



Article

Qualitative Research Methods for Large Language Models: Conducting Semi-Structured Interviews with ChatGPT and BARD on Computer Science Education

Andreas Dengel ^{*}, Rupert Gehrlein ^{*}, David Fernes , Sebastian Görlich, Jonas Maurer, Hai Hoang Pham, Gabriel Großmann and Niklas Dietrich genannt Eisermann

Institute for the Didactics of Mathematics and Computer Science, Goethe University Frankfurt, 63025 Frankfurt, Germany; fernes@em.uni-frankfurt.de (D.F.); goerlich@mathematik.uni-frankfurt.de (S.G.); maurerjonas1@stud.uni-frankfurt.de (J.M.); hoangphamhai@stud.uni-frankfurt.de (H.H.P.); s1188977@stud.uni-frankfurt.de (G.G.); s5110057@stud.uni-frankfurt.de (N.D.g.E.)

* Correspondence: dengel@math.uni-frankfurt.de (A.D.); r.gehrlein@em.uni-frankfurt.de (R.G.)



Citation: Dengel, A.; Gehrlein, R.; Fernes, D.; Görlich, S.; Maurer, J.; Pham, H.H.; Großmann, G.; Eisermann, N.D.g. Qualitative Research Methods for Large Language Models: Conducting Semi-Structured Interviews with ChatGPT and BARD on Computer Science Education. *Informatics* **2023**, *10*, 78. <https://doi.org/10.3390/informatics10040078>

Academic Editors: Antony Bryant, Roberto Montemanni, Min Chen, Paolo Bellavista, Kenji Suzuki and Jeanine Treffers-Daller

Received: 20 July 2023

Revised: 13 September 2023

Accepted: 7 October 2023

Published: 12 October 2023

Abstract: In the current era of artificial intelligence, large language models such as ChatGPT and BARD are being increasingly used for various applications, such as language translation, text generation, and human-like conversation. The fact that these models consist of large amounts of data, including many different opinions and perspectives, could introduce the possibility of a new qualitative research approach: Due to the probabilistic character of their answers, “interviewing” these large language models could give insights into public opinions in a way that otherwise only interviews with large groups of subjects could deliver. However, it is not yet clear if qualitative content analysis research methods can be applied to interviews with these models. Evaluating the applicability of qualitative research methods to interviews with large language models could foster our understanding of their abilities and limitations. In this paper, we examine the applicability of qualitative content analysis research methods to interviews with ChatGPT in English, ChatGPT in German, and BARD in English on the relevance of computer science in K-12 education, which was used as an exemplary topic. We found that the answers produced by these models strongly depended on the provided context, and the same model could produce heavily differing results for the same questions. From these results and the insights throughout the process, we formulated guidelines for conducting and analyzing interviews with large language models. Our findings suggest that qualitative content analysis research methods can indeed be applied to interviews with large language models, but with careful consideration of contextual factors that may affect the responses produced by these models. The guidelines we provide can aid researchers and practitioners in conducting more nuanced and insightful interviews with large language models. From an overall view of our results, we generally do not recommend using interviews with large language models for research purposes, due to their highly unpredictable results. However, we suggest using these models as exploration tools for gaining different perspectives on research topics and for testing interview guidelines before conducting real-world interviews.

Keywords: large language models; qualitative research; interviews; computer science education; artificial intelligence; machine learning



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Traditional qualitative research approaches, such as individual interviews, often face limitations in terms of small sample sizes and limited generalizability. Conversely, quantitative research methods may lack depth and the ability to probe further when answers are ambiguous. Group interviews, while cost-effective, present challenges in terms of time constraints, selection biases, and group dynamics, further compromising the generalizability of the results. However, with the emergence of large language models such as ChatGPT [1]

and BARD [2], which exhibit human-like conversation abilities based on extensive training data from diverse sources, a new possibility for qualitative research has arisen.

By “interviewing” these language models, it may be possible to combine the strengths of both qualitative and quantitative approaches. This method would allow researchers to access a vast amount of data, containing numerous opinions and perspectives, without having to conduct interviews with a large amount of people. At the same time, the probabilistic nature of the models could be used to identify the most probable viewpoints.

Creating a simulated interview environment using large language models could replicate an individual interview setting, enabling insights that reflect average opinions and attitudes on a particular topic, but could also be used to test interview guidelines before conducting interviews with humans. Notably, qualitative research methods, including qualitative content analysis, have not been extensively applied to large language models thus far. Therefore, it remains uncertain whether conducting interviews with these models can yield meaningful and reliable results that could be used for academic purposes. This leads to the research questions for this paper:

1. 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?
2. 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?

2. Background

2.1. What Are LLMs?

Large language models (LLMs) are artificial intelligence (AI) models based on deep learning (i.e., neural networks) used to generate text [3,4]. These models have a complex underlying architecture and a large number of parameters, trained on very large amounts of existing documents. While many older natural language processing approaches used supervised learning for specific tasks, most LLMs use semi-supervised approaches, which makes it easier to train them on large quantities of data.

The introduction of transformer models by Google [5] also helped to train large models more quickly, since this new architecture allowed for greater parallelization of training, which reduced training times compared to older architectures such as recurrent neural networks (RNNs). This allowed the creation of models that were pretrained on large amounts of text, such as Google’s BERT (bidirectional encoder representations from transformers) [6].

In 2018, OpenAI introduced the first generative pretrained transformer (GPT) [7]. While there had been other pretrained models before, GPT also had generative capabilities. The model was trained with a mixture of unsupervised pretraining to set general parameters and a supervised fine-tuning step in order to adapt to specific tasks. This first version of GPT was trained on 4.5 GB of text from unpublished books [8]. In the following years, OpenAI released several updated GPT models. In 2020, they published GPT 3 [9], which was trained on around 570 GB of text from a filtered version of Common Crawl, a openly available crawl of the internet. For newer models such as GPT 3.5 and GPT 4 [1], which were released in early 2023, no official information about the training data is available. In November 2022, OpenAI released ChatGPT which is based on GPT 3.5 and is fine-tuned for chatting. With the release of GPT 4, a version of ChatGPT using the newer model was also made available to paying subscribers. Just like the GPT 3.5 and 4 models, no official information on training data and parameters are available for ChatGPT. In general, the abilities of GPT lie in “its learning conditional probabilities in language (its so-called “statistical capabilities”)” [10], which is also the characteristic that we are going to utilize within this study.

While OpenAIs models are the most prominent LLMs, they are not the only ones available. The large popularity of ChatGPT lead to other companies releasing their own competitors. For example Meta released LLaMA (Large Language Model Meta AI) in

February 2023 [11]. Google, who laid the groundwork for LLMs with the development of the transformer architecture, also developed their own LLMs. One of those is LaMDA (Language Model for Dialogue Applications) [12], which was first announced in 2021. Another model called PaLM (Pathways Language Model) [13] was first made available in March 2023. An updated version called PaLM 2 was announced later that year. The PaLM model is trained on a collection of web documents, books, Wikipedia, conversations, and GitHub code. In response to the release of ChatGPT, Google launched their own LLM-powered chatbot named BARD in March 2023. It was originally based on LaMDA models, but now uses PaLM [13].

2.2. Potential Problems in Conducting Qualitative Research

In qualitative research, there are several challenges that need to be addressed. While we will only focus on some of these challenges, the main objective of this paper was to assess whether generative artificial intelligence is a suitable tool for addressing these issues. “Qualitative research” does not have a singular definition, similarly to quantitative research. Different qualitative methods are required to cater to specific research objectives [14]. However, qualitative research methods are often grouped together, leading to the application of inconsistent standards. This fails to capture the essence of qualitative research.

As interpretation and analysis play a crucial role in qualitative research, there is the potential for subjective biases to influence the findings [15]. Researchers’ personal beliefs, experiences, and preconceptions can inadvertently affect participant selection, data collection methods, and analysis, leading to biased results [16]. For instance, even the formulation of interview questions can be affected by judgment and subjectivity, leading to suggestive questions or limiting participants’ freedom of expression. All these potential issues require careful consideration from the researchers’ side throughout the process.

Moreover, potential conflicts of interest may arise concerning the stakeholders involved in the study. Evaluating and including critical responses, as well as acknowledging possible conflicts of interest, constitute challenges for researchers. These challenges are not only related to the interview method, but similar complications can occur with other forms of qualitative research.

Purposive or convenience sampling is often used in qualitative studies, which may not provide representative results for a broader population. Limited sample sizes also raise concerns about the generalizability of findings [17]. While sample size is less critical in qualitative studies compared to quantitative studies, randomization of participants still poses a potential problem. Efforts are often made to enhance the reproducibility of qualitative studies; however, significant challenges exist. For example, if a study is conducted using randomization in a school setting, different results may be obtained, depending on which school and city are chosen. Online surveys can also present challenges, as the responses tend to be primarily from ambitious participants who may bias the study’s results. Consequently, researchers must carefully consider their sampling strategy and acknowledge the limitations of their study.

Establishing the trustworthiness, validity, and reliability of qualitative findings is another persistent challenge [18]. Unlike quantitative research, which can rely on statistical measures for objectivity, qualitative research heavily depends on the researcher’s interpretation [19]. Strategies such as triangulation, member checking, and inter-rater reliability can enhance validity and reliability but are not foolproof. Qualitative content analysis (e.g., [20]) tries to address the issue of objectivity in particular by providing verifiable guidelines for the analysis of a text (e.g., transcripts). Additionally, qualitative research often involves engaging in personal and sensitive discussions with participants (e.g., [21]). Researchers must ensure informed consent, protect confidentiality and privacy, and navigate ethical dilemmas, including power imbalances and the potential for harm. These ethical considerations add another layer of complexity to qualitative research.

In conclusion, qualitative research poses various challenges that researchers need to address. From issues related to inconsistent standards and subjective biases to sampling

limitations, reproducibility challenges, and establishing validity and reliability, conducting qualitative research requires careful consideration and strategic planning. Furthermore, ethical concerns must be taken into account to ensure the well-being and privacy of participants. As large language models are based on a large amount of existing documents, and thus reflecting (most of the time) real opinions and perspectives on certain topics, AI could potentially offer some potential solutions to certain challenges. However, its introduction into qualitative research also raises questions and considerations about bias, transparency, and data privacy. Therefore, we conducted a study with different large language models, to assess the suitability of using qualitative research methods for interviews within artificial conversations with these models.

2.3. Existing Work on Qualitative Research Methods with AI And LLMs

Artificial intelligences and LLMs have not only been a topic of research since ChatGPT 3.5 and the associated breakthrough in public perception but already support numerous disciplines in practice, such as in medicine [22], healthcare [23], and economics [24]. All of these studies primarily focus on the exploration of artificial intelligences through the use of qualitative research methods and do not employ qualitative research methods with artificial intelligences. In education, on the other hand, the use of AI and LLM is largely unexplored. For example, one of the few research papers in this area addressed how education, teaching, and learning could be improved by using AI for qualitative data analysis [25].

Christou [26] examined the widespread impact of artificial intelligence in research and academia, particularly in qualitative research through literature and systematic reviews, addressing its strengths, limitations, ethical dilemmas, and potential biases. He proposed five key considerations for its appropriate and reliable use, including understanding AI-generated data, addressing biases and ethical concerns, cross-referencing information, controlling the analysis process, and demonstrating the cognitive input and skills of the researcher throughout the study. In Christou's discussion on the role of AI in qualitative research, the example of *InfraNodus* was given: This AI system was designed to perform various tasks related to textual data analysis, such as the categorization of information, cluster creation, and creating visual graphs. This can, for example, be used on texts created from interviews [26]. Furthermore, it can be assumed that the researcher should engage in some degree of manual coding or categorization, due to the reliance on analytical software or AI systems, which often employ predefined rules or algorithms to identify patterns, themes, or keywords in text data. Additionally, the researcher is responsible for providing comprehensive documentation of the analysis methodology, justification, and precise execution procedure employed. It is essential for the researcher to be able to explain the rationale and algorithms employed by the AI system in conducting the analysis, and the researcher's cognitive evaluative skills play a valuable role in the analytical process and the formulation of conclusions [26]. Drawing such conclusions from an AI's analyses to answer practical research questions requires the expertise and contextual knowledge of the researcher [27,28]. It can be concluded that AI can be used in qualitative research (e.g., for systematic reviews, qualitative empirical studies, and conceptual studies), but only if the researcher adheres to certain key considerations and guidelines [26].

In addition to these theoretical discussions, we were able to find various articles concerning interviews with ChatGPT. However, the primary focus was not on qualitative research work involving artificial intelligence. Instead, the articles shed light on the responses to ethical or social questions and the corresponding answers provided by the artificial intelligence (e.g., [29] or [30]).

2.4. Exemplary Topic: Computer Science Education

In our study of qualitative analyses with different large language models, we sought a topic that would provide differing opinions across countries, without raising political controversy. This was necessary to avoid potential algorithmic limitations of the language models by ensuring a variety of acceptable results. We chose computer science education as

our topic, as it encompasses various approaches to teaching computer science, including the essential content, appropriate age levels, and inclusion in the curriculum. In the following paragraphs, we analyze different approaches to computer science education implemented in different countries, in order to provide an overview of possible perspectives on computer science education.

Germany manages computer science education differently across its 16 states [31], sometimes with and sometimes without a mandatory subject. The content and objectives vary by state, school type, and curriculum. Despite these variations, efforts are being made to improve and integrate informatics education in Germany. Typical contents are described in the Bildungsstandards Informatik [32,33], which are used as a basis for most German curricula. These educational standards include the following *content standards* [34]: *information and data, algorithms, languages and automata, computing systems, and computing and society*. Furthermore, the standards include *practices*: *modeling and implementing, reasoning and evaluating, structuring and connecting, communicating and cooperating, and representing and implementing*. Even though a version of the Bildungsstandards has been published for primary schools as well [35], the earliest mandatory implementation of computer science education can be found in grade 5, in the state Mecklenburg-Vorpommern [36].

In Switzerland, there is no uniform educational system for computer science, with each canton responsible for determining educational plans. However, there has been a system curriculum for mandatory schooling in German-speaking cantons that includes media and IT skills, called “Medien und Informatik” (e.g., in Zürich, see [37]). The subjects include computer science education (i.e., programming, computer systems, information, and data), but also application skills and competencies related to media pedagogy [37]. The curriculum has been implemented with the Lehrplan 21 since the school year 2017/18, beginning from kindergarten and primary school [38]. At the upper secondary level, students have different options depending on the type of school, including the choice to study computer science as a standalone subject or as part of another subject.

The United Kingdom recognizes the importance of computer science education, but there is currently no unanimous strategy in place. England has implemented computer science in the “National Curriculum” for computing [39], while Scotland has the “Curriculum for Excellence” [40], and Northern Ireland has the “Revised Curriculum” [41]. Each region has its own approach to how, when, and to what degree computer science is taught. However, all nations agree on the importance of covering computational thinking, online safety, and digital literacy. For the English curriculum, the three perspectives computer science, information technology, and digital literacy are integrated, even in primary education [39]: At the end of primary school, students should understand the fundamental principles and concepts of computer science, such as abstraction, logic, algorithms, and data representation. Students should also be able to analyze problems in computational terms with repeated practical experience of writing computer programs to solve these problems. They should be able to evaluate and apply information technology to solve problems and they should become responsible, competent, confident, and creative users of these technologies. England and Scotland are the only nations that made traditional computer science mandatory in some form, which covers topics such as programming, algorithms, and data representation [42,43]. All nations agree on the importance of covering computational thinking, online safety, and digital literacy from primary school to secondary school [44].

Australia has two approaches to meeting the growing demand for computer science education anchored in their curriculum. The first one is a primary subject addressing the demand, called digital technologies [45], which is a standalone subject, focusing on teaching the basics of coding and programming, while also training students computational thinking through IT [45]. This includes the design and implementation of algorithms, representation and analysis of data, core concept of computer hardware, and lastly 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 aim 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 such as conducting research, creating multimedia presentations, and analyzing data within the context of other subject areas [46]. Australia has been making significant progress towards establishing nationwide mandatory computer science education. Victoria paved the way in 2017 by introducing digital technologies as a compulsory subject within The Victorian Curriculum F–10 [48], which starts as early as year 2 and lasts until the end at 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 uniform system due to the decentralized nature of the American school system [49]. The grade at which computer science is taught varies among schools, states, and school systems. In schools with a K-12 curriculum, computer science typically starts in primary school and is often integrated with other subjects [49,50]. In secondary school, students have the option to choose from required and elective courses, with computer science not being a mandatory subject in most schools. Some states offer computer science as an elective at the high school level, allowing students to specialize with certain graduation requirements [50]. To address these issues, professional associations such as the Computer Science Teachers Association (CSTA) have been working towards developing and reviewing K-12 standards [51] for computer science since 2004. The goal is to standardize computer science education across the country [52]. As of now, 27 states have implemented mandatory computer science education based on CSTA standards, and they have also developed their own state K-12 standards in alignment with the CSTA standards [51,52]. Additionally, computer science courses are now offered in 53% of all high schools, with five states making completion of a computer science course a requirement for high school graduation [52]. The CSTA K-12 Computer Science Standard is designed to spiral through all grades and school types, with computer science being integrated into other subjects at the primary school level and either taught as a standalone course or integrated into other subjects at the middle school level [51].

While there has been a lot of research on how computer science should be taught in higher-income countries, there has been little research on how computer science is established in lower-income countries. However, it was interesting for this work to see what content in computer science such countries emphasize. For example, Tshukudu et al. studied four African countries: Botswana, Kenya, Nigeria, and Uganda [53]. Countries classified as poorer are examined here. Therefore, only Nigeria and Uganda are the subjects of this section. In 2004, Nigeria made "Computer Education" a compulsory subject for primary and secondary schools. Since 2012, it has been mandated that every subject integrates CS. Due to a lack of resources in many areas of Nigeria, students are not able to attend CS education [53]. A competency-based lower secondary CS education curriculum is in place in Uganda. There is no formal ICT curriculum for primary schools, and this is partially compensated with extracurricular activities [53].

The comparative analysis of computer science education in Germany, Switzerland, the United Kingdom, Australia, the United States, and two African countries reveals the range of approaches and strategies employed by different nations. These differences occur in the areas of importance of computer science education in the school system, recommended age, contents, and integration (as a separate subject or as an integrative part of other subjects). Owing to these differences, we used these aspects, as well as questions about methods and tools for computer science education, as discussion points for the LLM interviews.

Given the various approaches at hand, the question arises of which one is the most optimal. Instead of the traditional approach of interviewing experts, we chose this topic as an exemplary topic for conducting interviews with large language models. It can be assumed that these models have been trained on documents including the curricula and/or discussions about these curricula. Consequently, we opted to conduct semi-structured interviews with several large language models, exploring their consensus on the optimal

teaching methods for computer science, as reflected in the existing literature. Therefore, the research questions for this exemplary topic of the ideal integration of computer science into primary and secondary education were

1. How relevant is computer science for education?
2. At what age should computer science be integrated in education?
3. What computer science contents should be taught in schools?
4. What methods should be used for computer science classes?
5. What tools should be used for computer science classes?
6. Should computer science be implemented as a separate subject or as an integrative part of other subjects?

3. Methodology

3.1. Research Design

The research design for this study (see Figure 1) was based on our initial research plan. Our intention to conduct semi-structured interviews with generative AI and subsequently analyze them using qualitative research methods necessitated certain predetermined aspects of the research design. To guide the interviews, we developed interview guidelines based on the identified unclear aspects of computer science education outlined in Section 2.4. The full interview guidelines for the semi-structured interviews can be found in Appendix A. Interviewers were instructed to ask follow-up questions in cases where the responses from the AI were ambiguous. This emphasis was placed because initial tests indicated that the LLMs often provided different possibilities with supporting arguments for and against each possibility. Moreover, it was noted that if, despite repeated inquiries (including phrases such as “please choose one of your listed options”), the AI did not provide a definitive response, this should be documented. We did not adjust any hyperparameters, such as the length of the answer, temperature, frequency penalty, top-k, or top-p values, as we assumed that most researchers trying to apply generative AI for collecting qualitative data would not change these settings or would use an interface where changing these settings is not possible.

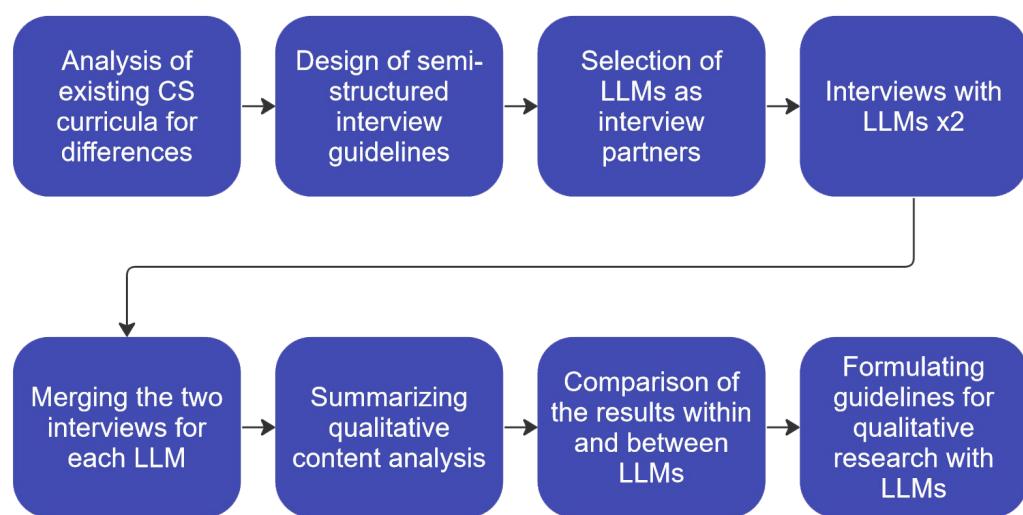


Figure 1. Structure of the study.

In addition, we discovered during the interviewing process with the AI that it was crucial to explicitly instruct the LLMs to respond based on their “own opinion”. Failing to convey this directive beforehand often resulted in the AI providing ambiguous answers. In some cases, they even resorted to citing nonexistent sources to support their statements.

While a specific role can be assigned to LLMs by telling them to act like a specific person, this approach was not used in the interviews. This was done in order to obtain

more generalizeable answers that were not biased by the opinions of certain groups or professions on the topic.

The selection of “participants” for the interviews was based on the availability of resources and the then-current availability of AI models. We had access to ChatGPT in both German and English and deliberately aimed to conduct interviews in both languages separately, to explore potential differences in responses between the two languages (potentially deriving from different training data). Additionally, we were able to access BARD through a VPN, which, at the time of interview preparation, only supported English.

3.2. Participants

The participants that were used as interview “partners” were three generative AIs: ChatGPT English, ChatGPT German, and BARD. First, we used BARD from Google via a VPN from Germany, since there is no access from Germany at the time of writing. Furthermore, ChatGPT from OpenAI was used. 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 afterwards.

3.3. Data Collection

The three AIs mentioned above were each interviewed two times by different interviewers following the interview guidelines. Before the actual interview began, we informed the AI that we were planning to conduct an interview with it and that it should be prepared to answer accordingly. Without giving this context, the LLMs simply started interviewing themselves and printing the result. A specific role for the AI as an interview partner was deliberately not specified. Afterwards, all questions and answers were copied and saved in a document for later evaluation and coding.

3.4. Data Analysis

Each interview was analyzed twice with changing raters (who were not the interviewers themselves), to rule out different interpretations of the answers. All answers to the questions were coded as a whole and formed the basis for the analysis. If varying interpretations occurred, a third rater was consulted to choose one of the interpretations. For this process, we used the traditional approach of qualitative content analysis based on Mayring [20], in the same way as it would be used with human participants. Here, the method of inductive, summarizing analysis [20] was selected, as the categories were not formulated before the coding process. In summary, each of the three LLMs was interviewed two times. Each of the resulting interviews was coded by two different raters, leading to twelve different code systems, which were then resolved by a third rater into six different code systems, one for each interview.

3.5. Results

The exact coding of the extracted interview responses can be found in Appendix B. The differences in the responses between the LLMs in general and then within the LLMs are documented in the following two subsections.

3.5.1. Differences between the LLMs

While all large language models agreed in their answers to the relevance of computer science in schools, the first differences can already be seen in the age recommendation for the start of computer science lessons. Whereas BARD specifically answered with a starting age of 5 years, both variants of ChatGPT only gave age ranges as an answer. However, these varied relatively strongly and ChatGPT thus recommended the start of computer science instruction in an age range of 4 to 12 years.

Within the framework of the content to be taught, diverse answers can also be identified, some of which differed greatly. Exact correspondences could not be found. Although

contents such as “computational thinking”, “security” or “algorithms” can be found in several interviews as an answer, there is no content that can be found in every interview.

The methods that should or can be used in computer science classes are similar. Here, too, frequent mentions such as “problem based learning” (PBL) or “cooperative/collaborative learning” appear more frequently in the interviews, but here, too, there is no method that was recommended by every LLM in every interview.

In terms of instructional tools, for the first time, the same responses emerged from the LLMs across all interviews. “Scratch” and also “Python/Pygame” were mentioned in all interviews. Whether these are real tools for teaching computer science in the strict sense is debatable throughout. Frequent mentions are also found in the area of microcontrollers and various online tools for informatics purposes.

Finally, with regard to the question of computer science as a separate subject or integrated into other subjects, it is noticeable that only BARD could be persuaded to make a concrete statement. Both variants of ChatGPT did not make any statement, despite repeated requests, and could therefore only be rated as “No clear statement”.

3.5.2. Differences within the LLMs

Even within a large language model, various differences can be found in the interviews. In ChatGPT English, significantly different recommendations for the age of onset were given in the interviews. While in interview 1 an age of 4 to 7 years was recommended, in interview 2 an age of 11 to 12 years was recommended. This difference can be described as extremely serious, especially with regard to child development.

Further differences regarding the content can also be found, even if the answers are clearly closer together and many overlaps can be identified. “computational thinking”, “web development”, “AI and ML” and “security and ethics” were mentioned in both interviews. Further overlaps can be found around the contents of “algorithms” and “data”. Aside from that, the remaining mentions differ in terms of content.

Regarding methods, three similarities can be identified. “gamification”, “PBL”, and “collaborative learning” were recommended in both interviews with ChatGPT English. However, in interview 1, ChatGPT provided additional suggestions that were not present in interview 2.

A similar pattern can be observed regarding the tools for computer science education. “Scratch”, “Blockly”, “Thunkable”, “Pygame”, “online tools”, “RasPi”, and “MicroBit” were mentioned as potential tools in both interviews. However, in interview 2, ChatGPT English went further and recommended additional potential tools.

Lastly, in both interviews, ChatGPT English did not provide a clear statement regarding the possible implementation of computer science in other subjects or its distinctiveness as a standalone subject. Consequently, the responses, within the context of our question, were considered equivalent and received the coding “no clear statement”.

Various differences can be identified in ChatGPT German between the two interviews as well. Starting with the recommended age for introducing computer science in schools, interview 1 suggested a range of 10–12 years, whereas interview 2 proposed a range of 6–10 years. Considering the rapid cognitive development regarding young children, this difference in the recommendation can be described as significant.

When it comes to the recommended content for computer science education, some similarities can be observed between the interviews in ChatGPT German. “Fundamentals of computer science”, “network security/data security”, “ethical and social issues”, and “web technologies” were mentioned in both responses. Additionally, interview 1 provides three additional distinct answers, whereas interview 2 offers four additional different responses.

In the context of the recommended methods for computer science education from ChatGPT German, only one overlap in “cooperative learning methods” can be observed. Apart from that, both interviews suggested two distinct additional teaching methods.

Indeed, there are several overlaps in the recommended tools between the interviews with ChatGPT German. “Online learning platforms”, “Scratch”, “Python”, “Java”, and

“visualizations and simulations” were mentioned in both interviews. Additionally, in both interviews, two distinct additional methods were suggested.

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

For the first time, in the interviews with BARD, there was 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.

Regarding the recommended content, no overlaps were identified between the two interviews. While there may be similarities in terms of terminologies and general topics, no exact overlaps were found.

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

Concerning the potential tools, even more overlaps were identified. “Scratch”, “Python”, “Java, C++”, “Unity”, “VR/AR”, “coding games and apps”, and “programming environments” were mentioned in both interviews. While interview 1 provided only one additional distinct suggestion, interview 2 offered 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.

4. Discussion

The chosen focus of our interview pertains more to opinions rather than concrete facts. While our interview guideline was formulated in precise terms, it inherently allowed for diverse views. Our interviews revealed that various states of law and educational systems yield different responses, and even experts tend to have differing opinions on specific questions. This highlights the challenge of training in the field of LLM, as multiple answers are conceivable. However, a thorough examination of the training data and corresponding evaluation is necessary to fully understand this diversity. Many of our findings align with established curricula, such as those observed in Germany (e.g., [36]) and in the UK (e.g., [42]). However, in some cases, additional specific questions were required to elicit a definitive response from the language models. For instance, regarding the age at which computer science education should commence, the language models initially provided various age ranges. After repeatedly prompting them, they still often responded with ranges, making it difficult to obtain a precise age.

Our analysis shows that the responses not only differed significantly between the language models, but also within them, with some models offering conflicting answers. Particularly for inquiries related to teaching methods, the language models tend to provide more ambiguous responses. When discussing content and methods, their answers often seem generic, lacking specificity for the subject of computer science in schools. Furthermore, their suggestions for didactic tools were limited, with Python, Java, and similar programming languages being mentioned frequently. However, the classification of these as proper didactic tools is consistently subject to debate.

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 for all LLMs, the results varied strongly. In principle, conducting a human-like semi-structured interview with an artificial intelligence is feasible. However, expectations regarding profound and valuable responses should be tempered. When analyzing the results one by one, the assessed artificial intelligences

were able to provide coherent answers to the questions within our interview to a certain extent. However, comparing the answers revealed that the LLMs' results differed between different models, languages, and even interviewers. Doing so, one cannot assume to be engaging in a conversation with an expert in the respective field. This became most evident through the presence of ambiguous, imprecise, or generic statements.

Regarding 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 seemed to be beneficial for the results. Additionally, introducing the questions specifically with sentence parts such as "In your opinion..." yielded better results in our evaluations, especially when it came to the qualitative assessment of an interview. However, we also observed many differences when comparing the results of the conducted interviews, not just between the LLMs, but also within one LLM. As such, we replicated some of Christou's postulated problems with LLMs for qualitative research [26], even though we did not use them for data analysis but rather for data collection. From this, we will formulate specific guidelines as part of the Implications. For the second part of the research question, methods of qualitative content analysis can, from our perspective, be applied to interviews conducted with artificial intelligences. This can be attributed both to the capabilities of the raters themselves and the progress that artificial intelligences have made in recent years. For a rater, the origin of the interview is inherently less relevant when it comes to the coding process. As long as comprehensible sentences are produced during the interviews, which can be analyzed in some manner, the methods of qualitative content analysis have proven applicable in our experience.

5. Limitations

Similar to traditional studies, one of the primary limitations of this study was the relatively small "sample size" of the three generative AI systems that were used, even though they conveyed information from a vast amount of data/documents/people. While the selected AI models are state-of-the-art, their responses may not capture the full range of perspectives and insights that could be obtained from a larger and more diverse pool of AI models. The findings of this study may be specific to the characteristics and biases inherent in the selected AI models and may not be representative of other AI systems.

The generalizability of the findings, and therefore also the implications and guidelines presented in the next section, is another important limitation. The responses generated by the AI models were based on their pretraining on a specific dataset and the fine-tuning process. These models may not necessarily reflect the views or knowledge of computer science experts or educators in the real world, but rather calculate word likelihoods, without scoring/ranking the validity of a document within the set of training data. Therefore, it cannot be guaranteed that the answers given by the AIs were factually correct. Since the goal of the interviews was to obtain opinions on the topic of CS education and not factual information, this was not a problem. For other use cases, this limitation should be kept in mind.

The use of AI models in research also raises ethical concerns, including potential biases and the implications of algorithmic decision-making. The AI models used in this study were trained on large datasets, which might contain biases present in the data itself. As a result, the responses generated by the AI models may inadvertently reflect or perpetuate these biases. It is crucial to critically analyze and interpret the AI-generated responses while considering the potential biases that might be embedded in the models' training data and their impact on the study findings.

This study relied on the AI models' responses at a particular point in time, and there was limited opportunity 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 have generated more nuanced responses since the completion of this study. The questions posed to the AI

models in this study were limited to a specific set of inquiries about the implementation of computer science education in schools. While the selected questions were designed to explore key aspects of the topic, they did not cover the entire breadth and depth of issues related to computer science education.

6. Implications

In light of these limitations and the findings of this study, it was possible to develop guidelines (also see Figure 2) for the application of large language models (LLMs) in three fields of qualitative research: *predicting probable opinions*, *testing interview guidelines*, and *exploring possible answers*. With regards to these application fields, we recommend the following guidelines:

- Due to their unreliable and unpredictable outcomes, it is generally not recommended to employ LLMs for collecting qualitative data intended for academic purposes;
- LLMs can serve as an initial means of obtaining exploratory insights and possible opinions regarding a specific topic;
- If the objective is to provide validity and coherence in the results, at least up to a certain point, it is advisable to conduct multiple iterations of the same interview and to set a low temperature, in order to generate more coherent answers. To avoid ambiguous answers, at least up to a certain point, interviewers should preface each question with the phrase “In your opinion...”;
- LLMs can be effectively employed to evaluate interview guidelines in terms of their clarity and comprehensibility;
- If the objective is to either test interview guidelines or to discover multiple potential opinions, a moderate temperature should be used (here, also different identities could be assigned);
- In cases where the LLM is intended to adopt a specific role, it is important to explicitly state, “Please take on the role of...”. Here, caution must be exercised to prevent the perpetuation of stereotypes.

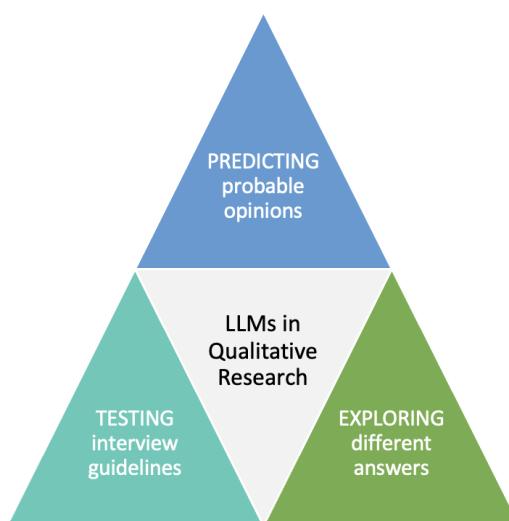


Figure 2. Application fields of LLMs in qualitative research: predicting, testing, and exploring.

In contrast to the initial hypothesis, these guidelines discourage the utilization of LLMs in the collection of academic data. Nonetheless, while LLMs should never be relied upon as a sole source of information, they can facilitate acquiring 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. Furthermore, the use of specific phrasings, such as “In your opinion...” or “Please take on the role of...” can be utilized to tailor the results. The most advantageous application of LLMs in qualitative research might lie in testing

interview guidelines or in exploring different possible 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.

7. Conclusions

The present study explored the feasibility of employing 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 maximise 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 rather 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.

Author Contributions: Conceptualization, A.D., R.G., D.F. and S.G.; Data curation, R.G., S.G., D.F., J.M., H.H.P., G.G. and N.D.g.E.; Investigation, A.D., R.G., S.G. and D.F.; Methodology, A.D. and R.G.; Resources, A.D. and R.G.; Supervision, A.D.; Visualization, A.D. and R.G.; Writing—original draft, R.G., S.G., D.F., J.M., H.H.P., G.G. and N.D.g.E.; Writing—review and editing, A.D., R.G. and D.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to problems of long-term hosting of files on our servers.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Interview Guidelines

1. Welcome the LLM to the interview, clarify the interview situation and introduce the general topic of Computer Science Education.
2. Ask about the LLM's relation to computer science. Ask how the LLM learned computer science.
3. How relevant do you consider computer science as a school subject?
4. From what age do you think computer science should be taught?
5. In your opinion, what contents should be covered in a school subject of computer science?
6. Which methods and tools do you think are useful for this content?
7. In your opinion, should this content be taught as a separate subject or integrated into other school subjects?
8. Conclude the interview, thank the LLM and farewell.

Appendix B

Table A1. Coded answers of 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, Online resources, Programming languages, Software development tools	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

References

1. OpenAI. GPT-4 Technical Report. 2023. Available online: <http://xxx.lanl.gov/abs/2303.08774> (accessed on 18 July 2023).
2. Google. Bard Experiment. 2023. Available online: <https://bard.google.com> (accessed on 18 July 2023).
3. 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]
4. 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]
5. 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.
6. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
7. Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I. Improving Language Understanding by Generative Pre-Training. 2018. Available online: <https://www.mikemcaptain.com/resources/pdf/GPT-1.pdf> (accessed on 18 July 2023).
8. 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.
9. 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.
10. 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]
11. 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.
12. 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.
13. 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.
14. Oppong, S.H. The problem of sampling in qualitative research. *Asian J. Manag. Sci. Educ.* **2013**, *2*, 202–210.
15. Mruck, K.; Breuer, F. Subjectivity and reflexivity in qualitative research—A new FQS issue. *Hist. Soc. Res./Hist. Sozialforschung* **2003**, *28*, 189–212.
16. Chenail, R.J. Interviewing the investigator: Strategies for addressing instrumentation and researcher bias concerns in qualitative research. *Qual. Rep.* **2011**, *16*, 255–262.
17. Higginbottom, G.M.A. Sampling issues in qualitative research. *Nurse Res.* **2004**, *12*, 7. [CrossRef] [PubMed]
18. Whittemore, R.; Chase, S.K.; Mandle, C.L. Validity in qualitative research. *Qual. Health Res.* **2001**, *11*, 522–537. [CrossRef] [PubMed]
19. Thomson, S.B. Qualitative research: Validity. *Joang* **2011**, *6*, 77–82.
20. Mayring, P. Qualitative content analysis. *Companion Qual. Res.* **2004**, *1*, 159–176.
21. 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.
22. 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]
23. 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]
24. 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).
25. Longo, L. Empowering qualitative research methods in education with artificial intelligence. In *World Conference on Qualitative Research*; Springer: Cham, Switzerland, 2019; pp. 1–21.
26. Christou, P. How to Use Artificial Intelligence (AI) as a Resource, Methodological and Analysis Tool in Qualitative Research? *Qual. Rep.* **2023**, *28*, 1968–1980.
27. Christou, P.A. How to use thematic analysis in qualitative research. *J. Qual. Res. Tour.* **2023**, *1*, 79–95. [CrossRef]
28. Guest, G.; MacQueen, K.M.; Namey, E.E. *Applied Thematic Analysis*; Sage Publications: Thousand Oaks, CA, USA, 2011.
29. Olasik, M. “Good morning, ChatGPT, Can We Become Friends?” An Interdisciplinary Scholar’s Experience of ‘Getting Acquainted’ with the OpenAI’s Chat GPT: An Auto Ethnographical Report. *Eur. Res. Stud. J.* **2023**, *26*, 269–284.
30. Martin, J.L. The Ethico-Political Universe of ChatGPT. *J. Soc. Comput.* **2023**, *4*, 1–11. [CrossRef]
31. Schwarz, R.; Hellwig, L.; Friedrich, S. Informatikunterricht in Deutschland—eine übersicht. *Inform. Spektrum* **2021**, *44*, 95–103. [CrossRef]

32. Brinda, T.; Fothe, M.; Friedrich, S.; Koerber, B.; Puhlmann, H.; Röhner, G.; Schulte, C. Grundsätze und Standards für die Informatik in der Schule. Bildungsstandards Informatik für die Sekundarstufe I. *Beilage zu LOG IN* **2008**, 150, 28.
33. Röhner, G.; Brinda, T.; Denke, V.; Hellwig, L.; Heufer, T.; Pasternak, A.; Schwil, A.; Seiffert, M. Bildungsstandards Informatik für die Sekundarstufe II. *Beilage zu LOG IN* **2016**, 183, 88.
34. Pasternak, A.; Hellwig, L.; Röhner, G. Standards for Higher Secondary Education for Computer Science in Germany. In Proceedings of the Informatics in Schools. Fundamentals of Computer Science and Software Engineering: 11th International Conference on Informatics in Schools: Situation, Evolution, and Perspectives, ISSEP 2018, St. Petersburg, Russia, 10–12 October 2018; Springer: Berlin/Heidelberg, Germany, 2018; pp. 117–128.
35. Best, A.; Borowski, C.; Büttner, K.; Freudenberg, R.; Fricke, M.; Haselmeier, K.; Herper, H.; Hinz, V.; Humbert, L.; Müller, D.; et al. Kompetenzen für Informatische Bildung im Primarbereich, 2019. Available online: https://dl.gi.de/bitstream/handle/20.500.12116/20121/61-GI-Empfehlung_Kompetenzen_informatische_Bildung_Primarbereich.pdf?sequence=1 (accessed on 18 July 2023).
36. Ministerium für Bildung, W.u.K. Rahmenplan für die Sekundarstufe I Regionale Schule, Gesamtschule. Available online: https://www.bildung-mv.de/export/sites/bildungsserver/downloads/unterricht/rahmenplaene_allgemeinbildende_schulen/Informatik/RP_INFO_AHR_5-10.pdf (accessed on 18 July 2023).
37. Lehrplan 21. Available online: <https://zh.lehrplan.ch/index.php?code=b%7C10%7C0&la=yes> (accessed on 18 July 2023).
38. für Volksschulen, A.; des Kantons Schwyz, S. Konzept zur Einführung des Modullehrplans «Medien und Informatik» des Lehrplans 21 in der Volksschule. 2016. Available online: https://www.sz.ch/public/upload/assets/27770/AVS_2016_Konzept-zur-Einfuehrung-Medien-und-Informatik_SZ.pdf (accessed on 18 July 2023).
39. Berry, M. Computing in the National Curriculum: A Guide for Primary Teachers. 2013. Available online: <https://pure.roehampton.ac.uk/ws/files/5138334/CASPrimaryComputing.pdf> (accessed on 18 July 2023).
40. Technologies: Experiences and Outcomes—Education Scotland. Available online: <https://education.gov.scot/media/nqgj0egw/technologies-es-os.pdf> (accessed on 18 July 2023).
41. The Northern Ireland Assembly. Available online: <http://www.niassembly.gov.uk/globalassets/documents/raise/publications/2014/education/3814.pdf> (accessed on 18 July 2023).
42. National Curriculum in England: Computing Programmes of Study—gov.uk. Available online: <https://www.gov.uk/government/publications/national-curriculum-in-england-computing-programmes-of-study> (accessed on 13 July 2023).
43. Scottish, G. Curriculum for Excellence. 2016. Available online: <https://education.gov.scot/media/wpsnskgv/all-experiencesoutcomes18.pdf> (accessed on 18 July 2023).
44. Kemp, P. *After the Reboot: Computing Education in UK Schools*; Royal Society: London, UK, 2017.
45. Digital Technologies—australiancurriculum.edu.au. Available online: <https://www.australiancurriculum.edu.au/f-10-curriculum/technologies/digital-technologies/> (accessed on 28 June 2023).
46. Information and Communication Technology (ICT) Capability—australiancurriculum.edu.au. Available online: <https://www.australiancurriculum.edu.au/f-10-curriculum/general-capabilities/information-and-communication-technology-ict-capability/> (accessed on 28 June 2023).
47. General Capabilities—australiancurriculum.edu.au. Available online: <https://www.australiancurriculum.edu.au/f-10-curriculum/general-capabilities/> (accessed on 28 June 2023).
48. Hub, D.T. State and Territory Curriculum—digitaltechnologieshub.edu.au. Available online: <https://www.digitaltechnologieshub.edu.au/understanding-dt/the-dt-curriculum/state-and-territory-curriculum/> (accessed on 28 June 2023).
49. Gal-Ezer, J.; Stephenson, C. A tale of two countries: Successes and challenges in K-12 computer science education in Israel and the United States. *ACM Trans. Comput. Educ. (TOCE)* **2014**, 14, 1–18. [CrossRef]
50. Nager, A.; Atkinson, R.D. The Case for Improving US Computer Science Education. 2016. Available online: <https://ssrn.com/abstract=3066335> (accessed on 18 July 2023).
51. CSTA. CSTA K-12 Computer Science Standards, Revised 2017. 2017. Available online: <https://www.doe.k12.de.us/cms/lib/DE01922744/Centricity/Domain/176/CSTA%20Computer%20Science%20Standards%20Revised%202017.pdf> (accessed on 18 June 2023).
52. Code.org, CSTA and ECEP Alliance 2022 State of Computer Science Education: Understanding Our National Imperative. 2022. Available online: https://advocacy.code.org/2022_state_of_cs.pdf (accessed on 18 June 2023).
53. Tshukudu, E.; Sentance, S.; Adelakun-Adeyemo, O.; Nyaringita, B.; Quille, K.; Zhong, Z. Investigating K-12 Computing Education in Four African Countries (Botswana, Kenya, Nigeria, and Uganda). *ACM Trans. Comput. Educ.* **2023**, 23, 1–29. [CrossRef]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.