INFOTECH
Journal of Applied Sciences, Informatics, Computing and Knowledge Society

Volume 4, No. 1, January/June, 2023

Qualitative Research Techniques for Language Models: Conducting Semi-Structured Conversations with ChatGPT and BARD in Computer Science Education

Thomas Mason[1]

[1] University of Queensland – Australia, thomasmason40@outlook.es, https://orcid.org/0009-0001-2108-8406

ABSTRACT

In the era of artificial intelligence, large language models like ChatGPT and BARD find diverse applications, from language translation to text generation. This study explores a novel approach to qualitative research by interviewing these models, which house vast amounts of data and perspectives. We investigate the applicability of qualitative content analysis to interviews with ChatGPT (English and German) and BARD, focusing on the relevance of computer science in K-12 education. Our findings reveal that model responses heavily depend on context, with the same model yielding varying results for identical questions. From this, we derive guidelines for conducting and analyzing interviews with large language models. While qualitative content analysis methods can be applied, careful consideration of contextual factors is crucial. The guidelines provided support researchers and practitioners in conducting nuanced interviews. Overall, we advise against relying on interviews with large language models for research due to their unpredictable nature. Instead, we propose using these models as exploration tools to gain diverse perspectives on research topics and validate interview guidelines before real-world applications.

Keywords: *Large Language Models; Qualitative Content Analysis; Contextual Factors*

Received:
03 March, 2023

Accepted for publication:
18 April, 2023

INFOTECH
Journal of Applied Sciences, Informatics, Computing and Knowledge Society

Volume 4, No. 1, January/June, 2023

Técnicas de Investigación Cualitativa para Modelos de Lenguaje: Realización de Conversaciones Semiestructuradas con ChatGPT y BARD en la Educación en Ciencias de la Computación

RESUMEN

En la era de la inteligencia artificial, grandes modelos de lenguaje como ChatGPT y BARD encuentran diversas aplicaciones, desde la traducción de idiomas hasta la generación de texto. Este estudio explora un enfoque novedoso para la investigación cualitativa mediante entrevistas con estos modelos, que albergan grandes cantidades de datos y perspectivas. Investigamos la aplicabilidad del análisis de contenido cualitativo a entrevistas con ChatGPT (inglés y alemán) y BARD, centrándonos en la relevancia de la informática en la educación K-12. Nuestros hallazgos revelan que las respuestas de los modelos dependen en gran medida del contexto, con el mismo modelo arrojando resultados variables para preguntas idénticas. De esto, derivamos pautas para realizar y analizar entrevistas con grandes modelos de lenguaje. Si bien se pueden aplicar métodos de análisis de contenido cualitativo, es crucial considerar cuidadosamente los factores contextuales. Las pautas proporcionadas respaldan a investigadores y profesionales para llevar a cabo entrevistas matizadas. En general, desaconsejamos depender de entrevistas con grandes modelos de lenguaje para la investigación debido a su naturaleza impredecible. En cambio, proponemos utilizar estos modelos como herramientas de exploración para obtener diversas perspectivas sobre temas de investigación y validar pautas de entrevistas antes de aplicaciones del mundo real.

Palabras clave: *Modelos de Lenguaje Grandes; Análisis de Contenido Cualitativo; Factores Contextuales*

INTRODUCTION

Conventional qualitative research methods, like individual interviews, often encounter limitations such as small sample sizes and restricted generalizability. On the other hand, quantitative approaches may lack depth and the ability to explore ambiguous responses. Group interviews, while cost-effective, pose challenges like time constraints, selection biases, and compromised generalizability. The advent of large language models like ChatGPT [1] and BARD [2], exhibiting human-like conversational abilities, presents a new avenue for qualitative research. "Interviewing" these models combines the strengths of qualitative and quantitative approaches, accessing a vast dataset of opinions without extensive human interviews. The probabilistic nature of the models identifies probable viewpoints. Simulating interviews with large language models replicates individual interview settings, offering insights into average opinions and attitudes. However, the application of qualitative research methods, particularly qualitative content analysis, to large language models is underexplored, raising uncertainties about their reliability for academic purposes. This paper addresses the following research questions:

1. What variations exist between and within large language models in human-like semi-structured interviews, considering (a) the model used and (b) the language used within the model?
2. What guidelines can be established for (a) conducting interviews with large language models and (b) applying qualitative content analysis to their results?

Background

Large language models (LLMs)

Large language models (LLMs) are AI models based on deep learning, employing neural networks to generate text [3,4]. These models have complex architectures and numerous parameters, trained on extensive document datasets. Unlike older natural language processing approaches using supervised learning, LLMs often employ semi-supervised approaches, facilitating training on large data volumes. Transformer models, introduced by Google [5], expedited large model training through enhanced parallelization, reducing training times compared to older architectures like recurrent neural networks (RNNs). This advancement allowed the development of models pretrained on substantial text datasets, exemplified by Google's BERT (bidirectional encoder representations from transformers) [6].

In 2018, OpenAI introduced the inaugural generative pretrained transformer (GPT) [7], distinguished by its generative capabilities. GPT underwent a two-step training process, involving unsupervised pretraining for general parameter setting and supervised fine-tuning for task adaptation. The initial GPT version was trained on 4.5 GB of text from unpublished books [8]. OpenAI subsequently released updated GPT models, including GPT 3 in 2020, trained on approximately 570 GB of text from a filtered version of Common Crawl. For newer models like GPT 3.5 and GPT 4 [1] released in early 2023, no official information on training data is available. In November 2022, OpenAI launched ChatGPT, derived from GPT 3.5 and fine-tuned for conversational purposes. GPT 4's release included an updated ChatGPT version for paying subscribers. Information on training data and parameters for GPT 3.5, GPT 4, and ChatGPT remains undisclosed. GPT's key strength lies in its capacity to learn conditional probabilities in language, a characteristic explored in this study [10].

While OpenAI's models, notably ChatGPT, dominate the landscape of Large Language Models (LLMs), other companies have introduced competitive alternatives. Meta released LLaMA in February 2023 [11], and Google, pivotal in transformer architecture development, produced LLMs like LaMDA [12] (announced in 2021) and PaLM [13] (available since March 2023, with an updated version, PaLM 2, announced later). PaLM is trained on a diverse dataset, encompassing web documents, books, Wikipedia, conversations, and GitHub code. In response to ChatGPT, Google unveiled their LLM-powered chatbot, BARD, initially based on LaMDA models but currently utilizing PaLM [13].

Challenges in Qualitative Research

Qualitative research entails diverse challenges, with this paper primarily focused on assessing whether generative artificial intelligence proves suitable for addressing these issues. The lack of a singular definition for "qualitative research" necessitates various methods tailored to specific research objectives. However, the grouping of qualitative research methods often leads to inconsistent standards, failing to capture the essence of qualitative research [14].

As qualitative research heavily relies on interpretation and analysis, there is a susceptibility to subjective biases that can impact the findings [15]. Researchers' personal beliefs, experiences, and preconceptions have the potential to inadvertently influence participant selection, data collection methods, and analysis,

leading to biased outcomes [16]. Even the formulation of interview questions can be influenced by judgment and subjectivity, resulting in suggestive questions or restricting participants' freedom of expression. These potential issues necessitate vigilant consideration by researchers throughout the entire process. Furthermore, potential conflicts of interest involving stakeholders in the study pose challenges, and researchers must address these by evaluating critical responses and acknowledging potential conflicts.

Purposive or convenience sampling, commonly employed in qualitative studies, may not yield results representative of a broader population. Limited sample sizes raise concerns about the generalizability of findings [17]. Although sample size is less critical in qualitative studies compared to quantitative studies, the randomization of participants introduces potential challenges. Reproducibility efforts encounter significant challenges, especially when conducting studies in different settings, such as schools, where varying results may arise based on the chosen school and city. Online surveys present their own challenges, as responses tend to be biased toward ambitious participants, potentially skewing study results. Therefore, researchers must carefully consider their sampling strategy and acknowledge the study's limitations.

Ensuring the trustworthiness, validity, and reliability of qualitative findings is an ongoing challenge [18]. In contrast to quantitative research relying on statistical measures for objectivity, qualitative research depends heavily on the researcher's interpretation [19]. While strategies like triangulation, member checking, and inter-rater reliability can enhance validity and reliability, they are not foolproof. Qualitative content analysis, for instance, addresses objectivity concerns by providing verifiable guidelines for text analysis, such as transcripts [20]. Additionally, qualitative research often involves engaging in personal and sensitive discussions with participants [21]. Researchers must ensure informed consent, protect confidentiality and privacy, and navigate ethical dilemmas, including power imbalances and the potential for harm, adding another layer of complexity to qualitative research.

In summary, qualitative research presents a myriad of challenges, encompassing issues like inconsistent standards, subjective biases, sampling limitations, reproducibility challenges, and the establishment of validity and reliability. Conducting qualitative research demands meticulous consideration and strategic

planning, with an additional imperative to address ethical concerns to safeguard participant well-being and privacy. The incorporation of large language models (LLMs) into qualitative research introduces the potential for solutions to these challenges, given that LLMs draw on extensive existing documents, often reflecting real opinions and perspectives. However, this introduction of AI into qualitative research raises questions about bias, transparency, and data privacy. Consequently, our study explores the viability of using qualitative research methods for interviews within artificial conversations with different large language models.

Existing Work on Qualitative Research Methods with AI And LLMs

Research on artificial intelligence (AI) and LLMs has gained prominence, particularly since ChatGPT 3.5, influencing various disciplines such as medicine, healthcare, and economics. However, existing studies predominantly explore AI and LLMs through qualitative research methods without employing these methods with AI. In the realm of education, the utilization of AI and LLMs remains largely unexplored. Notably, Christou [26] investigated the widespread impact of AI in research and academia, focusing on qualitative research through literature and systematic reviews. He outlined key considerations for the appropriate and reliable use of AI, emphasizing understanding AI-generated data, addressing biases and ethical concerns, cross-referencing information, controlling the analysis process, and showcasing the researcher's cognitive input and skills. Christou discussed the role of AI in qualitative research, citing the example of InfraNodus, an AI system designed for tasks like textual data analysis. While AI systems employ predefined rules or algorithms to identify patterns in text data, manual coding or categorization by the researcher is often necessary. The researcher's responsibility includes comprehensive documentation of the analysis methodology and justification, with a crucial role in explaining the rationale and algorithms employed by the AI system. Concluding, Christou highlighted that AI can be effectively used in qualitative research if researchers adhere to specific considerations and guidelines.

In addition to these theoretical discussions, we identified various articles related to interviews conducted with ChatGPT. However, the primary focus of these articles did not revolve around qualitative research involving artificial intelligence. Instead, these articles provided insights into responses to ethical or

social questions and the corresponding answers given by artificial intelligence, as exemplified by references [29] and [30].

Exemplary Topic: Computer Science Education

In our exploration of qualitative analyses with diverse large language models, we aimed to select a topic that would elicit varying opinions across countries without triggering political controversy. This approach was essential to mitigate potential algorithmic limitations in the language models, ensuring a range of acceptable results. We opted for computer science education as our topic, encompassing diverse teaching approaches, including essential content, suitable age levels, and curriculum inclusion. In the subsequent paragraphs, we examine distinct approaches to computer science education implemented in different countries to offer a comprehensive overview of potential perspectives on computer science education.

In Germany, the management of computer science education varies across its 16 states [31], sometimes mandating it as a subject and other times not. The content and objectives differ by state, school type, and curriculum. Despite these variations, efforts are underway to enhance and integrate informatics education in Germany. Bildungsstandards Informatik [32,33] outlines typical contents serving as a foundation for most German curricula. These educational standards cover content areas such as information and data, algorithms, languages and automata, computing systems, and computing and society. Additionally, the standards include practices like modeling and implementing, reasoning and evaluating, structuring and connecting, communicating and cooperating, and representing and implementing. While a version of Bildungsstandards has been published for primary schools [35], the earliest mandatory implementation of computer science education is found in grade 5 in the state of Mecklenburg-Vorpommern [36].

In Switzerland, there is no uniform educational system for computer science, with each canton responsible for determining educational plans. However, there exists a system curriculum for mandatory schooling in German-speaking cantons, incorporating media and IT skills known as "Medien und Informatik" (e.g., in Zürich, see [37]). This curriculum includes computer science education (programming, computer systems, information, and data) alongside application skills and competencies

related to media pedagogy [37]. Implemented with Lehrplan 21 since the school year 2017/18, starting from kindergarten and primary school, students at the upper secondary level have various options based on the type of school, including the opportunity to study computer science as a standalone subject or as part of another discipline.

The United Kingdom recognizes the significance of computer science education; however, there is currently no unified strategy in place. England has incorporated computer science into the "National Curriculum" for computing [39], while Scotland follows the "Curriculum for Excellence" [40], and Northern Ireland adopts the "Revised Curriculum" [41]. Each region has its distinct approach to how, when, and to what extent computer science is included in the curriculum. Nonetheless, all nations concur on the importance of covering computational thinking, online safety, and digital literacy.

In the English curriculum, the three perspectives of computer science, information technology, and digital literacy are integrated, even in primary education [39]. By the end of primary school, students are expected to comprehend fundamental principles and concepts of computer science, such as abstraction, logic, algorithms, and data representation. They should also possess the ability to analyze problems in computational terms and have practical experience in writing computer programs to address these issues. Additionally, students should be adept at evaluating and applying information technology to solve problems, becoming responsible, competent, confident, and creative users of these technologies. England and Scotland are the only nations that have made traditional computer science mandatory in some form, covering topics like programming, algorithms, and data representation [42,43]. All nations share a consensus on the importance of integrating computational thinking, online safety, and digital literacy from primary school to secondary school [44].

Australia employs two approaches to meet the increasing demand for computer science education embedded in its curriculum. The first approach involves a primary subject known as digital technologies [45], which is a standalone subject focused on teaching coding and programming basics while imparting computational thinking through information technology [45]. This includes designing and implementing algorithms, data representation and analysis, core computer hardware concepts, and cybersecurity [45]. The second approach, called information and communication technology (ICT) [46],

is a guideline rather than a specific school subject. It is commonly integrated and taught alongside other subjects across the curriculum as part of Australia's general capabilities system [47]. The objective is to incorporate ICT skills into various areas of learning rather than treating it as a separate discipline. This approach aims to teach students digital literacy and technology skills through activities like conducting research, creating multimedia presentations, and analyzing data within the context of other subject areas [46]. Australia has been making notable progress towards establishing nationwide mandatory computer science education. In 2017, Victoria led the way by introducing digital technologies as a compulsory subject within The Victorian Curriculum F–10 [48], starting as early as year 2 and lasting until the end of year 10. Queensland and South Australia [48] subsequently followed suit by integrating equivalent programs into their curricula.

Computer science education in the United States lacks a standardized system, primarily due to the decentralized nature of the American school system [49]. The commencement grade for computer science instruction varies among schools, states, and school systems. In K-12 curriculum schools, computer science typically initiates in primary school, often integrated with other subjects [49,50]. During secondary school, students can opt for required and elective courses, with computer science not being obligatory in most schools. Some states offer computer science as a high school elective, permitting students to specialize while fulfilling certain graduation requirements [50]. To address these disparities, professional associations like the Computer Science Teachers Association (CSTA) have been actively working on developing and reviewing K-12 standards for computer science since 2004 [51]. The aim is to standardize computer science education nationwide [52]. Presently, 27 states have implemented mandatory computer science education based on CSTA standards, developing their own state K-12 standards aligned with CSTA standards [51,52]. Additionally, computer science courses are now available in 53% of all high schools, with five states mandating completion of a computer science course for high school graduation [52]. The CSTA K-12 Computer Science Standard is designed to span across all grades and school types, integrating computer science into other subjects at the primary school level and either teaching it as a standalone course or integrating it into other subjects at the middle school level [51].

While extensive research has been conducted on how computer science is taught in higher-income countries, there is a limited focus on its establishment in lower-income countries. Nonetheless, this study finds it intriguing to explore the content emphasized in computer science in such countries. For instance, Tshukudu et al. examined four African countries: Botswana, Kenya, Nigeria, and Uganda [53], specifically considering countries classified as economically poorer. Therefore, this section focuses on Nigeria and Uganda. In 2004, Nigeria made "Computer Education" a mandatory subject for primary and secondary schools. Since 2012, it has been mandated that every subject integrates computer science, but resource limitations in many areas hinder students from attending computer science education [53]. Uganda has a competency-based lower secondary computer science education curriculum, with no formal ICT curriculum for primary schools, partially compensated through extracurricular activities [53].

The comparative analysis of computer science education in Germany, Switzerland, the United Kingdom, Australia, the United States, and two African countries underscores the diversity of approaches and strategies adopted by different nations. These variations encompass the significance of computer science education in the school system, recommended age, content, and integration (as a standalone subject or an integrative part of other subjects). Given these distinctions, we utilized these aspects, along with inquiries about methods and tools for computer science education, as focal points for the large language model interviews.

Given the array of approaches available, the inquiry arises as to which one proves most optimal. Departing from the conventional method of consulting experts, we selected this topic as an exemplar for engaging in interviews with large language models. It is reasonable to assume that these models have undergone training on documents that encompass curricula and/or discussions pertaining to these curricula. Consequently, we opted to conduct semi-structured interviews with several large language models, probing their consensus on the most effective teaching methods for computer science, as gleaned from existing literature. Thus, the research inquiries for this exemplar topic, exploring the ideal integration of computer science into primary and secondary education, were:

1. What is the significance of computer science in education?

2. At what point should computer science be incorporated into educational curricula?
3. What computer science topics should be included in school education?
4. What instructional methods are recommended for computer science classes?
5. What tools are advisable for computer science classes?
6. Should computer science be introduced as an independent subject or as an integrated component of other subjects?

METHODOLOGY

Research Design

The research design for this study (refer to Figure 1) was formulated based on our initial research plan. Our objective to conduct semi-structured interviews with generative AI and subsequently employ qualitative research methods for analysis dictated specific predetermined elements of the research design. To steer the interviews, we crafted interview guidelines grounded in the identified ambiguous aspects of computer science education, as detailed in Section 2.4. The complete set of interview guidelines for the semi-structured interviews is available in Appendix A. Interviewers were directed to pose follow-up questions in instances where the AI responses appeared ambiguous. This emphasis was underscored because preliminary tests indicated that large language models (LLMs) often presented different possibilities along with supporting arguments for and against each option. Additionally, it was observed that, even after repeated inquiries (including phrases like "please choose one of your listed options"), if the AI failed to furnish a definitive response, this circumstance should be duly documented. We refrained from adjusting any hyperparameters, such as the answer length, temperature, frequency penalty, top-k, or top-p values, presuming that most researchers attempting to apply generative AI for qualitative data collection would either not modify these settings or would use an interface where such adjustments are not feasible.

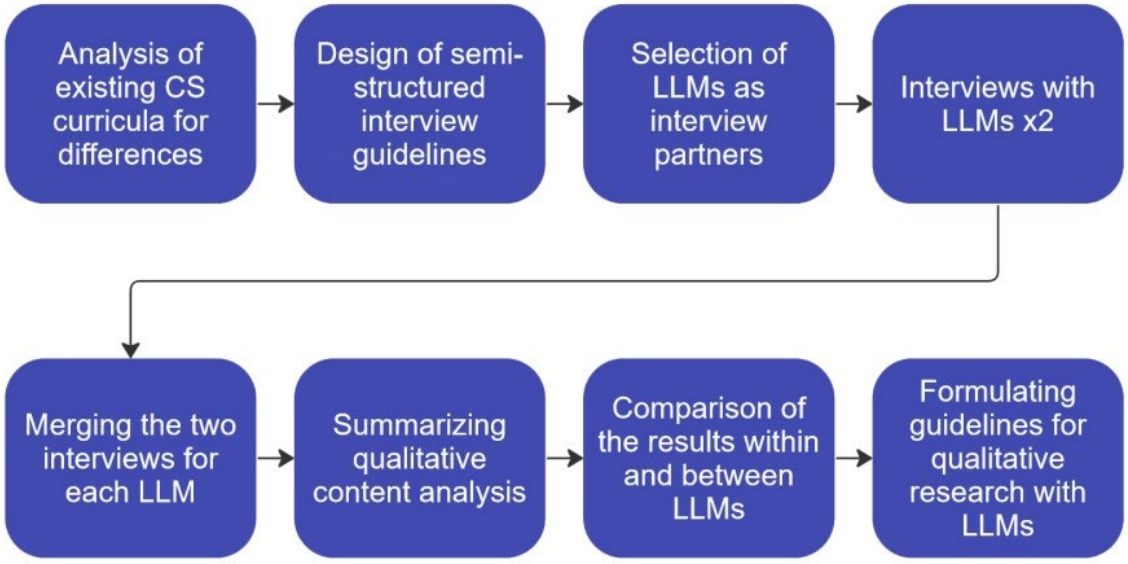

Figure 1 illustrates the study's framework.

Furthermore, during the AI interviewing process, we observed the crucial need to explicitly instruct the large language models (LLMs) to respond based on their "own opinion." Failure to convey this directive beforehand often resulted in ambiguous answers. In some instances, the AI even resorted to citing nonexistent sources to support their statements. While a specific role could be assigned to LLMs by instructing them to act like a particular person, this approach was not employed in the interviews. The goal was to obtain more generalizable answers unbiased by the opinions of specific groups or professions on the topic.

The selection of "participants" for the interviews was based on resource availability and the then-current accessibility of AI models. We had access to ChatGPT in both German and English and intentionally conducted separate interviews in both languages to explore potential differences in responses, potentially stemming from distinct training data. Additionally, we accessed BARD through a VPN, which, at the time of interview preparation, only supported English.

Participants

The interview "partners" were three generative AIs: ChatGPT English, ChatGPT German, and BARD. Initially, we used BARD from Google via a VPN from Germany, as there was no direct access from Germany at the time of writing. Furthermore, we utilized ChatGPT from OpenAI. Both interviews were

conducted in English. Finally, we conducted an interview with SchulKI, a ChatGPT API that can be used by students without an account in German, and translated the results into English afterward.

Data Collection

Each of the three AIs mentioned above underwent two interviews by different interviewers, following the provided interview guidelines. Before the actual interview, we informed the AI about our intent and instructed it to be prepared for the interview. Without this context, the LLMs initiated self-interviewing and printed the result. We deliberately did not specify a specific role for the AI as an interview partner. Subsequently, all questions and answers were copied and saved for later evaluation and coding.

Data Analysis

Each interview underwent two analyses with different raters (not the interviewers), aiming to eliminate diverse interpretations of the answers. All responses to the questions were coded as a whole, forming the basis for the analysis. In cases of varying interpretations, a third rater was consulted to choose one interpretation. For this process, we employed the traditional approach of qualitative content analysis based on Mayring [20], using the method of inductive, summarizing analysis [20] since categories were not formulated before the coding process. In summary, each of the three LLMs was interviewed twice, leading to twelve different code systems, which were then consolidated by a third rater into six distinct code systems, one for each interview.

Results

The precise coding of the extracted interview responses is detailed in Appendix B. The discrepancies in responses between the LLMs in general and within the LLMs are documented in the subsequent two subsections.

Disparities among the LLMs

Although all large language models (LLMs) concurred in acknowledging the importance of computer science in schools, variations surfaced in their recommendations for the age at which computer science lessons should commence. While BARD explicitly suggested a starting age of 5 years, both iterations of ChatGPT only provided age ranges as responses. However, these ranges exhibited considerable

divergence, with ChatGPT recommending the initiation of computer science instruction in an age range spanning from 4 to 12 years.

Regarding the content to be taught, diverse responses emerged, exhibiting significant discrepancies. No precise correspondences were identified, although certain themes such as "computational thinking," "security," or "algorithms" appeared in multiple interviews. However, there was no content universally present in every interview.

Similarities were observed in the methods recommended for computer science classes. Frequently mentioned approaches, such as "problem-based learning" (PBL) or "cooperative/collaborative learning," recurred across interviews. Nevertheless, there was no unanimity among the LLMs regarding a specific method, with variations evident in each interview.

For the first time, consistent responses across all interviews emerged concerning instructional tools. "Scratch" and "Python/Pygame" were cited in every interview. However, the classification of these tools as definitive instruments for teaching computer science remains debatable. Microcontrollers and various online tools designed for informatics purposes were also commonly mentioned.

Notably, on the question of whether computer science should be a separate subject or integrated into other subjects, only BARD provided a concrete statement. Both variants of ChatGPT refrained from making any definitive statement despite repeated inquiries, leading to their categorization as "No clear statement."

Discrepancies within the LLMs

Even within a single large language model, notable divergences were observed in the interviews. In ChatGPT English, substantial disparities in age recommendations were evident between interviews. The first interview suggested an age range of 4 to 7 years, while the second interview recommended an age range of 11 to 12 years, presenting a significant incongruity, particularly concerning child development. Further distinctions were identified in content recommendations, although there were clear commonalities and overlaps. Both interviews referenced "computational thinking," "web development," "AI and ML," and "security and ethics." Additionally, shared themes included "algorithms" and "data." However, variations in content recommendations surfaced beyond these commonalities.

Regarding methods, three consistent recommendations—"gamification," "PBL," and "collaborative learning"—appeared in both interviews with ChatGPT English. However, the first interview included additional suggestions that were not present in the second.

Similar patterns emerged concerning recommended tools for computer science education. "Scratch," "Blockly," "Thunkable," "Pygame," "online tools," "RasPi," and "MicroBit" were proposed as potential tools in both interviews. However, the second interview with ChatGPT English expanded on these recommendations by suggesting additional tools.

Ultimately, in both interviews, ChatGPT English refrained from providing a clear stance on the potential implementation of computer science, either within other subjects or as a distinct standalone subject. Consequently, these responses were deemed equivalent and received the coding "no clear statement" within the context of our question.

There are variations evident in ChatGPT German across the two interviews. Initially, there's a disparity in the suggested age for introducing computer science in schools, with interview 1 proposing a range of 10–12 years and interview 2 suggesting 6–10 years. Given the rapid cognitive development of young children, this difference in recommendation holds significance.

Regarding the recommended content for computer science education in ChatGPT German, some similarities emerge between the interviews. Both responses mention "Fundamentals of computer science," "network security/data security," "ethical and social issues," and "web technologies." Additionally, interview 1 provides three additional distinct answers, while interview 2 offers four additional different responses.

In terms of the recommended methods for computer science education in ChatGPT German, only one overlap in "cooperative learning methods" is observed. Apart from that, both interviews suggest two distinct additional teaching methods.

Several overlaps are identified in the recommended tools between the interviews with ChatGPT German. "Online learning platforms," "Scratch," "Python," "Java," and "visualizations and simulations" are mentioned in both interviews. Additionally, in both interviews, two distinct additional methods are suggested.

Lastly, ChatGPT German also does not provide a concrete answer regarding the categorization of computer science education as a standalone subject or integrated into other subjects.

In the interviews with BARD, there is not only agreement among all interviews regarding the relevance but also an exact match concerning the recommended age for the start of computer science education. Concerning the recommended content, no overlaps are identified between the two interviews. While there may be similarities in terms of terminologies and general topics, no exact overlaps are found.

Three overlaps are identified concerning the recommended methods. "Problem-based learning," "online courses," and "lectures" are mentioned as answers in both interviews. Besides these, two additional divergent suggestions for potential methods are provided in each interview.

Concerning the potential tools, even more overlaps are identified. "Scratch," "Python," "Java, C++," "Unity," "VR/AR," "coding games and apps," and "programming environments" are mentioned in both interviews. While interview 1 provides only one additional distinct suggestion, interview 2 offers three different responses.

Finally, the question regarding integration into another subject or standing as a standalone subject was particularly interesting. BARD provided a specific answer, the only one among the LLMs, but this response differed in the two interviews. While in interview 1 the recommendation leaned towards establishing computer science as a separate and independent subject, in interview 2 it was suggested that computer science should be integrated into other subjects.

DISCUSSION

The focus of our interview centered more on opinions rather than concrete facts, and while our interview guideline was formulated precisely, it inherently allowed for diverse views. The responses varied due to different states of law and educational systems, emphasizing the challenge of training in the field of LLM, as multiple answers are conceivable. However, a comprehensive examination of the training data and corresponding evaluation is necessary to fully understand this diversity. Several findings align with established curricula, such as those observed in Germany (e.g., [36]) and in the UK (e.g., [42]). Nevertheless, in some cases, additional specific questions were necessary to elicit definitive responses

from the language models, exemplified by the challenge in determining the precise age at which computer science education should commence.

The analysis indicates significant differences not only between the language models but also within them, with some models providing conflicting answers, especially in response to inquiries about teaching methods, where the language models often offer more ambiguous responses. Their answers on content and methods tend to be generic, lacking specificity for the subject of computer science in schools. Moreover, their suggestions for didactic tools were limited, primarily mentioning programming languages like Python and Java, which raises ongoing debates about their suitability as proper didactic tools.

For RQ1 (What differences can be observed between and within large language models regarding the results of human-like semi-structured interviews for (a) the used model, (b) the used language within the model?), we concluded that the results varied strongly for all LLMs. While conducting human-like semi-structured interviews with artificial intelligence is feasible, expectations regarding profound and valuable responses should be tempered. Analyzing the results individually, the assessed artificial intelligences provided coherent answers to some extent. However, comparing the answers revealed differences between different models, languages, and even interviewers. Engaging in a conversation with an expert in the respective field cannot be assumed, as evidenced by the presence of ambiguous, imprecise, or generic statements.

Concerning RQ2 (What are guidelines for (a) conducting such interviews with large language models and for (b) using qualitative content analysis on the large language models' results?), the answers are more complex. Prior contextualization of the interview for the artificial intelligence appeared beneficial. Introducing questions with sentence parts like "In your opinion…" improved results, especially for qualitative assessments. However, differences were observed not only between LLMs but also within one LLM. We encountered some of the problems postulated by Christou for qualitative research with LLMs [26], even though they were used for data collection rather than analysis. Specific guidelines will be formulated as part of the Implications. For the second part of the research question, qualitative content analysis methods can, in our perspective, be applied to interviews conducted with artificial

intelligences. This is attributed to both the capabilities of the raters and the progress that artificial intelligences have made in recent years. The origin of the interview is less relevant to the coding process, as long as comprehensible sentences are produced and can be analyzed.

Limitations

Similar to conventional studies, one significant limitation of this study is the relatively small "sample size" represented by the three generative AI systems used. Despite drawing information from an extensive range of data, documents, and perspectives, the chosen AI models, though state-of-the-art, may not fully encapsulate the diverse spectrum of viewpoints that could be gleaned from a larger, more varied pool of AI models. Consequently, the study's findings may be specific to the inherent characteristics and biases of the selected AI models and may not be universally representative of other AI systems.

The generalizability of the findings, as well as the implications and guidelines presented in the subsequent section, is another noteworthy limitation. The responses generated by the AI models originated from their pretraining on specific datasets and the subsequent fine-tuning process. These models might not inherently reflect the perspectives or knowledge of real-world computer science experts or educators but rather calculate word likelihoods without assessing the validity of a document within the training data. It is, therefore, important to acknowledge that the provided answers may not always be factually correct. Although this wasn't an issue for this study, given its focus on obtaining opinions about CS education rather than factual information, this limitation should be considered for other applications.

The use of AI models in research introduces ethical concerns, encompassing potential biases and the consequences of algorithmic decision-making. The AI models employed in this study were trained on substantial datasets that could harbor biases present in the data. Consequently, the responses generated by the AI models might inadvertently mirror or perpetuate these biases. A critical analysis and interpretation of the AI-generated responses are essential, taking into account potential biases embedded in the models' training data and their impact on the study's findings.

This study heavily relied on the AI models' responses at a specific point in time, with limited opportunities for real-time feedback or iterative refinement of the models' understanding or responses. The field of AI is evolving rapidly, and newer models or updates to existing models may have improved performance or generated more nuanced responses since the study's completion. The questions posed to the AI models were specific to inquiries about the implementation of computer science education in schools, designed to explore key aspects but not covering the entire breadth and depth of issues related to computer science education.

Implications

Taking into account the identified limitations and the outcomes of this study, guidelines were formulated (also refer to Figure 2) for the utilization of large language models (LLMs) in three specific realms of qualitative research: predicting probable opinions, assessing interview guidelines, and exploring potential responses. In consideration of these application areas, the following recommendations are put forth:

- Given their unreliability and unpredictable outcomes, it is generally discouraged to employ LLMs for the collection of qualitative data intended for academic purposes.
- LLMs can be utilized as an initial tool to gain exploratory insights and potential opinions on a specific topic.
- When aiming for validity and coherence in the results, particularly up to a certain extent, it is advisable to conduct multiple iterations of the same interview and set a low temperature to generate more coherent answers. To mitigate ambiguous responses, interviewers should preface each question with the phrase "In your opinion..."
- LLMs can effectively contribute to the evaluation of interview guidelines, particularly concerning clarity and comprehensibility.
- In scenarios where the objective is to test interview guidelines or uncover diverse potential opinions, a moderate temperature should be applied (consideration can also be given to assigning different identities).

- If the LLM is expected to adopt a specific role, explicit instruction such as "Please take on the role of..." is crucial. Caution must be exercised to prevent the perpetuation of stereotypes in such cases.

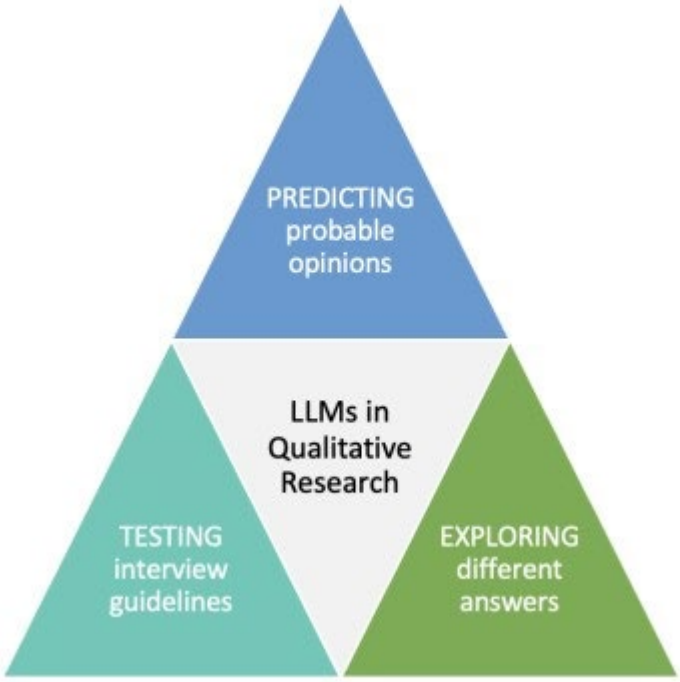

Figure 2. Utilization Scenarios of LLMs in Qualitative Research: Predicting, Testing, and Exploring.

Contrary to the initial hypothesis, these guidelines dissuade the use of LLMs for the collection of academic data in isolation. However, while LLMs should not be the sole source of information, they can assist in gaining an initial overview of a subject. Given that the generated content is influenced by probabilities, as well as random and variable factors, conducting multiple interviews becomes essential for a thorough prediction of answer likelihoods. Moreover, specific phrasings, such as "In your opinion..." or "Please take on the role of..." can be employed to tailor the results. The most advantageous applications of LLMs in qualitative research might lie in testing interview guidelines or exploring different potential answers to interview questions. Another notable advantage is the patience of an LLM, as interviewers can delve deep into questioning without concerns regarding human emotions.

CONCLUSIONS

The present study investigated the viability of using large language models (LLMs) as interview partners in qualitative research on computer science education. Through semi-structured interviews

with three LLMs, namely BARD, ChatGPT from OpenAI, and GPTSchule, valuable insights were obtained. However, it is essential to acknowledge the limitations and biases inherent in using LLMs for such research.

This study revealed that LLMs can serve as an initial means of obtaining exploratory insights into a specific topic. They can offer a broad range of responses, but it is crucial to conduct multiple iterations of the same interview to enhance the validity and reliability of the results. By preceding questions with phrases such as "In your opinion..." or "Please take on the role of...", interviewers can tailor the responses and mitigate potential biases.

However, it became evident that LLMs have certain limitations when it comes to generating precise and context-rich answers. In specific didactic inquiries, LLMs tend to remain somewhat vague, and their responses may lack creative approaches to continuously evolving teaching methods. Additionally, the lack of real-world experiences and awareness of specific educational systems limits the generalizability of the responses. This is especially important when trying to provide "true" answers, as argued by Sobieszek and Price: "The real reason GPT's answers seem senseless being that truth-telling is not amongst them. We claim that these kinds of models cannot be forced into producing only true continuation, but rather to maximize their objective function they strategize to be plausible instead of truthful" [10].

The findings of this study also highlighted potential ethical concerns related to using LLMs in qualitative research. These AI models are trained on large datasets, which may contain biases that can inadvertently influence their responses. Researchers must exercise caution and critically analyze the generated data to prevent the perpetuation of biases.

Based on the study's results, we recommend utilizing LLMs cautiously as an initial exploration tool. Researchers should not rely solely on LLM-generated responses but should combine them with more traditional qualitative research methods involving human participants. By doing so, the research can benefit from the strengths of both AI-driven insights and the depth and context provided by human experiences.

Appendix A

Interview Guidelines

1. Extend a warm welcome to the LLM, provide clarity on the interview setting, and introduce the overarching topic of Computer Science Education.
2. Inquire about the LLM's connection to computer science. Explore how the LLM acquired knowledge in computer science.
3. Assess the perceived relevance of computer science as a school subject.
4. Determine the suggested starting age for teaching computer science.
5. Solicit the LLM's opinion on the specific content that should be included in a computer science school subject.
6. Elicit insights on the preferred methods and tools deemed beneficial for teaching this content.
7. Seek the LLM's perspective on whether this content should be delivered as a standalone subject or integrated into other school subjects.
8. Conclude the interview, express gratitude to the LLM, and bid farewell.

Appendix B

Table A1. Encoded responses from the interviews

	ChatGPT English		ChatGPT German		BARD	
Interview	1	2	1	2	1	2
Relevance	Very high	Very high	Very high	Very high	Very high	Very high
Age	4–7	11–12	10–12	6–10	5	5
Contents	Computation-al Thinking, Web Development, AI and ML, Security and Ethics, Programming Concepts, Data structures and Algorithms, Software, Databases, Networks	Computation-al Thinking, Web Development, AI and ML, Security and Ethics, Hard- and Software, Programming, Algorithms, Data and Information, Cyber-Security, AR, VR, IoT, OS, Robotics, Problem solving	Programming Skills, Network Security, Information Management, AI and Robotics, Fundamentals of Computer Science	Networks, Data Security, Databases, Web Technologies, Fundamentals of Computer Science, Creative and Critical Thinking Skills, Media Literacy, Ethical and Social Issues	Basics of Computer Science, Data Structures, Software, Programming, Project Work, Website-, Game- and App Development, Hardware, AI and ML	Variables, Data Structures, Functions, Arrays, OOP, Algorithms, Loops, Branching, Ethical and Social Implications, Problem solving Strategies, Algorithmic Thinking
Methods	Gamification, PBL, Collaborative Learning, Peer Learning, Direct Instruction, Experimental Learning, Flipped Classroom	Gamification, PBL, Collaborative Learning	Independent Work, Project Work, Cooperative Learning Methods	Cooperative Learning Methods, Field Trips, Simulations and Visualizations	Problem Based Learning, Online Courses, Lectures, Project Based Learning, Programming Games and Apps	Problem Based Learning, Online Courses, Lectures, Hands-on-activities, Projects
Tools	Scratch, Blockly, Thunkable, Pygame, Online Tools, RasPi, MicroBit, Robotics, Visual Programming Languages, HTML, CSS, Javascript, Python, Java, GitHub, repl.it, Google Colab, Khan Academy, Code.org	Scratch, Blockly, Thunkable, Pygame, Online Tools, RasPi, MicroBit	Online Learning Platforms, Scratch, Python, Java, Visualizations and Simulations, Educational Games, Field Trips	Online Learning Platforms, Scratch, Python, Java, Visualizations and Simulations, Robots and Microcontrollers, Maker Spaces	Scratch, Python, Java, C++, Unity, VR/AR, Coding Games and Apps, Programming Environments, Hardware	Scratch, Python, Java, C++, Unity, VR/AR, Coding Games and Apps, Programming Environments, Online resources, Programming languages, Software development tools
Subject	No clear statement	No clear statement	No clear statement	No clear statement	Own Subject	Integrated in other Subjects

BIBLIOGRAPHICAL REFERENCES

OpenAI. GPT-4 Technical Report. 2023. Available online: http://xxx.lanl.gov/abs/2303.08774 (accessed on 18 July 2023).

Google. Bard Experiment. 2023. Available online: https://bard.google.com (accessed on 18 July 2023).

Shen, Y.; Heacock, L.; Elias, J.; Hentel, K.D.; Reig, B.; Shih, G.; Moy, L. ChatGPT and other large language models are double-edged swords. *Radiology* **2023**, *307*, e230163. [CrossRef] [PubMed]

Rillig, M.C.; Ågerstrand, M.; Bi, M.; Gould, K.A.; Sauerland, U. Risks and benefits of large language models for the environment.

Environ. Sci. Technol. **2023**, *57*, 3464–3466. [CrossRef] [PubMed]

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, Long Beach, CA, USA, 4–9 December 2017; Volume 30.

Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding.

arXiv **2018**, arXiv:1810.04805.

Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I. Improving Language Understanding by Generative Pre-Training. 2018. Available online: https://www.mikecaptain.com/resources/pdf/GPT-1.pdf (accessed on 18 July 2023).

Zhu, Y.; Kiros, R.; Zemel, R.; Salakhutdinov, R.; Urtasun, R.; Torralba, A.; Fidler, S. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 19–27.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1877–1901.

Sobieszek, A.; Price, T. Playing games with AIs: The limits of GPT-3 and similar large language models. *Minds Mach.* **2022**, *32*, 341–364. [CrossRef]

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. Llama: Open and efficient foundation language models. *arXiv* **2023**, arXiv:2302.13971.

Thoppilan, R.; De Freitas, D.; Hall, J.; Shazeer, N.; Kulshreshtha, A.; Cheng, H.T.; Jin, A.; Bos, T.; Baker, L.; Du, Y.; et al. Lamda: Language models for dialog applications. *arXiv* **2022**, arXiv:2201.08239.

Chowdhery, A.; Narang, S.; Devlin, J.; Bosma, M.; Mishra, G.; Roberts, A.; Barham, P.; Chung, H.W.; Sutton, C.; Gehrmann, S.; et al. Palm: Scaling language modeling with pathways. *arXiv* **2022**, arXiv:2204.02311.

Oppong, S.H. The problem of sampling in qualitative research. *Asian J. Manag. Sci. Educ.* **2013**, *2*, 202–210.

Mruck, K.; Breuer, F. Subjectivity and reflexivity in qualitative research—A new FQS issue. *Hist. Soc. Res./Hist. Sozialforschung* **2003**, *28*, 189–212.

Chenail, R.J. Interviewing the investigator: Strategies for addressing instrumentation and researcher bias concerns in qualitative research. *Qual. Rep.* **2011**, *16*, 255–262.

Higginbottom, G.M.A. Sampling issues in qualitative research. *Nurse Res.* **2004**, *12*, 7. [CrossRef] [PubMed]

Whittemore, R.; Chase, S.K.; Mandle, C.L. Validity in qualitative research. *Qual. Health Res.* **2001**, *11*, 522–537. [CrossRef] [PubMed]

Thomson, S.B. Qualitative research: Validity. *Joaag* **2011**, *6*, 77–82.

Mayring, P. Qualitative content analysis. *Companion Qual. Res.* **2004**, *1*, 159–176.

Surmiak, A.D. *Confidentiality in Qualitative Research Involving Vulnerable Participants: Researchers' Perspectives*; Forum Qualitative Sozialforschung/Forum: Qualitative Social Research (FQS): Berlin, Germany, 2018; Volume 19.

Laï, M.C.; Brian, M.; Mamzer, M.F. Perceptions of artificial intelligence in healthcare: Findings from a qualitative survey study among actors in France. *J. Transl. Med.* **2020**, *18*, 14. [CrossRef] [PubMed]

Haan, M.; Ongena, Y.P.; Hommes, S.; Kwee, T.C.; Yakar, D. A qualitative study to understand patient perspective on the use of artificial intelligence in radiology. *J. Am. Coll. Radiol.* **2019**, *16*, 1416–1419. [CrossRef] [PubMed]

Yang, Y.; Siau, K.L. A Qualitative Research on Marketing and Sales in the Artificial Intelligence Age. Available online: https://www.researchgate.net/profile/Keng-Siau-2/publication/325934359_A_Qualitative_Research_on_Marketing_and_Sales_in_the_Artificial_Intelligence_Age/links/5b9733644585153a532634e3/A-Qualitative-Research-on-Marketing-and-Sales- in-the-Artificial-Intelligence-Age.pdf (accessed on 18 July 2023).

Longo, L. Empowering qualitative research methods in education with artificial intelligence. In *World Conference on Qualitative Research*; Springer: Cham, Switzerland, 2019; pp. 1–21.