The Qualitative Report

Volume 28 | Number 12 Article 5

12-2-2023

Guidelines for the Integration of Large Language Models in Developing and Refining Interview Protocols

Jessica L. Parker EdD
Massachusetts College of Pharmacy and Allied Health Sciences, jessica.parker@mcphs.edu

Veronica M. Richard
Dissertation by Design, veronica@dissertationbydesign.com

Kimberly Becker
Iowa State University, kimberly@academicinsightlab.org

Recommended APA Citation

Parker, J. L., Richard, V. M., & Becker, K. (2023). Guidelines for the Integration of Large Language Models in Developing and Refining Interview Protocols. *The Qualitative Report*, *28*(12), 3460-3474. https://doi.org/10.46743/2160-3715/2023.6801

Guidelines for the Integration of Large Language Models in Developing and Refining Interview Protocols

Abstract

Rapid advancements in generative artificial intelligence (AI), specifically large language models (LLMs), offer unprecedented opportunities and challenges for qualitative researchers. This paper presents comprehensive guidelines for the ethical and effective use of LLMs in the development and refinement of interview protocols. Through a multidisciplinary lens, this paper explores potential pitfalls, ethical considerations, and best practices to ensure the responsible integration of LLMs in the research process. The guidelines proposed serve not only as a methodological roadmap for researchers but also as a catalyst for dialogue on the ethical dimensions of LLMs in qualitative research. Furthermore, the authors describe and share a web-based application developed to guide users through the stages of the protocol. Ultimately, the paper calls for a collective, informed approach to harness the capabilities of LLMs while upholding the integrity and ethical standards of scholarly research.

Keywords

large language models, ChatGPT, qualitative research, interview protocol refinement framework, interview protocol, generative artificial intelligence

Acknowledgements

In the development of this manuscript, we employed GPT-4, a generative artificial intelligence model, as a collaborative tool. GPT-4 was instrumental in various stages of the research and writing process, including brainstorming, table construction, and the creation of example scenarios.

The Qualitative Report 2023 Volume 28, Number 12, 3460-3474
https://doi.org/10.46743/2160-3715/2023.6801

Guidelines for the Integration of Large Language Models in Developing and Refining Interview Protocols

Jessica L. Parker[1], Veronica M. Richard[2], and Kimberly Becker[3]
[1]School of Healthcare Business, Massachusetts College of Pharmacy and Health Sciences, Boston, Massachusetts, USA
[2]Dissertation by Design, Raleigh, North Carolina, USA
[3]Academic Insight Lab, Raleigh, North Carolina, USA

Rapid advancements in generative artificial intelligence (AI), specifically large language models (LLMs), offer unprecedented opportunities and challenges for qualitative researchers. This paper presents comprehensive guidelines for the ethical and effective use of LLMs in the development and refinement of interview protocols. Through a multidisciplinary lens, this paper explores potential pitfalls, ethical considerations, and best practices to ensure the responsible integration of LLMs in the research process. The guidelines proposed serve not only as a methodological roadmap for researchers but also as a catalyst for dialogue on the ethical dimensions of LLMs in qualitative research. Furthermore, the authors describe and share a web-based application developed to guide users through the stages of the protocol. Ultimately, the paper calls for a collective, informed approach to harness the capabilities of LLMs while upholding the integrity and ethical standards of scholarly research.

Keywords: large language models, ChatGPT, qualitative research, interview protocol refinement framework, interview protocol, generative artificial intelligence

Introduction

The Castillo-Montoya (2016) Interview Protocol Refinement (IPR) framework has been influential in qualitative research, offering a four-phase approach to align interview questions with research objectives, create inquiry-based conversations, gather feedback, and pilot the protocol. This framework enhances the reliability of the interview schedule, allows for adjustments by identifying flaws through piloting, and mitigates researcher bias through reflexivity and supervisor or expert consultation. It also addresses qualitative research considerations like interviewer-interviewee trust, interview logistics, and question sequencing. By emphasizing the alignment of interview and research questions, the framework ensures data relevance and has been found valuable in enhancing the credibility of qualitative findings (see Khan et al., 2021; Ramonienė, 2023).

Using large LLMs such as ChatGPT in conjunction with the IPR framework (Castillo-Montoya, 2016) can augment novice researchers' critical thinking and reflection in interview protocol development and refinement. The integration of these advanced technological tools has the potential to streamline the alignment of interview and research questions but also offer a structured guide for crafting research protocols, thereby enhancing the methodological rigor of qualitative studies (Parker et al., 2023).

Furthermore, machine learning techniques can be employed to provide real-time, intelligent evaluation of these protocols, grounding the feedback in the principles of the IPR

framework. This amalgamation of the IPR framework and generative AI technologies also addresses pedagogical challenges like disciplinary heterogeneity by offering individualized, context-sensitive guidance. However, incorporating LLMs into academic research is not without its challenges and considerations, necessitating rigorous scrutiny for ethical and methodological soundness.

LLMs inherently lack the nuanced understanding and contextual awareness often crucial in qualitative research. In contrast, human researchers bring a depth of understanding across cultural, social, and ethical dimensions – nuances that generative AI technologies cannot fully replicate. Issues such as data privacy, informed consent, and responsible use of information are not merely technical considerations; they demand human discernment and a commitment to ethical responsibility (Lahman, 2018).

This paper aims to provide a systematic guide for qualitative researchers utilizing LLMs like ChatGPT in conjunction with Castillo-Montoya's (2016) IPR framework to develop and refine interview protocols. These guidelines build on the insights gained from our previous work (see Parker et al., 2023) and are intended to be a practical and ethical resource for qualitative researchers. An overview of the guidelines that will be discussed in this paper is displayed in Table 1.

Table 1
Overview of guidelines for researchers using LLMs to develop and refine interview protocols

Consideration	Explanation	Example
Ethical Considerations	The use of LLM tools like ChatGPT raises new ethical issues such as attribution practices and data privacy.	Ensuring no personal data is stored by the AI
Cultural Sensitivity[1]	ChatGPT's exposure to diverse texts can support testing questions across cultural contexts, ensuring respect and consideration.	Testing interview prompts and questions for multiple countries and cultural nuances
Quality of Output	Evaluation of ChatGPT's output, assessing reliability and validity; comparison with traditional methods.	Augmenting human-made protocols
LLMs as a Tool, Not Replacement	LLMs complement researchers, not replace them; human understanding of context remains vital.	Using AI to draft but humans finalize the output
AI Literacy	Understanding the basics of AI and LLM technology is essential for responsible use and meaningful interpretation of output.	Attending training sessions on AI basics, evolutions in the technology, limitations, and ethical considerations.

[1] Exposure to diverse texts is both boon and bane as that same exposure potentiates bias inherited from the training data (Ray, 2023).

Future Research Directions	Research is needed to identify best practices using LLMs in research planning and processes.	Exploring how AI can be used in qualitative data analysis

Ethical Considerations

The ethical landscape of generative AI in academia is complex, touching on issues of transparency, data privacy, and intellectual property, among others. The aim of this section is to provide a detailed overview of the ethical considerations that researchers must navigate when employing generative AI technologies in their work.

Before employing human-augmenting tools such as LLMs for academic research, graduate students should consult with their academic chairs or supervisors. This step is not only a matter of procedural integrity but also one of ethical transparency to ensure that the rationale behind the use of such tools is clearly articulated and approved.

Statement of Generative AI Use

Guidelines, such as those from the Association for Computing Machinery (ACM), emphasize the importance of transparency in AI systems (Association for Computing Machinery, 2018). Just as AI systems must be transparent, likewise, researchers who use AI tools should also be forthcoming about how they use such tools. There are two primary ways to ensure such transparency: (1) citation, and (2) acknowledgment. Following proper attribution practices helps maintain academic integrity and gives credit where it is due.

Citation. Although guidelines may evolve as this new technology becomes more ubiquitous, the American Psychological Association (APA) advises that the citation of LLMs should mirror the approach taken for software tools. The rationale behind this recommendation is that, despite their labels as "chatbots," the use of LLM-generated language is *not* a communication. Conversations with these models are not retrievable by other users, and the communique is not with another human. Essentially, using text generated during a chat session showcases the capabilities of the algorithm. Therefore, it is important to give credit to the creator of the algorithm. Below is guidance from the APA on how to cite ChatGPT (McAdoo, 2023).

- Reference List: OpenAI. (2023). ChatGPT-4 (Aug 29 version) [Large language model].
- Parenthetical citation: (OpenAI, 2023)
- Narrative citation: OpenAI (2023)

Acknowledgment. When employing LLMs in the process of manuscript writing or revision, which goes beyond the instances highlighted in the previous section (e.g., seeking suggestions on wording/phrasing, refining the conciseness of content, or other interactions reminiscent of collaborating with a writing assistant or editor), it becomes more apt to acknowledge the AI rather than to cite it. Suitable places for such acknowledgment include the cover letter accompanying a manuscript submission to a journal editor, within the acknowledgments section of the manuscript, or directly within the specific section where the AI tool played a role. In addition, when introducing the use of such tools to supervisors, committees, and/or institutional review boards, it is imperative to offer a detailed statement regarding the AI's deployment. This statement ought to outline the following:

- Title of the exact AI tool used, including the model number

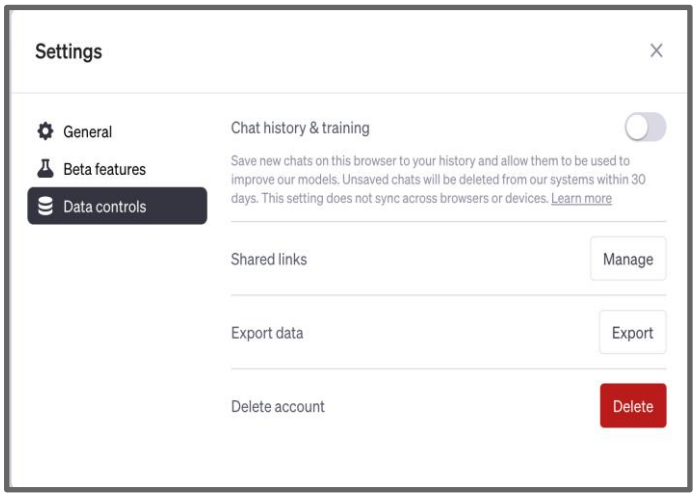
- Protocols established for maintaining confidentiality
- Specific purpose behind utilizing the AI
- Steps undertaken to ensure ethical compliance

Data Privacy and Intellectual Property

Safeguarding intellectual property is a paramount ethical concern when engaging with LLMs. When using ChatGPT, researchers are advised to deactivate its chat history and training (Figure 1). By doing so, researchers ensure no personally identifiable information or intellectual property is inadvertently stored or used in its training model. Disabling these functions serves as a precautionary measure to protect the proprietary nature of the research questions, methodologies, and preliminary findings, thereby aligning with the broader academic ethos of maintaining the integrity of scholarly work.

Researchers such as Hall et al. (2014) and Foley (2015) delve into the complexities of intellectual property rights in academic settings, emphasizing the need for stringent protective measures; however, in the emergent environment and application of generative AI tools in contemporary research environments, the discourse on this topic is expected to expand and evolve.

Figure 1
Screenshot of privacy toggle in ChatGPT's user settings



Cultural Sensitivity

Drawing upon the culturally responsive, relational, and reflexive ethics (CRRRE) as delineated by Lahman (2018) and Lahman et al. (2011) and denoted as the three Rs, we argue that LLMs can be leveraged to enhance the cultural sensitivity of interview protocols. LLMs are trained on vast datasets that include text from a wide range of contexts, cultures, and languages. This extensive training allows the LLMs to learn about various cultural norms, values, and linguistic nuances. Further, the natural language processing capabilities of LLMs enable them to understand, generate, and process human language in a way that can be tailored to cultural differences. Thus, by fostering a responsive approach, LLMs can aid in the real-time adaptation of interview protocols to align with the cultural norms and expectations of diverse participant groups.

Responsiveness to Cultural Contexts

Castillo-Montoya (2016) emphasizes the importance of starting interviews with non-threatening questions that familiarize participants with describing experiences. For instance, asking about the neighborhood where a participant grew up can set the tone for a conversation and provide insights into their sociopolitical consciousness. LLMs can be programmed to generate such questions, considering cultural nuances and social structures, making the interview more responsive to diverse cultural contexts.

> *Example scenario:* Consider a research project aimed at investigating the perception and experience of neighborhood safety among residents from different cultural backgrounds within a cosmopolitan city. A set of introductory questions tailored to different cultural contexts could be generated. For participants from a collectivist culture, where community relationships and collective actions are highly valued, an introductory question might be: "Can you describe a community gathering or event in your neighborhood that made you feel a sense of belonging and safety?" Conversely, for participants from individualistic cultures, where personal autonomy and privacy are often emphasized with a question like: "Can you recall an instance where you or your immediate neighbors took steps to enhance safety in your neighborhood?"
>
> *Example Prompt:* Generate introductory questions for a research study investigating neighborhood safety among residents from different cultural backgrounds within a cosmopolitan city. The study aims to understand how cultural factors influence individuals' sense of security and community engagement. For collectivist cultures, emphasize community relationships and collective actions. For individualistic cultures, emphasize personal autonomy and individual actions. Ensure the questions are open-ended, non-threatening, and designed to make participants feel comfortable sharing their experiences.

While LLMs offer the capability to generate culturally nuanced questions, we argue that researchers must be acutely aware of their own limitations in fully grasping the diverse perspectives of various cultures (Lahman, 2018). This recognition calls for (a) keeping an aspect of flexibility in interview protocols to allow for learning in the field, and (b) acknowledging our expectations and assumptions may or may not align with participants' cultural understandings (Lahman, 2018; Rogoff, 2003). To this point, Brinkmann and Kvale (2015) highlight the difficulties of identifying subcultural differences and understandings.

Relationality and Language

Generally speaking, relationality refers to a researcher's intent to exhibit genuine care and respect toward participants (Ellis, 2007; Lahman, 2018), which includes the interview protocol and language used in interview questions and prompts. The language employed in interview questions and prompts should be clear, comprehensible, and devoid of specialized jargon or academic vernacular (Brinkmann & Kvale, 2015). In this context, LLMs can be instrumental, aiding in adapting questions to be more relatable and easily grasped across varied cultural contexts. This focus on relational language dovetails with the CRRRE's emphasis on fostering rapport, care, and respect during interviews (Lahman, 2018; Lahman et al., 2011).

As with the responsiveness to cultural contexts, it is the researcher's responsibility to be critical of not only the output of LLMs, such as ChatGPT, but also of the prompts used to

generate output. Being critical of how we care for and respect our participants, in this case through our language in research tools, should be of the utmost importance.

Reflexivity in Protocol Design

Reflexivity in protocol design can be conceptualized as an ongoing exercise of introspection (Lahman, 2018) throughout the states of protocol creation and iteration. While qualitative researchers aim to uncover nuanced insights about participants and their lived realities (Jacob & Furgeson, 2012), there is equal merit in reflecting on our methodological decisions and biases as investigators.

Here, LLMs offer a complementary tool, assisting researchers in recognizing cognitive biases and identifying gaps or overrepresentations in the interview protocol. Thinking in a reflexive manner, using LLMs provides opportunities for thoughtful and self-questioning processes in which researchers "interrogate" how the interview protocol conveys what is valued, what is honored, and what is left out (Braun & Clarke, 2022). In this process, researchers gain awareness and are positioned to discover new possibilities (Braun & Clarke, 2022) for revising protocols that serve to enrich the data collected.

However, the advent of generative AI introduces an additional layer of complexity to reflexive research. It is incumbent upon the researcher to scrutinize interview questions and prompts generated or refined by an LLM. A concrete step here would be to perform a "bias audit" on the protocol items, examining them for both overt and subtle biases that may be inherent in the AI's training model. For instance, the researcher could consult with domain experts or ethicists to validate the fairness and inclusivity of the items.

> *Example Scenario:* A researcher seeks to capture the daily struggles, coping mechanisms, and support systems of single parents in urban settings. As a married individual without children, the researcher recognizes potential biases in understanding the intricacies of single parenting in urban environments. The researcher decides to use an LLM to develop interview questions and explore potential biases or assumptions present in them.
>
> *Example Prompt*: You are a qualitative researcher who is interested in understanding the challenges faced by single parents in urban settings. Generate a set of interview questions that capture the challenges, coping strategies, and support systems of these parents. Questions should be unbiased, open-ended, and sensitive to the diverse realities of single parents in urban settings. After generating these questions, highlight potential biases or assumptions present in them, and suggest alternative phrasing.

The inclusion of LLMs in protocol design underscores the potential of the technology to aid researchers in responsively navigating diverse cultural contexts, cognitive biases, and assumptions present in interview protocols. This not only enriches the quality of the data collected but also maximizes ethical and respectful engagement with participants across diverse cultural backgrounds. However, the real strength of qualitative research lies in its human core – our capacity for empathy, understanding, nuanced interpretation, and critical reflection.

While LLMs can assist, guide, and augment our processes, the final responsibility of crafting authentic and meaningful research remains firmly in the hands of the human researcher. With the output provided by the LLM, it is the researcher's responsibility to consider the assumptions and biases highlighted. In gaining awareness and thinking critically

about these points, the researcher can then make key, thoughtful decisions about language and wording.

Quality of Output

While LLMs offer the advantage of speed and scalability in generating text, the quality of their output needs to be evaluated and benchmarked. As Brinkmann (2018) cautioned, efficiency is not everything. These generative AI models, despite their sophistication, are not infallible. For instance, ChatGPT has been shown to "hallucinate" or confidently provide false or misleading information. Moreover, the model's limitations extend to potential biases in its training data, which can inadvertently perpetuate stereotypes or misinformation. Thus, human oversight is critical when using these models to generate and refine interview protocols.

Traditional human-generated interview protocols often benefit from multiple rounds of expert review and validation. In contrast, the emergent nature of LLMs in qualitative research is uncharted territory, and established methods for evaluating their output quality are still largely unknown. This presents a landscape of both uncertainty and potential for researchers. This section presents considerations for qualitative researchers evaluating the quality of an LLM's output while developing and refining interview protocols.

The identified criteria for evaluating output quality (Table 2) draws upon the IPR framework (Castillo-Montoya, 2016), embodying elements essential for rigorous qualitative inquiry.

Table 2
Criteria for evaluating output quality

Criteria	Sub-Criteria	Description
Conceptual Depth	Alignment with Research Objectives	Generated content must closely align with the research objectives.
	Interpretive Depth	Questions and prompts should reflect a nuanced understanding of the subject matter to facilitate depth in the collected data.
	Elicit Storytelling	Questions and prompts should be open-ended, encouraging interviewees to explain, describe, and reflect.
	Sensitivity to Context	Questions and prompts generated should be sensitive to socio-cultural and situational contexts, avoiding generic or overly broad queries that yield limited insights.

Reliability or Dependability	Consistency	Multiple pilots with identical or similar queries should produce consistent topical outputs, affirming the LLM's reliability.
	Replicability	A comprehensive log of prompts and corresponding outputs should be maintained and made accessible, facilitating the replication of the study or processes by other researchers to maximize dependability.
Validity or Credibility	Face Validity	An initial heuristic evaluation should confirm that the generated output appears valid; that is, it directly contributes to the research objectives.
	Content Validity	Generated items should comprehensively cover the subject matter to prevent gaps in the research findings.
Language and Semantics	Clarity	Output should be straightforward, avoiding ambiguity and jargon that can confuse respondents.
	Precision	Terms and phrases should be used consistently and directly pertain to the subject matter.
	Complexity	While simple language is often preferable, some research contexts may require complex queries. In such cases, complexity should not compromise clarity or precision.
Factual Integrity or Credibility	Verification	Generated content should be verified against trusted sources or expert opinions to assess its factual integrity or credibility.
	Hallucination	LLMs can produce outputs that are well-written but may not be accurate or appropriate, known as “hallucinations.” Output should be closely reviewed and monitored to identify potential hallucinations.

When considering the quality of the output, it is critical to also note the role of *input quality*. Given the principle of "garbage-in, garbage-out," researchers must be meticulous in crafting the queries or prompts fed into the LLM. Inaccurate or vague input can invariably lead to unreliable and invalid output, compromising the research objectives. For instance, consider a scenario where a researcher uses ChatGPT to develop an interview protocol for a study with poorly defined research questions or objectives. The output generated by the chatbot will likely

produce generalized, off-topic interview questions and prompts irrelevant to the research topic. Therefore, we recommend keeping a detailed log of prompts (i.e., inputs) and related outputs to increase one's credibility and reflexivity in using LLMs in interview protocol creation and refinement. This log serves as a critical aspect of the audit trail, affording researchers the opportunity to share their detailed processes, clarify the relationship between input and output, and identify potential biases and assumptions. Given the lack of standardized guidelines for using LLMs in academic research, maintaining such a detailed log not only helps ensure the quality and reliability of the research but also sets a precedent for the responsible and ethical use of generative AI in scholarly endeavors. Expert reviews can provide an additional layer of validation or credibility, supplementing the output quality criteria outlined above.

LLMs as a Tool, Not a Replacement

The key benefit of generative AI tools is that they are designed to complement human researchers rather than to serve as substitutes. A hybrid human-AI approach is an optimal way to uphold methodological rigor, meticulous adherence to established research protocols and standards, and maximize ethical integrity in the research process (Porsdam Mann et al., 2023).

Advantages and Structured Frameworks

LLMs offer many advantages, including but not limited to rapid text generation, efficient data sorting, and preliminary analysis capabilities (AlZaabi et al., 2023). Their prowess in managing large datasets often surpasses the efficiency of manual methods. Furthermore, when LLMs are integrated into structured research frameworks, such as Castillo-Montoya's (2016) IPR framework, they can optimize specific research activities, thereby enhancing the overall quality and credibility of the study.

Limitations and the Need for Human Oversight

Despite these considerable advantages, LLMs are not without limitations. They often lack the capacity for contextual comprehension and nuanced understanding, which are quintessential qualitative research elements (AlZaabi et al., 2023; Dignum, 2018). Additionally, LLMs are devoid of ethical judgment, a fundamental cornerstone in academic research (Porsdam Mann et al., 2023). Consequently, human oversight, especially when guided by established frameworks and ethical guidelines, becomes indispensable for mitigating these limitations.

Symbiotic Relationship and Collaborative Approach

The optimal utilization of LLMs is realized when they are employed in a symbiotic relationship with human researchers. While LLMs can significantly assist in the drafting or coding phases of research, human expertise remains irreplaceable for tasks that require nuanced revision, interpretation, critical analysis, and ethical evaluations. This collaborative approach preserves and enhances academic rigor and ethical integrity, which are foundational pillars in qualitative research.

Protocol Development and Human Finalization

In the specific context of developing interview protocols, LLMs can be invaluable. However, finalizing the protocol rests squarely on the human researcher. This maximizes the

chances that the questions and prompts generated are ethically sound, methodologically rigorous, and contextually appropriate. As researchers have noted, the final test is in the actual interview and analysis of the data. If the questions or prompts do not elicit the information needed to answer the research question(s), it is the researcher's responsibility to change and adapt.

Complementarity and Synergy

The complementarity between LLMs and human researchers enriches the research process by merging computational efficiency with human insight and reflexivity. This synergy is instrumental in sustaining the academic and ethical standards vital to research. It also aids researchers in navigating the intricate ethical and methodological complexities that inevitably arise when integrating advanced computational tools into qualitative research paradigms.

AI Literacy

AI literacy is a prerequisite for the ethical and effective use of LLMs in qualitative research. Otherwise, researchers risk misusing the technology, misinterpreting its outputs, and compromising ethical standards.

Foundational Knowledge

It is crucial for researchers to grasp the basic tenets of generative AI and LLMs, including how these models are trained, their capabilities, and their limitations. Understanding the architecture, for instance, can help researchers understand why an LLM might generate a specific type of output or why it might be biased. Table 3 outlines several aspects of foundational knowledge that researchers should consider familiarizing themselves with prior to integrating LLMs into their research processes. It should be noted that that table is by no means exhaustive but serves as a preliminary guide for inquiry.

Table 3
Foundational concepts to explore before using LLMs in research

Model Architecture	Transformers	LLMs like ChatGPT are based on transformer architectures. Transformers are particularly strong at handling elements in sequence (e.g., interview questions are inherently sequential).
	Neural networks	A neural network is the foundational structure of an LLM. Neural networks learn from the data they are trained on, cannot generate or analyze information beyond their training data, and can produce misleading outputs.
Training and Data	Supervised learning	LLMs are generally trained in a supervised manner, where the model learns to predict the next word in a sequence based on the sentence patterns it has seen so far.

	Data sets	Quality and diversity of data sets are crucial. The model can only generate outputs based on the data it was trained on.
Capabilities	Natural language understanding and generation	LLMs can understand and generate human-like text, making them useful for tasks such as content creation and text summarization.
	Task agnostic	LLMs can be fine-tuned for specific tasks but are generally task-agnostic, meaning they can perform a variety of natural language tasks without task-specific training.
Limitations	Lack of understanding	While LLMs can generate text based on patterns, they do not understand the text in a way that humans do.
	Biases	LLMs can inherit the biases present in their training data, potentially leading to biased outputs.
	Hallucinations	LLMs are known to "hallucinate" or confidently assert incorrect facts.
Evaluation metrics	Accuracy	A measure of how well the model's outputs align with human-labeled ground truth during testing.
	Quality	This can be subjective but involves assessing the readability, relevance, and factual accuracy of the generated text.

Future Research Directions

The use of LLMs to develop and refine interview protocols heralds a promising future for qualitative research. However, the journey toward seamless integration of these generative AI models within qualitative research beckons a more in-depth exploration, rigorous evaluation, and thoughtful consideration of cultural, responsible, relational, reflexive, ethical, and methodological implications.

Evaluative Frameworks for LLM Prompting and Output

Future research is needed to develop a robust evaluative framework to assess human-AI-generated interview protocols' quality, reliability, and validity. Detailed methodological studies are needed to examine the culturally relevant, responsive, reflexive, and ethical efficacy and quality of LLM-generated and human-directed protocols. What prompt and follow-up prompts produce appropriate protocols for different research participants and contexts? What aspects of the protocol development and piloting process warrant more human-generated revisions? Overall, gaining clarity on LLM prompting and output with guidelines would enable qualitative researchers to use such tools to increase their trustworthiness and reflexive practices. Establishing evaluative frameworks for LLM prompting and output can enhance trustworthiness and encourage reflexive practices. These frameworks feature specific prompts

designed to guide critical, process-oriented thinking. Moreover, they aim to offer reflection prompts that specifically address assumptions and biases.

Mitigating Potential Biases

Future research is needed to investigate potential biases in LLM-generated content and develop strategies to mitigate or rectify such biases. Such research could involve developing algorithms or methodologies to identify, mitigate, or rectify such biases.

Comparative studies that assess the potential for bias in LLM-generated questions against human-generated questions could offer valuable insights.

Empirical Validation and Comparative Studies

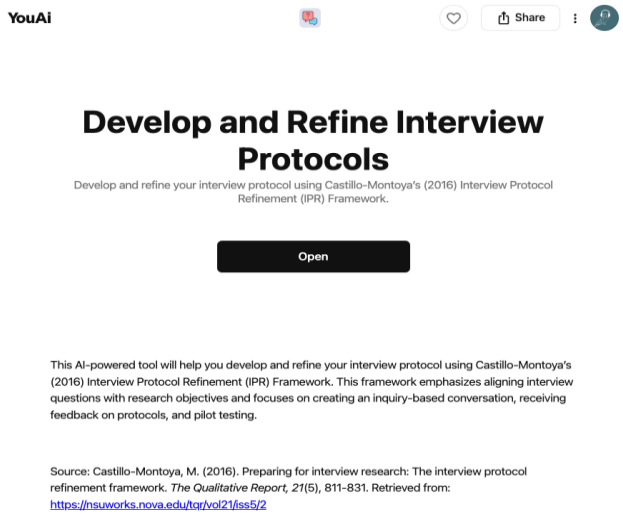
Future work could focus on empirically validating the effectiveness of integrating LLMs into the research process. Comparative studies may offer insights into the quality of data collected using traditional methods versus those incorporating LLMs.

Future Trends

Technological advancements are creating unprecedented opportunities for qualitative research. One such advancement is web applications that are programmed to do the prompt engineering (i.e., the kind of prompts described in this article) for the user. To scaffold scholars' use of an LLM to develop and refine interview protocols, we developed such an app, which guides users through the stages of the protocol. The application was built to interact with ChatGPT-4 via a platform called YouAI (https://youai.ai/), which hosts context-aware AI applications. This application (https://academicinsightlab.org/develop-and-refine-interview-protocols) is programmed with the prompt engineering shown in Figure 2.

Figure 2

Screenshot of the AI application for developing and refining interview protocols



We present this app with the caveat that it is a new and emerging technology that has grown exponentially yet is still nascent. Users are encouraged to apply the guidelines described herein as they explore its relevance to their research.

Conclusion

As generative AI and LLMs continue to evolve, their integration into qualitative research methods will undoubtedly become more sophisticated. As we delineated in this paper, this technology is not without inherent ethical and methodological challenges. By extending Castillo-Montoya's (2016) framework, we have provided a structured guide for scholars to navigate the intricate relationship between generative AI and interview protocol design. This endeavor, however, goes beyond mere methodological recommendations. It beckons a broader discourse within the academic community on the ethical ramifications of melding AI with qualitative methodologies. It is our fervent hope that this article serves as an impetus for a collective, informed, and conscientious approach – ensuring that while we maximize AI's potential, we remain unwavering in our commitment to the core tenets of ethical and impactful scholarly research.

References

Association for Computing Machinery. (2018). *ACM Code of Ethics and Professional Conduct.* https://www.acm.org/code-of-ethics

AlZaabi, A., ALAmri, A., Albalushi, H., Aljabri, R., & AAlAbdulsalam, A. (2023). ChatGPT applications in academic research: A review of benefits, concerns, and recommendations. *bioRxiv.* https://doi.org/10.1101/2023.08.17.553688

Braun, V., & Clarke, V. (2022). *Thematic analysis: A practical guide*. SAGE.

Brinkmann, S. (2018). The interview. In N. K. Denzin, & Y. S. Lincoln (Eds.), *The SAGE handbook of qualitative research* (5th ed., pp. 576-599). SAGE.

Brinkmann, S., & Kvale, S. (2015) *Interviews: Learning the craft of qualitative research interviewing* (3rd ed.). SAGE.

Castillo-Montoya, M. (2016). Preparing for interview research: The interview protocol refinement framework. *The Qualitative Report*, *21*(5), 811-831. https://doi.org/10.46743/2160-3715/2016.2337

Dignum, V. (2018). Ethics in artificial intelligence: Introduction to the special issue. *Ethics and Information Technology*, *20*(1), 1-3. https://doi.org/10.1007/s10676-018-9450-z

Ellis, C. (2007). Telling secrets, revealing lives. *Qualitative Inquiry*, *13*(1), 3-29. https://doi.org/10.1177/1077800406294947

Foley, H. C. (2015). A new approach to intellectual property management and industrially funded research at Penn State. *Research-Technology Management*, *55*(5), 12-17. https://doi.org/10.5437/08956308X5505008

Hall, J., Matos, S., Bacher, V., & Downey, R. (2014). Commercializing university research in diverse settings: Standardized intellectual property management. *Research-Technology Management, 57*(5), 26-34. https://doi.org/10.5437/08956308X5705250

Jacob, S. A., & Furgerson, S. P. (2012). Writing interview protocols and conducting interviews: Tips for students new to the field of qualitative research. *The Qualitative Report*, *17*(42), 1-10. https://doi.org/10.46743/2160-3715/2012.1718

Khan, N., Wei, H., Yue, G., Nazir, N., & Zainol, N. (2021). Exploring themes of sustainable practices in manufacturing industry: using thematic networks approach. *Sustainability*, *13*(18), 10288. https://doi.org/10.3390/su131810288

Lahman, M. K. E. (2018). *Ethics in social science research: Becoming culturally responsive*. SAGE.

Lahman, M. K. E., Geist, M., Rodriguez, K., Graglia, P., & Deroch, K. (2011). Culturally responsive relational reflexive ethics in research: The three R's of ethics. *Quality and Quantity: International Journal of Methodology*, *45*(6), 1397-1414,

https://doi.org/10.1177/1077800410392330

McAdoo, T. (2023, April 7). How to cite ChatGPT. *APA Style Blog.* https://apastyle.apa.org/blog/how-to-cite-chatgpt

Parker, J. L., Richard, V. M., & Becker, K. (2023). Flexibility & iteration: Exploring the potential of large language models in developing and refining interview protocols. *The Qualitative Report, 8*(9), 2772-2791. https://doi.org/10.46743/2160-3715/2023.6695

Porsdam Mann, S., Earp, B., Møller, N., Vynn, S., & Savulescu, J. (2023). AUTOGEN: A personalized large language model for academic enhancement-ethics and proof of principle. *The American Journal of Bioethics*, *23*(7), 24-36. https://doi.org/10.1080/15265161.2023.2233356

Ramonienė, L. (2023). Sustainability motives, values and communication of slow fashion business owners. *Journal of Philanthropy and Marketing*, *28*(2), e1788. https://doi.org/10.1002/nvsm.1788

Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*, *3*(1), 121-154. https://doi.org/10.1016/j.iotcps.2023.04.003

Rogoff, B. (2003). *The cultural nature of human development*. Oxford University Press.

Author Note

Dr. Jessica Parker is a researcher and educator who is passionate about demystifying the research and writing process for scholars. She is the founder and CEO of Dissertation by Design and the co-founder of Academic Insight Lab. Her research interests are at the intersection of technology and education; she is particularly intrigued by the potential of generative AI for academic purposes, exploring how this technology can revolutionize the way we conduct research, teach, and learn. Jessica has worked with a diverse range of researchers and scholars and continues to teach doctoral students of Health Sciences at Massachusetts College of Pharmacy and Health Sciences (MCPHS) University. Please direct correspondence to jessica.parker@mcphs.edu.

Dr. Veronica Richard is a qualitative methodologist at Dissertation by Design. She has held various academic positions, including adjunct, assistant, and associate professor roles at the University of Northern Colorado, Indiana University Northwest, and Concordia University Chicago. Her experience spans working with undergraduate, master's, and doctoral students, primarily in the fields of literacy and research methods. Throughout her career, Veronica has been an integral part of nine research teams, many of which were grant-funded and dedicated to supporting at-risk youth. Veronica earned her Ph.D. from the University of Northern Colorado in 2010, specializing in Applied Statistical Research and Research Methods with a cognate in Reading. Please direct correspondence to veronica@dissertationbydesign.com.

Dr. Kimberly Becker is an applied linguist who specializes in disciplinary academic writing and English for research publication purposes. She is the co-founder of Academic Insight Lab and holds a Ph.D. in applied linguistics and technology (Iowa State University, 2022) and an M.A. in teaching English as a second language (Northern Arizona University, 2004). Kimberly's research and teaching experience as a professor and communication consultant has equipped her to support native and non-native English speakers in written, oral, visual, and electronic communication. Her most recent publications are related to the use of ethical AI for automated writing evaluation and a co-authored e-book, *Preparing to Publish*, about composing academic research manuscripts. Please direct correspondence to kimberly@academicinsightlab.org.

Acknowledgements: In the development of this manuscript, we employed GPT-4, a generative artificial intelligence model, as a collaborative tool. GPT-4 was instrumental in various stages of the research and writing process, including brainstorming, table construction, and the creation of example scenarios.

Article Citation

Parker, J. L., Richard, V. M., & Becker, K. (2023). Guidelines for the integration of large language models in developing and refining interview protocols. *The Qualitative Report*, *28*(12), 3460-3474. https://doi.org/10.46743/2160-3715/2023.6801