Research ISSN 2525-8222
DOI: http://dx.doi.org/10.33361/RPQ.2024.v.12.n.30.724

AN ANALYSIS OF THE GENERATIVE AI USE AS ANALYST IN QUALITATIVE RESEARCH IN SCIENCE EDUCATION

UMA ANÁLISE DO USO DE IA GENERATIVA COMO INVESTIGADORA EM PESQUISAS QUALITATIVAS EM EDUCAÇÃO CIENTÍFICA

Juliana Rodrigues dos Anjos[1]
Maira Giovana de Souza[2]
Agostinho Serrano de Andrade Neto[3]
Bruno Campello de Souza[4]

Abstract: The article evaluates the effectiveness of generative artificial intelligence models, specifically ChatGPT 4.0 and Claude 2.0, in conducting qualitative research within the field of scientific education. By applying the Cognitive Networks Mediation Theory (CNMT) to analyze interviews from two students, it was found that Claude 2.0 surpassed ChatGPT 4.0 in recognizing cognitive mediations and distinguishing between pre- and post-test conditions. Although both models concurred on the concept of conceptual evolution, Claude 2.0 demonstrated a greater capacity for detail, notably by referencing specific interview excerpts to support its analyses upon request. In contrast, ChatGPT 4.0 exhibited difficulties in these areas, even when given additional prompts. The study concludes by acknowledging the utility of AI, particularly Claude 2.0, for qualitative research, while also emphasizing the critical role of human oversight in detecting potential deficiencies within these models.

Keywords: Qualitative analysis; Science Education; CNMT; ChatGPT; Claude.

Resumo: O artigo avalia a eficácia de modelos de inteligência artificial generativa, especificamente ChatGPT 4.0 e Claude 2.0, na realização de pesquisas qualitativas na área de educação científica. Ao aplicar a Teoria da Mediação das Redes Cognitivas (CNMT) para analisar entrevistas de dois estudantes, constatou-se que Claude 2.0 superou o ChatGPT 4.0 na capacidade de reconhecer mediações cognitivas e distinguir entre condições de pré e pós-teste. Embora ambos os modelos tenham concordado sobre o conceito de evolução conceitual, Claude 2.0 demonstrou maior capacidade de detalhamento, notavelmente ao referenciar trechos específicos das entrevistas para apoiar suas análises quando solicitado. Em contraste, ChatGPT 4.0 exibiu dificuldades nessas áreas, mesmo quando fornecido com prompts adicionais. O estudo conclui reconhecendo a utilidade da IA, particularmente de Claude 2.0, para pesquisas qualitativas, enquanto também enfatiza o papel crítico da supervisão humana na detecção de potenciais deficiências dentro desses modelos.

Palavras-chave: Análise qualitativa; Educação científica; TMC; ChatGPT; Claude.

---

[1] Master's Degree in Science and Mathematics Education from the Lutheran University of Brazil (ULBRA), Post-Graduate Program in Science and Mathematics Teaching (PPGECIM) at ULBRA, Farroupilha Avenue, 8001, Canoas, Rio Grande do Sul, Brazil. Email: juranjos@rede.ulbra.br

[2] Master's Degree in Science and Mathematics Education from the Lutheran University of Brazil (ULBRA), Post-Graduate Program in Science and Mathematics Teaching (PPGECIM) at ULBRA, Farroupilha Avenue, 8001, Canoas, Rio Grande do Sul, Brazil. Email: maira.souza@rede.ulbra.br

[3] Ph.D. in Physics from the University of São Paulo (USP), Post-Graduate Program in Science and Mathematics Teaching (PPGECIM) at the Lutheran University of Brazil (ULBRA), Farroupilha Avenue, 8001, Canoas, Rio Grande do Sul, Brazil. Email: agostinho.serrano@ulbra.br

[4] Ph.D. in Cognitive Psychology from the Federal University of Pernambuco (UFPE), Department of Administrative Sciences (DCA) at UFPE, Recife Campus, Prof. Moraes Rego Avenue, 1235, Cidade Universitária, Recife, Pernambuco, Brazil. Email: bcampello@uol.com.br

ISSN 2525-8222
DOI: http://dx.doi.org/10.33361/RPQ.2024.v.12.n.30.724

1 Introduction

At the cusp of the 21st century's third decade, the increasing development of Information Technology (IT) has impacted the landscape of research methods in social sciences. This research focuses on the emergence and impact of Generative Artificial Intelligence (GenAI) platforms, such as OpenAI's ChatGPT (OPENAI, 2022) and Anthropic's Claude (ANTHROPIC, 2023), in the development of qualitative research. The incorporation of IT in education have already caused remarkable changes (Bruun; Duka, 2018) in society. Similarly, the implementation of GenAI in tasks usually human-centered is promising to impact different fields, such as qualitative data analysis.

In this context, this study aims to explore two GenAI chatbots can be used in qualitative studies alongside human analysts, specifically within a Physics education research analyzing students' conceptual understanding. Through an innovative theoretical approach, we compared the performance of ChatGPT 4.0 and Claude 2.0 in the analysis, aiming to comprehend how the GenAI can improve, challenge of change the qualitive analysis methods.

At the core of this exploration lies a central research question reminiscent of past academic endeavors (Belotto, 2018): How the GenAI can enhance the human analysis in the domain of qualitative research? Through this comparative analysis, we aim to explore the emerging role of GenAI in qualitative research, as well as contribute to the controversy regarding the joint use of technology and human outlook in academic research.

2 Bibliographic Review

Generative artificial intelligence (GenAI), particularly in the context of producing textual analysis, is an emerging field, especially within the realms of social sciences and humanities, which traditionally emphasize qualitative analysis. This trend began to gain momentum in 2023 with the popularization of ChatGPT. However, it is worth noting that computer software has long played a role in both quantitative and qualitative research (Santos; Santos; Boss, 2023). For the scope of this paper, a SCOPUS search was conducted using the keywords "ChatGPT" and "qualitative analysis". This yielded a total of 21 papers. Other databases, such as Google Scholar and ERIC, were excluded due to the return of diverse papers, including numerous preprints, opinion pieces, and editorials that had not undergone double-blind peer review. The criteria for inclusion and exclusion are as follows:

Table 1: Inclusion and exclusion criteria for the bibliographic search

Inclusion Criteria	Exclusion Criteria
- Full articles that mentioned in a searchable index the word ‘ChatGPT’ and ‘qualitative analysis’. - Articles that used the word ‘ChatGPT’ many times (5+) in the text, and actually dialogued with the context of LLMs.	- Articles that only mentioned ChatGPT casually (1-2 times) and the paper was actually about another subject. - Editorials, Opinions, Letters, etc. - Pre-prints (no peer review). - Papers where the focus was ML pre-ChatGPT. - Papers that not dwell on qualitative analysis.

Source: authors

A total of six papers, as of September 21st, 2023, were finally selected. Those papers and their contribution will be discussed in detail.

In his insightful paper, “A Critical Perspective Over Whether and How to Acknowledge the Use of Artificial Intelligence (AI) in Qualitative Studies” (Christou, 2023a), Prokopis Christou critically evaluates the contention surrounding the proper acknowledgment of AI in qualitative research. Notably, Christou (2023a) transparently acknowledges and documents the concurrent utilization of both AI and the researcher's cognitive abilities in his theoretical discussions.

Christou (2023a) underscores that technology, particularly artificial intelligence, has become a prevalent, albeit under-acknowledged, element in contemporary qualitative analysis and scientific research. This fact is evidenced by the prior use of technology for data interpretation, proposition development, and insight generation (Ching *et al*2018, as cited in Christou, 2023a). Generative AI systems, such as ChatGPT, though comparatively recent, are only new embodiments of existing practices across global research centers.

However, the integration of AI in qualitative research has not escaped criticism (Miloyan *et al*. 2019, as cited in Christou, 2023a). The central concern within the scholarly community, as highlighted by Christou, revolves around the potential omission or intentional disregard in referencing its use.

Unlike traditional neural networks or deep learning tools, where researchers maintain control over data processing methodologies, Generative AI operates with a degree of autonomy. While such tools can be prompted to undertake “deep qualitative analysis”, the specific analytical frame or bias is inherently determined by the system, drawing from its pre-existing datasets and human-augmented training.

Contrasting the controllable outputs of specialized neural networks with the more autonomous Generative AI highlights an essential distinction. While the former might be categorized as computational tools, the latter introduces an “additional voice” to the qualitative analysis process, bearing its own intrinsic biases and perceptions.

> Whether or not AI is recognized as a source of information or an analysis tool, it is increasingly being used by researchers. AI-generated or informed literature reviews, systematic reviews, conceptual papers, study approaches, and analyses (such as thematic analysis), have all been used increasingly in recent years (Christou, 2023a, p.5).

The integration of Generative AI tools in qualitative research has sparked considerable debate and critique from within the qualitative research community. Concerns regarding potential biases, inaccuracies—termed “hallucination” by the LLM community—and an inability to capture finer details are commonly cited. Nonetheless, as any analytical tool or person can present similar shortcomings, it underscores the importance of thorough verification. As Christou suggests, the primary researcher must not only be intimately familiar with the data but also rigorously cross-check both the input and output of any technological tool, Generative AI included.

The capability of AI to grasp the intricacies of theoretical frameworks, often foundational to qualitative empirical studies, has been questioned. There is skepticism regarding AI’s aptitude to encapsulate and apply intricate theoretical concepts accurately. However, our findings suggest this view warrants a more nuanced perspective. We demonstrate that when a Generative AI tool is informed of a specific, unfamiliar theoretical reference, it can astoundingly produce not only a competent interpretation for its qualitative application but also apply it in an analysis aligned with the informed reference.

Christou (2023a) reflects on the reticence among researchers concerning their disclosure about AI usage in academic publications, primarily stemming from apprehensions of skepticism from peers. This hesitance might lead many to discreetly incorporate AI tools, such as ChatGPT, in their work. Intriguingly, it is speculated that the utilization of ChatGPT within academic settings is more prevalent than publicly acknowledged. Barocas and Selbst (2016), as cited by Christou (2023a), unveiled a tendency among researchers to downplay or even hide AI use, possibly to evade critiques against mechanized formulation of their propositions or theoretical discussions. Such hesitance seems misplaced, as AI can potentially augment and enrich analytical depth, thus enhancing rather than diminishing the researcher's cognitive contributions.

CC BY NC ND

A separate ethical consideration arises concerning the attribution of credit to AI tools. While academic integrity demands acknowledgment of others' contributions, such recognition does not conventionally extend to tools, AI or otherwise. Yet, to maintain scientific rigor, transparency about methodologies and tools employed is paramount. In conclusion, Christou (2023a) strongly advocates for comprehensive disclosure regarding AI use in qualitative research, be it for bibliographic reviews, text generation/correction, or in-depth qualitative or theoretical analysis.

Atkinson (2023) critically examines the integration of Artificial Intelligence (AI) and Machine Learning Techniques (MLT) specifically within the realm of systematic literature review (SLR). Evidencing the transformative impact of AI on diverse research methodologies, Atkinson, referencing Longo (2020), underscores the surge of data available to today's scholars. Contrary to preliminary assumptions of AI's utility being confined to reference retrieval from databases, Atkinson's discourse extends further. He emphasizes that AI, harnessing the capabilities of MLT, aids not just in data extraction but crucially in abstracting and interpreting the essence of scholarly content.

This nuance posited by Atkinson becomes particularly salient in the context of our exploration of Generative AI in qualitative research. While Atkinson's subsequent deliberations delve deep into qualitative research protocols, one focal point merits attention: the innovative usage of ChatGPT. This tool, once prompted with an SLR's guiding research question, suggests potential search terminologies and strategies. ChatGPT 4.0 users can input and process a diverse array of file formats, though with character constraints that vary based on the file type.

The limits encountered, especially concerning ChatGPT's reading capacity, can be attributed to the "context window"—the model's capacity to simultaneously consider a specific quantity of textual tokens for generating responses. GPT-3, for instance, operates within a 2048-token window, approximately equal to 450-500 words. ChatGPT 4.0, on the other hand, boasts an 8192-token capacity. Notably, the Claude 2.0 model by Anthropic can comprehend an astounding 100,000 tokens or approximately 20,000 words. In this study, we harness both ChatGPT 4.0 and Claude 2.0 for qualitative analyses.

Hamilton *et al*. (2023) embark on a discerning examination of AI's role, particularly using ChatGPT, in augmenting qualitative analysis—a focal theme for our special call for papers. At its core, the authors employ a pre-processed dataset, optimized for ChatGPT (version 3.5), to elicit analysis in response to expertly crafted prompts. Their methodology encompassed coding qualitative interviews from 71 guaranteed income pilot recipients, and subsequently comparing ChatGPT's coding results with those executed by humans.

However, their endeavors are grounded in a lineage of work highlighting AI's burgeoning role in research. Both Tabone and Winter (2023) and Mesec (2023) have illuminated AI's efficacy in textual data analysis. Yet, as Hamilton et al. (2023) articulate, their primary concern is not the raw interview content per se but rather the comparative methodology between human and AI-generated outputs.

Their data pre-processing involved a sequence of steps: transcription via Rev.com, anonymization, upload to a qualitative coding platform, and subsequent manual extraction of 1125 pivotal statements. These were then fed into ChatGPT, considering the constraints of character limits for input prompts and the broader "context window", which varies depending on the underlying AI architecture. Their operational strategy encompassed feeding ChatGPT in tranches—50 statements per prompt, over 22 or 23 sessions. Yet, a glaring oversight emerges in their methodology; the authors seem to disregard the possibility of exceeding ChatGPT 3.5's 4096-token context window, potentially compromising the holistic data interpretation they sought.

Hamilton *et al*. (2023) compared the coding capabilities of human analysts and AI, specifically ChatGPT. Both identified five overlapping themes in their analysis. Additionally, each method discovered unique themes not identified by the other: human analysts identified six unique themes, while ChatGPT identified five. The human analysis exhibited depth, nuance, and flexibility that were notably lacking in the AI's interpretation. This human insight, grounded in contextual understanding and expertise, added significant depth to the interpretation, leading to a more comprehensive understanding of the data.

However, a word of caution is warranted regarding the study's conclusions. Given the possibility that the researchers might have inadvertently exceeded ChatGPT 3.5's context window, the AI's resultant analysis could be perceived as lacking depth, nuance, and context. Our forthcoming results section offers a divergent perspective on this issue.

In a recent paper, Chubb (2023) utilizes ChatGPT for qualitative research, marking this as the second investigation of its kind. Chubb highlights the myriad academic applications of ChatGPT, noting its use:

> [...] in rapid document writing and translation (TATE et al., 2023), qualitative data analysis (TABONE; WINTER, 2023), artistic composition (KIRMANI, 2022), climate change research (BISWAS, 2023), text assessment (ZHONG et

> al., 2023), potential in scholarly peer review (HOSSEINI; HORBACH, 2023), and programming language decoding (SOBANIA et al., 2023), among others. (Chubb, 2023, p. 2)

Interestingly, Chubb opted to use ChatPDF over ChatGPT. While she does not specify her reasoning, it can be inferred that the inability of ChatGPT 3.5 to handle PDFs at the time prompted a shift to tools like ChatPDF for broader data processing— whether for academia, business, or personal use.

ChatPDF, a third-party platform, facilitates the analysis of single PDFs using ChatGPT 3.5. This tool, while claiming to utilize GPT-3.5’s API provided by OpenAI, faces limitations in viewing entire PDFs. The site’s FAQ explains that ChatPDF views only select paragraphs, attempting to sidestep the context window constraint of 4096 tokens. By indexing paragraphs (through an undisclosed method) and targeting those deemed “relevant”, it offers a potentially compromised analytical scope. This method’s efficacy is not universally guaranteed, requiring careful consideration of its outcomes, as also indicated in Hamilton (2023).

Chubb’s work delineates her threefold contribution: elucidating her analytical framework using AI, demonstrating ChatPDF’s capacity to transform interview transcripts into narrative vignettes, and shedding light on the merits and drawbacks of AI in qualitative research. Central to her approach is a thorough understanding of raw data, echoing sentiments by Christou (2023a) and Hamilton *et al*. (2023).

Chubb’s process involves direct PDF uploads to ChatPDF, distinguishing her from Hamilton *et al*. (2023) who pre-processed data due to input constraints. A key distinction is Chubb’s emphasis on prompt engineering— tailoring prompts to optimize results from the LLM. Through iterative refinement, Chubb discussed the importance of prompt choice in her work. Intriguingly, some papers offer diverse and even unexpected prompts, exemplified by Yang et al.’s (2023) whimsical suggestion: “take a deep breath and...”. Further, she suggests that ChatGPT itself could assist in crafting the ideal prompt.

Post-output, Chubb fine-tunes the resulting vignettes and evaluates their accuracy, addressing the known “hallucination” issue of LLMs. A sample of AI-rendered vignettes was verified by participants, affirming their accurate representation. Chubb (2023) extensively comments on ChatGPT’s responses to AI-assisted qualitative analysis and the tools appropriate for it. She underscores the necessity of prompt engineering by contrasting different prompt outcomes. Moreover, the topic of AI bias, which is a

CC BY NC ND

ISSN 2525-8222
DOI: http://dx.doi.org/10.33361/RPQ.2024.v.12.n.30.724

perennial concern for many researchers, is discussed. Citing Bozkurt et al. (2023), she mentions the potential of LLMs to redress historical and geographical injustices.

Major AI stakeholders are actively measuring and mitigating biases in their LLMs, refining them through RLHF (Reinforcement Learning from Human Feedback). However, LLM outputs warrant scrutiny for both accuracy and potential bias. Concluding her paper, Chubb emphasizes that LLMs serve as tools augmenting researchers, not substituting them, especially in qualitative analysis.

Christou (2023b) delves into the contemporary utilization of AI in research, focusing on both its advantages and limitations, while anchoring the discussion in theoretical critical principles. Central to this is the increasing employment of Generative Pre-trained Transformers (GPTs), a subset of deep learning models, by the qualitative research fraternity. These models serve diverse purposes, ranging from the production and summarization of information to its translation and analysis (Conneau; Lample, 2019; Lund; Wang, 2023; Lund *et al*. 2023 apud Christou, 2023b).

A pivotal point in Christou's work is the ontological exploration of deep learning models, particularly GPTs. These models are identified as a distinct AI variant cohabiting with other AI forms. Given their training on vast quantities of human-generated content, they can be ontologically situated within a broader system of human-machine interactions. When GPTs emulate human responses to textual prompts, they act both as language usage outcomes and as tools enhancing human understanding. Through such a lens, qualitative researchers can comprehend the potentials and pitfalls of GPTs more intimately.

The author also underscores the pitfalls associated with the "hallucinations" of LLMs, stemming from inaccurate or misconceived training data. Such models, while advanced, might overlook the intricacies of academic prose or struggle to discern reliable sources from unreliable ones. Christou (2023b) thus advocates for prompt engineering, which steers LLM outputs in alignment with expert knowledge. Furthermore, he suggests researchers furnish LLMs with theoretical contexts essential for qualitative analyses, always mindful of context window constraints.

Additionally, Christou (2023b) champions the triangulation method, a staple in certain qualitative research paradigms, advocating the cross-referencing of AI-generated outputs to bolster consistency and validity. In summation, Christou (2023b) paints a largely optimistic picture of AI's role in qualitative research, provided researchers are astutely aware of potential hallucinations, inaccuracies, ethical quandaries, and biases. A researcher’s unique perspective, he argues, should not remain implicit but be actively articulated, especially when guiding an LLM using prompt engineering techniques.

In another noteworthy contribution, Fonseca, Chimenti, and Suarez (2023) underline the potential of deep learning language models, particularly in the realm of social sciences in developing nations. These models, akin to the BERT and GPT structures, not only empower researchers to interpret and elucidate phenomena in previously inconceivable ways but also bring operational benefits to theory-generating research. Such models offer practical solutions in reducing the costs tied to textual content coding. Furthermore, they facilitate the efficient and scalable management of vast, diverse data garnered from a range of sources.

Of significant emphasis is the opportunity these computational methodologies present for qualitative research in the Global South. They enable researchers to harness datasets from developed regions, thereby achieving a depth and breadth of analysis often reserved for higher-budget settings. As a result, the integration of these computational tools may help bridge the disparity in resources and capabilities within the global qualitative social science research arena.

3 Theoretical Referential used for the analysis

In this section we are going to discuss the theoretical referential used by the interpreters, be them the human or the AI interpreters to assess the students’ interviews.

Drawing from some of the references used in the previous sections, we will ask the chatbot Claude 2.0 to essentially explain how we could use a novel theoretical referential to analyze learning, or conceptual evolution, a customary target of qualitative research in education. First, we made sure that both ChatGPT 4.0 and Claude 2.0 did not previously know the referential, and they both answered negatively when questioned if it was present in their training data. Second, we double checked by asking if they could explain it and they used different authors with similar names, denying to know the authors of the referential we used, the Cognitive Networks Mediation Theory (Souza *et al*. 2012).

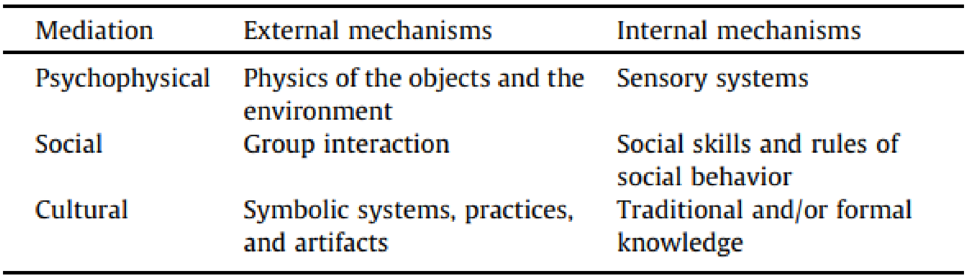
After being sure the referential was not present in their training data, we sent a PDF file with the referential (Souza *et al*. 2012). This paper is not going to explain the details of the referential, as it was only mentioned here to subside the discussion of whether the chatbots could use a novel referential to perform a qualitative analysis. Instead, the following prompt was sent to Claude 2.0, with the PDF file:

> This is a paper discussing a novel theoretical referential, CNMT, cognitive networks mediation theory. Read the paper, act as a qualitative researcher and explain how one should use it to interpret or code a student interview, grounded in this referential. What points should we look for in the interview if the analysis of the student understanding should be done grounded in this referential? (Authors, 2023).

And Claude 2.0’s response:

> Here are some key points to look for when coding a student interview through the lens of the Cognitive Mediation Networks Theory (CMNT):
> - Look for **evidence of the student using external tools or resources to enhance their thinking and learning.** This could include things like taking notes, using a calculator, looking something up online, etc. The CMNT views these kinds of interactions with external structures as "cognitive mediation."
> - Pay attention to how adept the student seems to be at **integrating internal mental processes with the external tools they are using.** According to the CMNT, individuals develop “internal mechanisms” that allow them to effectively deploy external tools for cognition. Look for signs that the student has strong internal mechanisms for the external tools they use.
> - Note the **types of external cognitive mediation the student relies on. Are they more basic physical tools like pencil and paper? Or are they more complex digital technologies like smartphones and internet searches?** The CMNT proposes that digital technologies constitute a more advanced form of mediation leading to enhanced cognition.
> - **Listen for the student talking about how they cognitively approach tasks and problems.** Does their thinking style align with the modes of thinking predicted to arise from digital immersion according to the CMNT? For example, visual-spatial thinking, accessing large information sets, using computer metaphors, etc.
> - **Pay attention to the level of the student's cognitive performance as evidenced in the interview**. The CMNT hypothesizes that greater utilization of digital cognitive mediation leads to gains in cognitive abilities. Look for signs that heavy usage of digital technology co-occurs with high cognitive achievement.
> In general, use the CMNT lens to understand **how the student deploys external cognitive tools, how adept they are at integrating internal and external processing, their degree of digital immersion, their cognitive skills and strategies, and their overall level of cognitive performance.** The goal is to code the interview for alignment with the CMNT's model of digitally mediated cognition (Authors, 2023)[5]

Table 3: Provided description of each mediation to be used in the analysis



Mediation	External mechanisms	Internal mechanisms
Psychophysical	Physics of the objects and the environment	Sensory systems
Social	Group interaction	Social skills and rules of social behavior
Cultural	Symbolic systems, practices, and artifacts	Traditional and/or formal knowledge

Source: Souza *et al.* (2012).

---

[5] In line with the recent discussion in the academic community, all responses given by the chatbots were considered authored **not** by the LLMs (as they are not co-authors, but tools), instead authored by the paper authors themselves.

ISSN 2525-8222
DOI: http://dx.doi.org/10.33361/RPQ.2024.v.12.n.30.724

In essence, that is what is expected of someone using CNMT to analyze an interview, showing that the chatbot can explain how the qualitative analysis should be done. For the results below, a clear chat window was used, and not a subsequent conversation based on this one. The only information provided to the LLMs regarding what are the mediations discussed in the theoretical referential is reproduced as it is in Table 3 above.

4 Methods

The research objective was to assess the ability of the language models ChatGPT 4.0 and Claude 2.0 to perform qualitative analyses and determine which of them is more effective considering the analyses conducted by human beings. Therefore, data from previous two studies, already analyzed by human beings, were selected to be also analyzed by the GenAI.

Before 2023, two studies were conducted to investigate the different mediations (psychophysical, social, cultural, and/or hypercultural) that facilitate the development of mental images in students when responding to questions related to the nature of light (Investigation A) and concepts of special relativity (Investigation B), both studies were conducted with high school students from the state education system.

The investigation A, regarding the nature of light, was conducted remotely during the COVID-19 Pandemic, with thirteen participants analyzed. Students engaged in simulations, presentations, videos, readings, and a mobile application for the analysis of emission and absorption spectra. On the other hand, Investigation B was conducted in-person, before the pandemic, with fourteen students. The students interacted with computer simulations, videos, presentations with gifs and animations, mock-ups, and developed group activities, focused on space contraction and time dilation phenomena.

At the conclusion of both interventions, semi-structured interviews were conducted with the students. The interviews were developed using the *Report Aloud* protocol (Trevisan *et al*. 2019), and all of them were fully recorded. The analysis of the interviews was carried out through recordings and video transcriptions. Each test conducted during data collection, including pre-tests and post-tests, was discussed during the interviews, where students were invited to express their conceptions and imaginations regarding these topics.

ISSN 2525-8222
DOI: http://dx.doi.org/10.33361/RPQ.2024.v.12.n.30.724

In the transcriptions, the names of the students were replaced with the letter "S" followed by a number, such as S11 (Student 11). From the interview transcriptions and gesture analysis, in a triangulation process, the evaluation of the mediations used and possible conceptual evolution regarding each research theme began. The approach aimed to identify the predominant mediations, considering the possibility of multiple mediations in explaining a single concept. It is important to note that both studies were conducted and analyzed before the introduction of Generative Artificial Intelligence.

In the analysis conducted for this study, students S11 (investigation A) and S18 (investigation B) were chosen due to demonstrate the greatest use of different mediations among other participants. This investigation was performed using the ChatGPT 4.0 language model with the "advanced data analysis" mode and Claude 2.0 for the examination of their interview data, predominant mediations, and conceptual evolution.

Student S11 exhibited a pronounced inclination towards employing hypercultural mediation in their responses during the interview. Moreover, it was observed that their conceptual development is closely tied to their interaction with hypercultural tools, for instance. In the same way, student S18 showed to use a lot the hypercultural mediation, however, using also in a significant way the social mediation. As well as S11, it was possible to note that S18's conceptual evolution was closely related to the interaction with these mediations.

Claude 2.0 and ChatGPT 4.0 were compared in how they conducted interviews analysis. To investigate this matter, the complete interviews transcriptions and the paper that underlies the research framework, CMNT (Souza *et al*. 2012), were submitted to the chatbots requesting analyses of both materials. The data from Study A were analyzed by Claude 2.0 on August 15th and by ChatGPT 4.0 on August 29th both in 2023. Concerning the Study B, the analysis using Claude 2.0 was conducted on September 19th and using ChatGPT 4.0 on September 30th also both in 2023[6].

It is worth noting that both models were selected because they supposedly can either hold all given information in a context window (Claude 2.0) or navigate the PDF files information (ChatGPT 4.0 with advanced data analysis). ChatPDF was not used in this study because, as we already discussed, it does not look at the entire information at once.

---

[6] The complete dialogue with ChatGPT 4.0 and Claude 2.0 for both interview analysis is available at: https://drive.google.com/file/d/1svbNt9zuuVJrooiFspdEqAaeXRN0RKXb/view?usp=drive_link

ISSN 2525-8222
DOI: http://dx.doi.org/10.33361/RPQ.2024.v.12.n.30.724

For the analysis of the results provided by the chatbots Claude 2.0 and ChatGPT 4.0, the data were organized into three distinct categories. These categories encompass: the first one dedicated to the analysis of mediations, the second one focusing on the analysis of conceptual evolution, and the third one addressing the source from which the chatbots obtained information, by analyzing the interviews. Each category has a primary prompt that was employed in the study as an analytical instrument used by the chatbots (Table 2). The same prompts were used for both approaches, with ChatGPT 4.0 and Claude 2.0.

Table 2: Prompts used for both LLMs (adapted for *investigation A/investigation B*)

CATEGORY	PROMPT
Mediations	In file *S11/S18*, there is an interview with a student, and in the Campello file, there is the concept of a theory addressing cognitive changes associated with the emergence of technologies. It posits that cognition can occur through different mediations, such as psychophysical (real-world objects), social (interaction between people), cultural (symbols and books), and hypercultural (technological tools such as simulations, games, slides, applications, etc.). The work involving this theory serves as a theoretical framework. I request that you first read the article (disregarding the results, focusing on the introduction) to grasp the theoretical framework. Subsequently, I would like you to conduct a deep qualitative analysis of the interview (which is in Portuguese) grounded in this theoretical framework. Please perform to the best of your ability, assuming the role of a full-fledged researcher in the fields of physics and cognitive sciences.
Conceptual evolution	Now I want you to tell me how good and solid, using *light/space* and time concepts, the student is, make an in-depth assessment of his post-instruction knowledge compared to pre-instruction. Tell me how it evolved:
Source	Where in his interview did you gather your analysis of the student's knowledge evolution?

Source: authors

The category related to mediations aims to investigate the various forms of mediation reported and employed by the students during their responses to questions related to the topic, both in the pre-test and post-test. On the other hand, the category that focuses on conceptual evolution involves the analysis of the knowledge acquired by students after instruction, in comparison to their prior knowledge. This approach allows researchers to understand the factors that contributed to an improvement in students' understanding. Finally, the category referred to as “sources” identifies specific points in

ISSN 2525-8222
DOI: http://dx.doi.org/10.33361/RPQ.2024.v.12.n.30.724

the interview that highlight the changes, or the absence thereof, in students' knowledge throughout the instructional process.

5 Results and discussion

In this section, we present the results obtained from the use of the two GenAI chatbots. The analysis categories and their respective results are separated for both Investigation A and Investigation B.

5.1 Category 1: Mediations

Firstly, Claude 2.0 conducted an analysis of the article and the interview transcript, grounding itself in the theoretical framework of the Cognitive Mediation Networks Theory (CMNT). Subsequently, it carries out its own analysis based on this theoretical foundation.

> **CLAUDE 2.0 (1.A):** In the pre-test, the student demonstrates difficulty in explaining the concepts of light, frequency, and wavelength, indicating little familiarity with these concepts, showing that his initial understanding was limited.
> In the pre-test, his explanations are primarily based on visual and sensory aspects. This seems to indicate psychophysical mediation, with little cultural, social, or hypercultural mediation.
> In the post-test, his explanations demonstrate a deeper understanding of phenomena, including concepts like the electromagnetic spectrum, frequency, wavelength, and absorption/emission. This suggests that hypercultural mediation occurred during the experiment, through the use of simulations and videos. These simulations appear to have been central to his new understanding of luminous phenomena, exemplifying hypercultural mediation.
> Overall, the qualitative analysis based on CMNT suggests that the student started with limited prior knowledge, based on perception and direct experience, evolving to a deeper understanding after hypercultural mediations, especially through the use of computer simulations. This process exemplifies CMNT (Authors, 2023).

The files containing both interviews also included information about the gestures performed by the student, along with their respective descriptions. This inclusion may have significantly contributed to the understanding of what the student was expressing, even considering that artificial intelligence still lacks the capability to directly analyze images (it can, however, analyze the *description* of the images).

The gestures produced by the students are properly identified in the interview document through markers '[#]' (using the symbol '#') letters that correlate with the mental representation of the gestures performed by each participant and the time in the video that the gesture was performed. For example, let us consider a scenario in which a student makes a gesture while discussing the Electromagnetic Wave. In the transcription, this gesture is encoded as '[#EW; 03:31]' and followed by a detailed visual description of the gesture made by the student – as "moves left hand as undulating like a wave".

The use, by the LLMs, of these gestural descriptions occurred spontaneously, without the need for explicit prompts or explanations in the interview file. In other words, the LLMs interpreted them as gestural descriptions and autonomously integrated in the analysis.

Overall, the artificial intelligence Claude 2.0 analysis accurately identified the student's use of mediations in both the pre-test and post-test. It is important to note that in the interview, this usage was not explicitly separated but only mentioned.

The only discrepancy between human-made analysis and Claude's analysis pertains to the *YouTube* videos. While Claude 2.0 classified the viewing of these videos as a form of hypercultural mediation, in our analysis, they were understood as a manifestation of cultural mediation, albeit still related to hyperculture. Interestingly, classifying *YouTube* videos as a cultural mediation was an evolving process and discussion in the research group throughout the years, and initially they were indeed classified as hypercultural tools, just like Claude 2.0 did.

It is important to highlight that Claude 2.0 already made reference to a 'deeper understanding of the phenomena', specifically regarding the post-test. An aspect identified by it, not mentioned in the interview.

Now let's compare with ChatGPT 4.0 analysis. The same prompt was used, with the same files, but ChatGPT 4.0 with advanced data analysis could not read the PDF file and actually see the entire interview (it even mentions that only one question is being read). That happened likely because, as we could see in a report of the file reading process, only 500 characters were initially read. Therefore, we switched to another format, a word "docx" file, and obtained better results with it. The files were essentially the same, just converted from PDF to WORD format. The prompt used was the same as the one used in Claude 2.0.

In an initial stage, ChatGPT 4.0 conducts an analysis of the theoretical framework, addressing the key aspects of the theoretical content at hand. Subsequently, it initiates its analysis of student S11. ChatGPT 4.0 describes the student's highlighted responses in both the post-test and the pre-test as a starting point for initiating the analysis.

> **ChatGPT 4.0 (1.A):** Definition of Light:
> PRE-TEST: S11 defines light simply, associating it with what is visible.

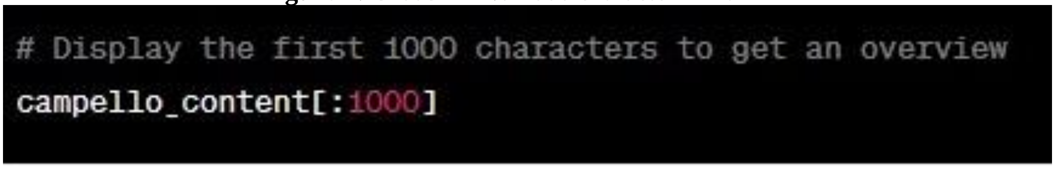
> POST-TEST: He now has a more complex understanding, describing light as electromagnetic waves and also mentioning its spectral composition.
> Object Colors:
> PRE-TEST: S11 believes an object's color is determined by how it reflects light.
> POST-TEST: (The full response has not yet been viewed, but he is expected to have expanded his understanding based on his response about light emission sources).
> In summary, the evolution of S11's responses suggests an enrichment in his understanding, which may be influenced by interactions with cognitive mediations as proposed by CMNT (Authors, 2023).

ChatGPT 4.0, despite making insightful inferences about cognitive development between pre-test and post-test, a pattern also identified in previously conducted non-AI research, it fails in its initial response to address possible mediations. The chatbot only mentions mediations when explicitly requested. It briefly mentions that hypercultural mediation assists in this conceptual evolution in a general sense but does not provide specific details or contexts. Furthermore, it is worth mentioning that ChatGPT 4.0 asked a question regarding the student's involvement with simulations or applications during the intervention, despite this information being explicitly stated in the interview, as the student frequently refers to the use of the computer.

The interview moment in which the student mentions the simulation: *"That simulation we did, where it emitted the sunlight on it, and then you could see the absorption causing the electron to move further from the electron cloud, and then to emit it would return to its original position, emitting light"* (S11).

Even changing from PDF to WORD format, it is worth mentioning that the ChatGPT 4.0's character limit of 1000 characters (for a WORD file, see Figure 1) posed challenges for reading the interviews in their entirety.

Figure 1: ChatGPT 4.0 1000 character limit



```
# Display the first 1000 characters to get an overview
campello_content[:1000]
```

Source: The authors (2023)

In the same way, for Investigation B, firstly Claude 2.0 analyzed the article and interview transcript, to, after that, conduct its own analysis of the interview according to the theoretical framework presented:

> **CLAUDE 2.0 (1.B):** The student describes how their understanding of relativity concepts evolved from the pre-test to the post-test after instruction. In the pre-test, they lacked an intuitive grasp of relativity and gave responses based on an "absolute" notion of motion. After instruction, they started to incorporate a relativistic perspective.

ISSN 2525-8222
DOI: http://dx.doi.org/10.33361/RPQ.2024.v.12.n.30.724

> The interview reveals the cognitive mediation process underlying this evolution. Initially, the student relied primarily on internal psychophysical mediation - their own perception of motion as absolute. Instruction introduced new external mediations - simulations, images, metaphors, and social interaction. These allowed reconceptualization through hypercultural mediation. […]
>
> Social mediation with peers, through discussion and debate, facilitated perspective taking. Cultural mediations like metaphors (twins paradox, shrinking spacecraft) provided concrete embodiments of abstract principles. Fictional hypercultural sources like Harry Potter also contributed anchoring metaphors.
>
> By integrating these new mediations, the student gradually overcame their initial internalized psychophysical intuitions. They developed more expert-like relativistic reasoning. This aligns with the theory's premise that hypercultural mediations can enhance cognition if effectively internalized.
>
> The interview illustrates the mechanisms by which hypercultural tools reshape thinking. It provides qualitative evidence for the theory's validity in explaining cognitive impacts of digital technologies. The student's growth shows how multifaceted mediations help assimilate counterintuitive physics into an enriched mental model of the world (Authors, 2023).

Claude 2.0 was capable of perceive the conceptual evolution of the student and related it to the use of different mediations. However, it classified some resources incorrectly. While Claude 2.0 assumed that the shrinking spacecraft, showed in an animated GIF, was a cultural mediation, in our analysis, it consists of a hypercultural mediation. Although, it is important to note that this classification was discussed in the research group, not being recognized easily. In another point, Claude 2.0 also classified the Harry Potter movie as a hypercultural mediation, even though it is a cultural mediation. In general, the IA could recognize the role of the mediations in student's conceptual evolution, even though classifying wrongly some of these mediations.

ChatGPT 4.0, on the other hand, firstly provided a short explanation about the paper containing the theoretical framework and the file with the interview. For the analysis, the AI focused only on the first question of the test, which was discussed in the beginning of the interview. It did not analyze any other part of the interview, which contains more seven questions, all of them approached during the dialogue with the student.

The only mediation that ChatGPT 4.0 correctly identified was the psychophysical one. It completely ignored the hypercultural mediations mentioned by the student later in the interview. Moreover, the AI wrongly classified the Galileo's test as a cultural mediation and the social interaction with the interviewer as a social mediation, ignoring the moments where the student mentioned the activities they did with the classmates:

> **ChatGPT 4.0 (1.B):** From the provided excerpt, there isn't direct evidence of S18 using technological tools like simulations or apps to aid their understanding.

ISSN 2525-8222
DOI: http://dx.doi.org/10.33361/RPQ.2024.v.12.n.30.724

> […]
> Their [*the student's*] direct experiences and physical understanding (psychophysical), their interaction with the interviewer (social), and their engagement with established scientific concepts (cultural) all play roles in shaping their cognition (Authors, 2023).

This happened, probably, because ChatGPT 4.0 could not analyze the entire interview as Claude 2.0 was capable of. This analysis of only some excerpts of the document cannot provide the IA with enough information to perform a deeper analysis of the student's conceptions and mediations used.

5.2 Category 2: Conceptual Evolution

The next question is related to the idea of conceptual evolution. Each test yielded a different response, although both agree that changes occurred between the pre-test and the post-test.

> **Claude 2.0 (2.A):** [...] In the pre-test, the student exhibits very poor and limited knowledge. They mention "light waves" and the "brightness of the Sun" but admit to not having a clear understanding of what light is. Their explanation is purely based on visual perception. They mention frequency and wavelength but also admit to not grasping these concepts well. Their knowledge of wavelength and reflection is non-existent. In the post-test, they demonstrate a significantly solid comprehension of luminous phenomena. They explain light as electromagnetic waves, introducing concepts like the electromagnetic spectrum and visible spectrum. They discuss frequency and wavelength, correctly relating them to color. They explain the absorption and emission of light by atoms clearly, discussing electronic transitions. Their explanation of reflection has also evolved, understanding how different materials reflect different wavelengths […] (Authors, 2023).

Claude 2.0 accurately identified the importance of mediations in the evolution of concepts. It emphasized that psychophysical mediation played a fundamental role in the initial stage when the student had a superficial knowledge of the subject with simpler concepts. Subsequently, by using hypercultural tools, the student was able to refine their concepts, developing a more sophisticated scientific language and a deeper understanding of the topic.

> **ChatGPT 4.0 (2.A):** Overall Analysis:
> Pre-instruction: S11 demonstrated a basic and phenomenological understanding of light. He relied primarily on direct observations and had rudimentary familiarity with some technical concepts.
> Post-instruction: There is a clear evolution in his understanding. S11 now employs technical and scientific terminology, demonstrating a more theoretical and interdisciplinary grasp of light. He combines concepts from chemistry and physics and appears to have a more solid understanding of light-related phenomena (Authors, 2023).

It is observed that Chat GPT 4.0 also captured at the same idea; however, due to not reading the entire question, it is unable to develop many responses and, thus, does not

ISSN 2525-8222
DOI: http://dx.doi.org/10.33361/RPQ.2024.v.12.n.30.724

identify the mediations that played a more significant role in this process. Relying on more general arguments.

For the Investigation B, Claude 2.0 could accurately identify the conceptual evolution of the student S18 concerning the relativistic space and time:

> **CLAUDE 2.0 (2.B):** Based on the interview, the student showed clear evolution in their conceptual understanding and reasoning about space and time from pre to post-instruction:
> Pre-Instruction:
> - Had an absolute view of motion - speed and distance were fixed irrespective of frame of reference
> - Did not grasp relativity of simultaneity - time was absolute across frames
> - Did not apply relativistic velocity addition
> - Lacked intuitive understanding of length contraction and time dilation
>
> Post-Instruction:
> - Began to incorporate relativistic perspective taking
> - Recognized need to consider reference frame when judging motions
> - Applied relativistic velocity addition in some situations
> - Showed embryonic grasp of length contraction and time dilation, though still shaky
> - Could articulate explanations using metaphors like twins paradox
> - Formed mental models linking speed and space-time warping
>
> Overall, the pre to post evolution shows a clear shift from a novice Newtonian view to the beginnings of expert-like relativistic reasoning. However, the post-instruction mental models are still tentative and context-dependent. […] Their grasp of relativistic concepts is thus evolving but remains quite novice-like. More robust hypercultural mediation and social negotiation is likely needed to develop greater fluency in applying relativistic perspectives across contexts (Authors, 2023).

Here the analysis was quite good. Claude 2.0 could correctly identify that the student evolved in their conceptions about space and time and constructed mental models to that. It also separated the pre-test and post-test evidence, even though these are not clearer separate in the transcription. Moreover, the IA was capable of perceive that the student departed of a novice Newtonian notion to a more sophisticated conceptions, even though, they have not a deep understanding of it. Furthermore, Claude 2.0 suggested that more hypercultural and social interaction are needed aiming to provide a more robust construction of the concepts. It is important to note that in our previous analysis these two mediations played the more significant role in student's conceptual evolution concerning relativistic space and time.

Besides providing a longer analysis compared to Claude 2.0, ChatGPT 4.0 was not capable of identifying all conceptions S18 showed during the interview. The IA could correctly perceive that the student presented a conceptual evolution, but in a shallow way of analysis. Concerning the concept of space, ChatGPT 4.0 correctly identified the concepts of S18, which are from an absolute perspective in pre-test and changed to a relative view in post-test. It is worth noting, also, that ChatGPT 4.0 could identify the student's reasoning during the interview, when S18 changed the answers while thinking over the question.

However, it is important to note that, as well as Claude 2.0, ChatGPT 4.0 showed misconceptions about the question dealing with throwing a ball in the bus situation. The AI identified that the change in S18's response, considering the necessity of using more force is correct, even though the pre-test's answer was correct, the necessary force does not change, involving the inertia concept. Moreover, dealing with time, the IA affirmed this concept was not directly approached during the interview, even though there were two questions in the test dealing with this concept specifically.

> **ChatGPT 4.0 (2.B):** S18's post-instruction response showcases a grasp of relative motion. They now understand that the bus moving forward would impact the ball's motion, implying that they need to exert more force if they want the ball to move in the opposite direction. This reveals an understanding that motion is not absolute but relative to the surrounding environment. […] S18's post-instruction knowledge shows a clear evolution from their pre-instruction understanding. While initially rooted in absolutes, their post-instruction responses showcase a budding understanding of the relativity of motion. They begin to factor in the effects of external movements (like the bus's motion) on individual actions (like throwing the ball). This suggests that they're starting to grasp more complex physics concepts, even if they haven't fully mastered them. The fact that they're questioning their own understanding ("Now I don't know if I agree with this...") further indicates an active engagement with the learning process, which is a positive sign for future learning and comprehension (Authors, 2023).

Probably, once again, this happened due to the incapacity of ChatGPT 4.0 to analyze all document at the same time, as the questions concerning time concept were approached later in the interview. This resulted in an incomplete analysis of the conceptual evolution of S18's concerning space and, more evidently, time.

5.3 Category 3: Sources of evidence

In a last prompt, we aimed to ask the chatbots to identify where, in the interview, they could detect the evidence obtained to assess students' evolution. To access the conceptual evolution of a student, the entire piece must be visited, therefore, this can help the researcher bring evidence and detect hallucinations (if they were made).

> **Claude 2.0 (3.A):** Pre-test: "I pictured a wall, like this one (points to the wall), where there used to be a window [#JA; 02:06], where the Sun used to shine" and "I don't know. I think there I just thought of it as light that hits and comes back to us. In a more generic way, you know."
> Post-test: "Light is electromagnetic waves, in this case, coming from the sun. In this case, this light is white because it contains all the colors of the visible spectrum added together" and "A red shirt absorbs all the colors being emitted on it, reflecting back or emitting only the red light, which is what we see, and thus, we believe the color of the object is red." […] From these excerpts, I

ISSN 2525-8222
DOI: http://dx.doi.org/10.33361/RPQ.2024.v.12.n.30.724

> could observe the difference in the quality of explanations and the student's conceptual mastery between the pre-test and post-test (Authors, 2023).

While the interview document lacks separate sections for the pre-test and post-test, Claude 2.0 has effectively identified the excerpts corresponding to each phase. This enables a clear distinction between the student's explanations and conceptual mastery, highlighting the disparity in quality between the pre-test and post-test responses. Additionally, it demonstrates the AI's capability to comprehensively read the entire document, as evidenced by its provision of examples from the beginning, middle, and end of the interview. It is important to note that only a portion of the analyses considered by Claude 2.0 has been displayed in the excerpt above.

For ChatGPT 4.0, this part was not very effective or not easily utilizable, given that, as it did not read the entire interview, it remains unable to correctly comprehend the indicated excerpts. However, it is interesting to note that it addressed what was said without directly mentioning the excerpts.

> **ChatGPT 4.0 (3.A):** Definition of Light:
> PRE-TEST: S11's definition of light is based on a direct and tangible observation of the world around him. He relates light to what is visible on the wall when sunlight passes through a window. POST-TEST: His definition evolves into a more technical understanding, identifying light as "electromagnetic waves" and commenting on the spectral composition of white light. […]
> Evaluation: In the POST-TEST, S11 demonstrates an understanding more aligned with scientific concepts of light. S11 has shown a clear and significant evolution in his understanding of light from the PRE-TEST to the POST-TEST. Initially, his responses were more based on direct observations and intuitions. However, in the POST-TEST, he presents a more technical and conceptual understanding of light and its properties (Authors, 2023).

Significant differences were observed between the Claude 2.0 model and ChatGPT 4.0 in analyzing the interview. The interview covered both the pre-test and the post-test. Claude 2.0 managed to distinguish these moments without the need for explicit mentions in the input, allowing the tool to promptly identify which interactions were related to each phase of the test. Additionally, Claude 2.0 identified the specific mediation that contributed to the conceptual development of student S11 from the outset of the conversation.

Analyzing student S18, from Investigation B, Claude 2.0 could provide the sources of its analysis, once again, separating the pre-test and post-test results. However, it did not present any quotation of the interview, as it did for student S11:

> **CLAUDE 2.0 (3.B):** Pre-Instruction:
> - In the ball throwing on a bus question, they initially thought the force would be the same on a moving or stationary bus, indicating a lack of consideration of reference frames.

cc BY NC ND

ISSN 2525-8222
DOI: http://dx.doi.org/10.33361/RPQ.2024.v.12.n.30.724

> - In the velocity addition question with the car and helicopter, they first thought the velocity could not be negative, not grasping relative motion.
> - In the Lorentz contraction questions, they believed length and time were absolute across frames of reference.
> Post-Instruction:
> - In the ball throwing questions, they applied relativistic thinking in some cases, recognizing the need for added force.
> - In the velocity addition question, they realized negative velocity made sense from the helicopter's frame.
> - In Lorentz contraction, they started invoking time dilation and length contraction qualitatively.
> - They described visualizing relativistic effects in simulations.
> - They gave metaphorical explanations like the twins paradox.
> Looking at how their reasoning on similar questions evolved shows the shift from novice absolutist views towards more expert relativistic perspectives. But the post-instruction understanding was still inconsistent, revealing the intermediate nature of their conceptual grasp (Authors, 2023).

It is important to highlight that here Claude 2.0 showed some misconceptions. In the ball throwing on a bus question, the student had responded correctly in the pre-test and incorrectly in the post-test. This question involves the concept of inertia, as the bus moves in a constant velocity, the force for throw the ball would be the same, no matter if it is stopped or moving in a constant velocity. The AI presented a naïve idea usually also presented by the high school students.

However, concerning the other concepts, the Claude 2.0’s analysis was consistent. It correctly analyzed the evolution in the conception of relative velocities and relative motion, moreover Claude 2.0 identified that the student evolved in spacetime comprehension, recognizing the relative nature of it using the mediations to that – simulations (hypercultural) and explanations (social).

Now, let’s analyze the ChatGPT 4.0 answer. Once again, ChatGPT 4.0 provided a long analysis for the question. However, showing the same misconception as before, related to throwing a ball in a bus situation. It incorrectly associated the student’s reasoning for this question with their conceptual understanding of space. Besides providing the quotation of the excerpts where the student mentioned what the AI was approaching – something Claude 2.0 did not for S18 – it only used the initial part of the interview, ignoring the segment where the student discussed other concepts, such as the time. This result probably reflects the incapacity of ChatGPT 4.0 to analyze the entire document at the same time.

> **ChatGPT 4.0 (3.B):** S18 now recognize that the movement of the bus will influence how the ball moves within it. They understand that the forward motion of the bus would impact the ball's motion, which is why they believe more force would be needed to throw it.
> They’re starting to grasp the idea of relative motion, understanding that the bus's movement has an effect on objects inside it.

> The evolution in their understanding is evident when comparing their reasoning from the pre-test to the post-test. In the pre-test, they saw motion and space as absolute, but by the post-test, they began to understand these concepts in relative terms, factoring in the context of the moving bus (Authors, 2023).

Throughout the conversation, it became evident that Claude 2.0 played a more effective role in the qualitative assessment of the interview based on the provided theoretical framework since it does not limit itself to analyzing just the words or terms but also their origin, which is crucial in research like this.

On the other hand, ChatGPT 4.0 could not make this distinction efficiently, resulting in a mix of dialogues related to both the pre-test and the post-test. This made it challenging to identify the mediations, which remained ambiguous even at the end of the conversation, not aligning with previous analyses.

Another notable difference is that, even when explicitly prompted to separate the pre-test and post-test sections, ChatGPT 4.0 could not fully identify the questions within each phase. For instance, concerning the Investigation A, the three questions from the pre-test were replicated in the post-test, and ChatGPT 4.0, besides not correctly identifying the mediations related to the answers to these questions, also did not adequately evaluate the last question in the post-test.

The only predominant mediation that ChatGPT 4.0 managed to identify was in the absorption part. However, it is worth noting that this concept was difficult for S11 to grasp, and although they maintained the idea of psychophysical mediation in both the pre-test and post-test, other less relevant mediations emerged in the post-test, which ChatGPT 4.0 failed to identify.

Analyzing Investigation B’s results, ChatGPT 4.0 only considered the first questions of the test, ignoring the last ones involving the concept of time. Therefore, ChatGPT 4.0 could not correctly analyze the conception evolution of the student and identified many mediations mentioned by S18 during the interview wrongly and ignoring some of them.

Lastly, it is crucial to note that neither of the Large Language Models (LLMs) exhibited hallucinations throughout the entire analysis. Claude 2.0 successfully distinguished between the pre-test and post-test phases without explicit instructions and identified specific mediations. However, ChatGPT 4.0 struggled to make this distinction effectively, mixing the two phases and failing to identify mediations, even with explicit subsequent instructions.

ISSN 2525-8222
DOI: http://dx.doi.org/10.33361/RPQ.2024.v.12.n.30.724

6 Conclusion

In this article, we conducted an examination comparing the qualitative analysis performed by two chatbots using generative artificial intelligences (AIs), ChatGPT 4.0 and Claude 2.0. The research objective consisted of assessing the ability of these AIs to perform in-depth qualitative analysis and compare them to each other.

For the study, two interviews previously analyzed by human researchers were used, which dealt with the conceptual evolution of high school students, specifically regarding the concepts of light and Special Relativity. These analyses were anchored in the theoretical framework of the Cognitive Networks Mediation Theory (Souza *et al*. 2012). The same interviews were then presented to the chatbots, with the request to perform a qualitative analysis based on the provided theoretical framework.

This study complements and goes further than previous usage discussed in the literature review where either pre-processed data was fed into the Generative AI (Hamilton *et al*. 2023) or GenAI was used to produce a tool (vignettes) for further use in the analysis (Chubb, 2023). Therefore, our raw results (interview) were fed into GenAI to produce the final outcome (analysis).

The results obtained indicated that the Claude 2.0 model was able to appropriately distinguish pre-test and post-test moments in the interviews and accurately identify the predominant mediations. These conclusions were largely aligned with the analyses previously conducted by humans.

In contrast, ChatGPT 4.0 faced difficulties in this task, mixing moments and failing to identify mediations specifically, even when explicit instructions were provided. Additionally, ChatGPT 4.0 demonstrated signs of using only parts of the interviews in its analysis, resulting in incomplete feedback. As argued, this is likely due to the more limited context window of ChatGPT 4.0 as compared to Claude 2.0 (with 100.000 tokens of context window). The hindsight of this limited context window was discussed in the literature review as pertinent to other works (Hamilton *et al*. 2023; Chubb, 2023).

Regarding the concept of light in the first investigation (Investigation A), both Claude 2.0 and ChatGPT 4.0 reached the conclusion that there were conceptual changes between the pre-test and post-test. However, Claude 2.0 was able to accurately identify the importance of mediations in the evolution of concepts, highlighting the fundamental role of psychophysical mediation in the initial phase when the student had a superficial knowledge and simpler concepts on the subject. Subsequently, with the introduction of hypercultural tools, the student was able to refine his concepts, developing a more sophisticated scientific language and a deeper understanding of the topic. On the other hand, ChatGPT 4.0, due to its inability to read the entire question, failed to develop detailed responses and, therefore, did not identify the mediations that played a more significant role in this process.

In the second investigation (Investigation B) related to Special Relativity, Claude 2.0 was able to accurately identify the conceptual evolution of student S18 regarding relativistic space and time, correctly highlighting the transition from an initial Newtonian notion to more sophisticated, albeit incipient, conceptions. ChatGPT 4.0, on the other hand, was unable to identify all the conceptions presented by the student during the interview, focusing only on the initial questions and ignoring those related to the concept of time. This can also be attributed to the limitation of ChatGPT 4.0 in analyzing the entire document simultaneously.

In terms of evidence of conceptual evolution (category 3), in Investigation A, Claude 2.0 effectively identified the sections corresponding to each phase (pre-test and post-test) in the interview, even without explicit mentions in the transcription. This allowed for a clear distinction of explanations and the conceptual domain of the student between the two moments. ChatGPT 4.0, on the other hand, could not make this distinction efficiently, mixing dialogues from both the pre-test and post-test, it adversely affected the analysis.

In Investigation B, Claude 2.0 again separated the results of the pre-test and post-test, although it did not provide interview citations as it did for student S11. Overall, Claude 2.0 consistently analyzed the evolution in the conception of relative velocities and motion, as well as the student S18's relative spacetime understanding. ChatGPT 4.0, once again, limited itself to the initial questions of the interview, ignoring the portion where the student discussed other concepts, such as time. This limitation of ChatGPT 4.0 can be attributed to its inability to analyze the entire document simultaneously.

Regarding the qualitative assessment of the interview based on the provided theoretical framework, the Claude 2.0 model stood out as it did not limit itself to analyzing words or terms but also considered their origin, which was crucial for this research. ChatGPT 4.0 could not make this distinction efficiently, which hindered the identification of mediations and sources of evidence. This once again emphasizes the importance of AI models being able to holistically understand data for conducting deep and theoretically grounded qualitative analyses, provided they possess a sufficient context window.

Overall, this study concludes that the Claude 2.0 model proved to be more effective in conducting theoretically oriented qualitative analyses compared to ChatGPT 4.0. Claude 2.0’s ability to understand the entire textual content simultaneously, without context window restrictions, allowed for a more holistic and comprehensive analysis of the data. The authors emphasize that conversational AI models have the potential to assist in qualitative analyses, with a special focus on the Claude 2.0 model compared to ChatGPT 4.0, which presents similar limitations to those of a human, such as the interchange of mediations or lack of attention to the theoretical framework in identifying the origins of mediations.

Therefore, this study concludes that the Claude 2.0 model surpassed ChatGPT 4.0 in terms of qualitative analysis. In conclusion, there are indications suggesting the potential of employing Large Language Models (LLMs) as an additional analyst for conducting qualitative analysis, akin to engaging another interpreter, to enhance triangulation among various interpreters – as discussed by Christou (2023b).

This preliminary notion, underscored at the closure of this paper, invites further exploration by the scholarly community. However, it is emphasized that the researcher’s familiarity with raw data and the theoretical framework remains essential, and this study offers valuable insights into the boundaries between human and artificial cognition in qualitative research (Christou, 2023a).

Claude 2.0’s superior performance on the analysis developed indicates its potential to be used as a tool in the process of qualitative analysis of data. However, it is important to note that the present study was performed within a limited context, using the data obtained from two studies and considering only two students. Therefore, we highlight that more studies are necessary to investigate in depth the capacity of GenAI to perform qualitative analysis.

Finally, one word of caution: currently (this is likely subject to change) Claude 2.0 is not currently available in Brazil; this analysis was possible due to the researchers having obtained access to it before this limitation. As other models are available, and with even larger context window, Brazilian researchers should have increasingly access to those powerful tools for qualitative analysis.

References

ANTHROPIC. Claude v.2.0. Available in: https://www.anthropic.com/index/100k-context-windows. Accessed on: October 9, 2023.

ATKINSON, C. F. Cheap, Quick, and Rigorous: Artificial Intelligence and the Systematic Literature Review. **Social Science Computer Review**, Thousand Oaks, v. 42, n. 02, p. 08944393231196281, aug. 2023.

BAROCAS, S.; SELBST, A. D. Big data's disparate impact. **California Law Review**, Berkeley, v. 104, n. 3, 671-732, 2016.

BELOTTO, M. J. Data analysis methods for qualitative research: Managing the challenges of coding, interrater reliability, and thematic analysis. **The Qualitative Report**, Miami, v. 23, n. 11, p. 2622-2633, 2018.

BISWAS, S.S. Potential use of chat GPT in global warming. **Annals of biomedical engineering**, Piscataway, v. 51, n. 6, p. 1126-1127, 2023.

BOZKURT, A.; XIAO, J.; LAMBERT, S.; PAZUREK, A.; CROMPTON, H.; KOSEOGLU, S.; FARROW, R.; BOND, M.; NERANTZI, C.; HONEYCHURCH, S.; BALI, M.; DRON, J.; MIR, K.; STEWART, B.; COSTELLO, E.; MASON, J.; STRACKE, C. M.; ROMERO-HALL, E.; KOUTROPOULOS, A.; TOQUERO, C. M.; SINGH, L.; TLILI, A.; LEE, K.; NICHOLS, M.; OSSIANNILSSON, E.; BROWN, M.; IRVINE, V.; RAFFAGHELLI, J. E.; SANTOS-HERMOSA, G.; FARRELL, O.; ADAM, T.; THONG, Y. L.; SANI-BOZKURT, S.; SHARMA, R. C.; HRASTINSKI, S.; JANDRIĆ, P. Speculative futures on ChatGPT and generative artificial intelligence (AI): A collective reflection from the educational landscape. **Asian Journal of Distance Education**, Delhi, v. 18, n. 1, 2023.

BRUUN, E. P.G.; DUKA, A. Artificial intelligence, jobs and the future of work: Racing with the machines. **Basic Income Studies**, Berlin, v. 13, n. 2, p. 20180018, 2018.

CHING, T. et al. Opportunities and obstacles for deep learning in biology and medicine. **Journal of The Royal Society Interface**, London, v. 15, n. 141, p. 20170387, 2018.

CHRISTOU, P. A. A critical perspective over whether and how to acknowledge the use of artificial intelligence (AI) in qualitative studies. **The Qualitative Report**, Miami, v. 28, n. 7, p. 1981-1991, 2023a.

CHRISTOU, P. A. How to use artificial intelligence (AI) as a resource, methodological and analysis tool in qualitative research? **The Qualitative Report**, Miami, v. 28, n. 7, p. 1968-1980, 2023b.

CHUBB, L. A. Me and the Machines: Possibilities and Pitfalls of Using Artificial Intelligence for Qualitative Data Analysis. **International Journal of Qualitative Methods**, Thousand Oaks, v. 22, n. 1, p. 16094069231193593, 2023.

CONNEAU, A.; LAMPLE, G. Cross-lingual language model pretraining. **Advances in neural information processing systems**, Cambridge, v. 32, n.01, 2019.

FONSECA, A. L. A. da; CHIMENTI, P. C. P. S.; SUAREZ, M. C. Using deep learning language models as scaffolding tools in interpretive research. **Revista de Administração Contemporânea**, Maringá, v. 27, v. 03, p. e230021, 2023.

HAMILTON, L.; ELLIOTT, D.; QUICK, A.; SMITH, S.; CHOPLIN, V. Exploring the Use of AI in Qualitative Analysis: A Comparative Study of Guaranteed Income Data. **International Journal of Qualitative Methods**, Thousand Oaks, v. 22, n. 01, p. 16094069231201504, 2023.

HOSSEINI, M.; HORBACH, S. PJM. Fighting reviewer fatigue or amplifying bias? Considerations and recommendations for use of ChatGPT and other Large Language Models in scholarly peer review. **Research Integrity and Peer Review**, Palo Alto, v. 8, n. 1, p. 4, 2023.

KIRMANI, A. R. Artificial intelligence-enabled science poetry. **ACS Energy Letters**, Washington, v. 8, n. 1, p. 574-576, 2022.

LONGO, L. Empowering qualitative research methods in education with artificial intelligence. In: **World Conference on Qualitative Research**. Cham: Springer International Publishing, Porto, v. 1068, p. 1-21, 2019.

LUND, B. D.; WANG, T. Chatting about ChatGPT: how may AI and GPT impact academia and libraries?. **Library Hi Tech News**, Leeds, v. 40, n. 3, p. 26-29, 2023.

LUND, B. D; WANG, T; MANNURU, N.R; NIE, B; SHIMRAY, S; WANG, Z. ChatGPT and a new academic reality: Artificial Intelligence-written research papers and the ethics of the large language models in scholarly publishing. **Journal of the Association for Information Science and Technology**, Syracuse, v. 74, n. 5, p. 570-581, 2023.

MESEC, B. **The language model of artificial inteligence chatGPT-a tool of qualitative analysis of texts**. [Preprint]. 2023. Available in: https://doi.org/10.22541/au.168182047.70243364/v1.

MILOYAN, B.; MCFARLANE, K.; VASQUEZ-ECHEVERRIA, A. The adapted Autobiographical interview: A systematic review and proposal for conduct and reporting. **Behavioural Brain Research**, Amsterdam, v. 370, p. 111881, 2019. Available: https://doi.org/10.1016/j.bbr.2019.03.050.

OpenAI. **Introducing ChatGPT**, 2022. Retrieved from https://openai.com/blog/chatgpt.

SANTOS, D. J. M.; SANTOS, P. V.; BOSS, S. L. B. A análise de conteúdo com apoio do software gratuito RQDA. **Revista Pesquisa Qualitativa**, São Paulo, v.11, n.28, p. 805-824, set./dez. 2023.

SOBANIA, D.; BRIESCH, M.; HANNA, C.; PETKE, J. **An analysis of the automatic bug fixing performance of chatGPT**. [Preprint] arXiv:2301.08653, 2023. Available in: https://arxiv.org/pdf/2301.08653.pdf. Melbourne. Accessed on: October 9, 2023.

SOUZA, B. C.; SILVA, A. S.; SILVA, A. M.; ROAZZI, A.; CARRILHO, S. L. S. Putting the Cognitive Mediation Networks Theory to the test: Evaluation of a framework for understanding the digital age. **Computers in Human Behavior**, Québec, v. 28, n. 6, p. 2320-2330, 2012.

TABONE, W.; WINTER, J. Using ChatGPT for human–computer interaction research: A primer. **Manuscript submitted for publication**, 2023.

TATE, T.; DOROUDI, S.; RITCHIE, D.; XU, Y. **Educational research and AI-generated writing**: Confronting the coming tsunami [preprint], 2023. Available in: http://dx.doi.org/10.35542/osf.io/4mec3.

TREVISAN, R.; SERRANO, A.; WOLFF, J. F. S.; RAMOS, A. F. Peeking into their mental imagery: The Report Aloud technique in science education research. **Ciência e Educação**, Bauru, v. 25, n. 3, pp. 647-664, 2019. Available in: https://doi.org/10.1590/1516-731320190030004.

ISSN 2525-8222
DOI: http://dx.doi.org/10.33361/RPQ.2024.v.12.n.30.724

YANG, C.; WANG, X.; LU, Y.; LIU, H.; LE, Q. V.; ZHOU, D.; CHEN, X. **Large language models as optimizers**. [Preprint] arXiv:2309.03409, 2023. Available in: https://arxiv.org/pdf/2309.03409.pdf. Accessed on: October 9, 2023.

ZHONG, Q.; DING, L.; LIU, J.; DU, B.; TAO, D. **Can chatGPT understand too? a comparative study on chatgpt and fine-tuned bert**. [Preprint] arXiv:2302.10198, 2023. Available in: https://arxiv.org/pdf/2302.10198.pdf. Accessed on: October 9, 2023.

Received on: October 11, 2023.

Accepted on: March 4, 2024.