?', *Journal of Applied Learning and Teaching*, 6.

Sallam, Malik. (2023). "ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns." In *Healthcare*, 887. MDPI.

Sanderson, Conrad, David Douglas, Qinghua Lu, Emma Schleiger, Jon Whittle, Justine Lacey, Glenn Newnham, Stefan Hajkowicz, Cathy Robinson, and David Hansen. (2023). 'AI ethics principles in practice: Perspectives of designers and developers', *IEEE Transactions on Technology and Society*.

Schwartz, Shalom H. (2012). ‘An overview of the Schwartz theory of basic values’, *Online readings in Psychology and Culture*, 2: 2307-0919.1116.

Seymour, William, and Max Van Kleek. (2020). “Does Siri have a soul? Exploring voice assistants through shinto design fictions.” In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 1-12.

Shams, Rifat Ara, Mojtaba Shahin, Gillian Oliver, Jon Whittle, Waqar Hussain, Harsha Perera, and Arif Nurwidyantoro. (2021). ‘Human values in mobile app development: An empirical study on bangladeshi agriculture mobile apps’, *arXiv preprint arXiv:2110.05150.*

Shams, Rifat, Muneera Bano, Didar Zowghi, Qinghua Lu, and Jon Whittle. (2023). “Exploring Human Values in AI Systems: Empirical Analysis of Amazon Alexa.” In *Empirical Requirements Engineering Workshop (EmpiRE’23) at International Requirements Engineering Conference RE’23.* Hanover, Germany: IEEE.

Tafferner, Zoltán, Balázs Illés, Olivér Krammer, and Attila Géczy. (2023). ‘Can ChatGPT Help in Electronics Research and Development? A Case Study with Applied Sensors’, *Sensors*, 23: 4879.

Thoppilan, Romal, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, and Yu Du. 2022. ‘Lamda: Language models for dialog applications’, *arXiv preprint arXiv:2201.08239.*

Tlili, Ahmed, Boulus Shehata, Michael Agyemang Adarkwah, Aras Bozkurt, Daniel T Hickey, Ronghuai Huang, and Brighter Agyemang. (2023). ‘What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education’, *Smart Learning Environments*, 10: 15.

Van Dis, Eva AM, Johan Bollen, Willem Zuidema, Robert van Rooij, and Claudi L Bockting. (2023). ‘ChatGPT: five priorities for research’, *Nature*, 614: 224-26.

Wardat, Yousef, Mohammad A Tashtoush, Rommel AlAli, and Adeeb M Jarrah. (2023). ‘ChatGPT: A revolutionary tool for teaching and learning mathematics’, *Eurasia Journal of Mathematics, Science and Technology Education*, 19: em2286.

Weber, Max. (1949). ‘” Objectivity” in social science and social policy’, *The methodology of the social sciences*: 49-112.

White, Jules, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C Schmidt. (2023). ‘A prompt pattern catalog to enhance prompt engineering with chatgpt’, *arXiv preprint arXiv:2302.11382.*

Yan, Da. (2023). ‘Impact of ChatGPT on learners in a L2 writing practicum: An exploratory investigation’, *Education and Information Technologies*: 1-25.

Zowghi, Didar, and Francesca da Rimini. (2023). ‘Diversity and Inclusion in Artificial Intelligence’, *arXiv preprint arXiv:2305.12728.*