Evaluating ChatGPT to support data visualization design

George Moreno de Oliveira[1], **Simone Diniz Junqueira Barbosa**[1]

[1]Departamento de Informática
Pontifícia Universidade Católica do Rio de Janeiro - PUC-Rio
Rio de Janeiro, RJ – Brasil

{gmoreno,simone}@inf.puc-rio.br

Abstract. Introduction*: Large Language Models (LLMs) like ChatGPT offer new avenues for diverse tasks, including complex activities such as data visualization design. While existing studies explore LLMs for specific design activities, particularly product generation and ideation, a comprehensive investigation into their support for the entire data visualization design process, especially for non-experts, remains largely unexplored.* ***Objective****: This work addresses this gap by investigating LLMs' capability to assist beginners in applying design methods and tools throughout the data visualization design process. Our research was guided by the question: "How can LLMs support the visualization design process?"* ***Methodology****: A preliminary study explored prompt strategies for generating data visualization recommendations and established criteria for evaluating model response quality. Methodologically, we analyzed design techniques, created prompts for non-specialist designers, and evaluated the process of building design guides with experts, defining a usage context and evaluation criteria for assessing user interaction and perception.* ***Results****: Our findings indicate that ChatGPT can support both abstract and tangible design activities at varying levels. However, it is crucial that users possess sufficient domain knowledge to critically assess the quality of the model's responses, as relying solely on LLMs without this expertise is not advisable.*

Keywords *Data Visualization, Design Process, Large Language Models.*

1. Introduction

Data visualization is a powerful means of "visual representation and presentation of data to facilitate understanding" [Kirk 2016]. To enhance this understanding, visual complements such as charts and images are commonly employed [Knaflic 2015]. The creation of these visual artifacts can range from manual techniques, like paper and pencil, to advanced digital tools that enable interactive, detailed, and easily shareable visualizations [Munzner 2014].

The proliferation of both proprietary and free software solutions, including widely used tools like Microsoft Excel and PowerPoint, and Google Sheets and Slides, has democratized data visualization. These tools empower non-programmers to generate visualizations from data, enabling common users to create financial spreadsheets, student grade charts, and digital dashboards. This accessibility has made the process of choosing visualization types, information layouts, and final presentations a fundamental aspect of data visualization design. This design process guides decisions from initial curiosity to data manipulation and the final presentation choices [Kirk 2016].

The design process itself is multifaceted, involving various methods and tools to aid in its different stages. For instance, understanding the problem context for a visualization might involve user interviews [Bischof et al. 2011], competitive analysis [Chen et al. 2025], and/or surveys. While there are myriad design methods and tools in areas such as service design [Alves e Jardim Nunes 2013, Stickdorn et al. 2021], digital product design [Carvalho 2022, Kumar 2013, Rodrigues Catalano e Rossi Lorenzi 2023], and specifically data visualization design [Parsons 2022, McKenna et al. 2014, Kirk 2016, Munzner 2014], each offers similar structured approaches, in terms of content, to guide designers.

Recently, the emergence of Large Language Models (LLMs) such as ChatGPT, Gemini, Copilot, Llama, and Claude has opened new avenues for non-expert users. These tools, often featuring conversational chat interfaces, allow users to quickly obtain information without requiring programming knowledge. Although LLMs provide rapid access to diverse information, their reliance on embedded knowledge limits user intervention when a system fails to provide an adequate response [Nguyen et al. 2022].

Considering the potential of LLMs to facilitate access to information through Natural Language Processing (NLP) across various contexts, this work investigates their capacity to support data visualization design. Specifically, we explore how these models can assist in applying design methods and tools. Our research was guided by the overarching question: "How can LLMs support the data visualization design process?" This led to three specific sub-questions:

1. How can LLMs *assist* designers in the data visualization design process, without just doing it for them?
2. At which points in the design process can LLMs provide assistance?
3. How can LLMs help non-expert data visualization users execute design process steps?

To address these questions, this study evaluates the ability of LLMs, specifically ChatGPT – chosen for its pioneering conversational interface and widespread use among interviewed participants –, to support the application of design methods and tools in data visualization. This evaluation involved participants with varying levels of expertise, conducting two distinct design activities, and comparing the support provided by ChatGPT with that of human data visualization design experts.

Our methodological approach to answer the research questions comprised three main parts: (1) an analysis of design methods and tools relevant to data visualization design, drawing from existing literature; (2) the synthesis of prompts and scenarios for conducting activities with non-expert designers and constructing expert-guided guides; and (3) user observations with all participants, complemented by an evaluative questionnaire based on six developed criteria. Through this implementation, we sought to identify how ChatGPT could support design activities, leading to a clearer understanding of how designers' prior knowledge influenced their critical evaluation of model responses – a crucial aspect for effective use of LLM.

This paper is structured as follows: Section 2 details the research methods, including the LLM prompt construction and the design of the user studies. Section 3 presents the qualitative findings from both the LLM usage by non-specialist designers and the expert-guided guide construction, followed by a comparative discussion of their

support capabilities. Section 4 highlights the specific contributions of this work to the field of Human-Computer Interaction. Finally, Section 5 summarizes the main conclusions and outlines avenues for future work.

2. Methodology

To investigate "how LLMs can support the data visualization design process," this study proposes an evaluation of model responses in design assistance. Our methodology, illustrated in Figure 1, comprised four main stages: (1) identifying relevant design process stages, methods, and tools; (2) constructing prompts specifically designed to facilitate these selected methods and tools; (3) conducting observation sessions where users with varying levels of expertise interacted with the models; and (4) comparing LLM-generated responses with those provided by human experts. The detailed steps and refinements, particularly in prompt creation, are elaborated in the following subsections.

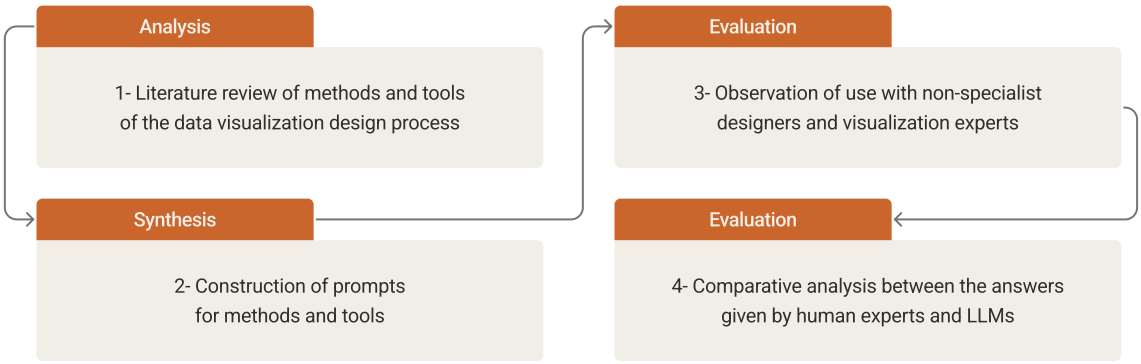

Figura 1. Overview of the research methodology, presented in four main stages.

2.1. Literature review of methods and tools of the data visualization design process

Design processes are adaptable across various contexts, stages, and projects. Selecting and correctly applying design tools can significantly alter outcomes. For instance, using a questionnaire instead of direct observation would produce different insights. While simple online searches can offer quick overviews or practical examples of design activities, designers often encounter extensive guides, some generic, others specific to data visualization. Notably, works by [McKenna et al. 2014] and [Parsons 2022] collectively identified 102 design methods and tools, 58 of which are applicable to data visualization design.

These 58 methods are categorized into four process stages: understanding (39), ideation (25), production (34), and launch (29) [McKenna et al. 2014]. A single method, like a questionnaire, can be used across multiple stages (*e.g.*, initial discovery, ideation input, production validation, or post-launch evaluation). Based on this, [Parsons 2022]'s research methodology, using a survey, allowed data visualization designers to classify methods based on familiarity and frequency of use. This enabled us to filter the 58 methods, focusing on the quadrant representing the most frequently used and familiar methods, which resulted in six core methods, as shown in Figure 2.

Among these most frequent and familiar design methods and tools, we selected concept sketches/wireframing and user interviews as the bases for building prompts that

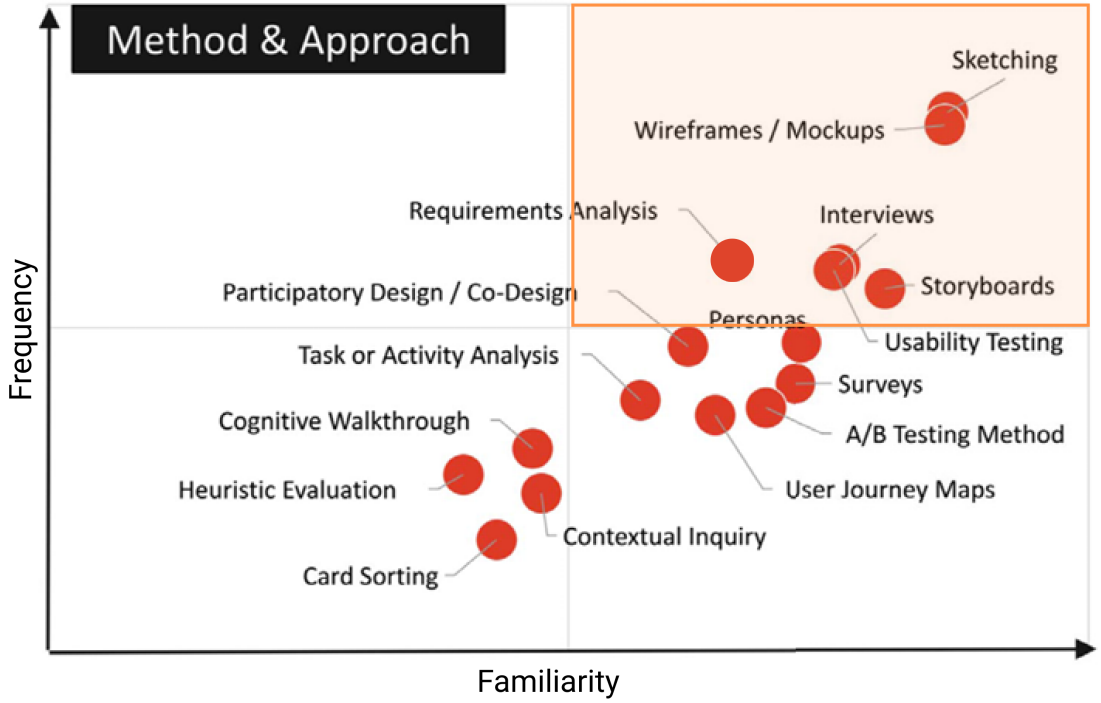

Figura 2. Quadrant showing the most frequent and familiar methods from [Parsons 2022].

aim to guide designers on how to apply these methods. We chose them not only for their familiarity and frequency of use, but also because one is inherently textual (interview scripts) and the other visual (wireframes).

2.2. Construction of prompts for methods and tools

Prompts serve as instruments for communication between users and pre-trained models. In this work, focusing on chat-based interfaces, we utilized textual interaction, where users communicate with the model solely through text. According to OpenAI's guidelines,[1] the initial step in prompt construction involves defining the model's role, summarizing its subsequent activities. For our context, a suitable role was:

> *"You are a design assistant who should directly guide a designer who wants to carry out [...] aiming to perform a data visualization design activity."*

This "persona" (as OpenAI calls it) helps the model understand its role as a data visualization design assistant and generate contextually relevant responses. For activity-oriented prompts, where the model should perform a task rather than just answer a question, it is crucial to include specific details, delimiters, output format specifications, and, if possible, examples. An illustrative example is:

[1] https://platform.openai.com/docs/guides/prompt-engineering/six-strategies-for-getting-better-results

> *"As an assistant for data visualization design activities and techniques, create a list of ideas, tips, or guiding questions to support the creation of a concept sketch for a data visualization."*

This example conveys the model's role, specific task details, and the desired output format (a list). While these elements can be added incrementally, for user convenience during our observation sessions, we provided them in a single initial command.

2.3. Observation of use with non-specialist designers

User interaction with LLMs can vary based on the interface (chat or API). Chat interfaces facilitate use for individuals with limited programming knowledge, often accessible via the internet without demanding significant computational resources. As our study was conducted in February 2025, ChatGPT used the 4.o mini language model, the free version that all participants could access.

In this context, we conducted direct observation sessions. This involved presenting participants with a task and observing their interaction in a simulated or real-world scenario. While useful for identifying if ChatGPT adequately responds to user needs and for refining prompts, direct observation can introduce observer interference.

Participants were divided into two main groups based on the activity performed, without prior distinction by data visualization knowledge. The sole requirement was some familiarity with the chosen design methods and tools, ensuring their focus remained on evaluating the model's support rather than learning the methods themselves.

2.3.1. Interaction script

We developed guides for the two selected data visualization design methods, along with their respective prompts. These materials aimed to facilitate easy access to the commands, activity scenarios, and target audience. An example scenario for a participant acting as a visualization designer using the model for concept sketching was:

> *"You are participating in a data visualization project about the financial problem caused by the emergence and legalization of betting houses. Data is available on the amount invested in bets before and after legalization; the age of the most frequent users; social class (divided from A - very rich to E - very poor), and betting location. Your role, at this moment, is to build a visualization sketch. How could you use the guide and ChatGPT for this?"*

For this study, we developed one scenario per design method, enabling participants to understand and attempt the activity based on the provided context. During their interaction with the guide and the model, participants verbalized their thoughts, expressing frustrations, satisfaction, or other emotions/desires related to the interaction. A pre-observation interview was conducted to collect demographic information (age, research/work area, years of experience, and design activity practice). After each activity, participants were given some time for free discussions and critique, following a standard timeline:

- Initial conversation - 5 minutes;
- Activity presentation - 10 minutes;
- Initial prompt application - 3 minutes;
- Activity execution - 35 minutes;
- Free discussions and critique - 5 minutes;
- Post-use questionnaire - 2 minutes.

These sessions were conducted remotely, allowing for video and audio recording of participant interactions for subsequent analysis.

2.3.2. Questionnaire for Use Evaluation with Provided Criteria

The evaluation questionnaire aimed to assess ChatGPT's responses in supporting data visualization ideation using criteria from [Kim et al. 2025] and Grice's communication maxims [Grice et al. 1975]. The objective was to understand both novice and expert users' perceptions after interacting with the model during a design activity supported by its corresponding guide. We adapted these models to create simpler yet comprehensive evaluation criteria:

- C1: Is relevant and focuses on what the question asks;
- C2: Mentions data not provided in the question;
- C3: Provides sufficient information to build the visualization;
- C4: Provides more information than necessary to build the visualization;
- C5: Is easy for non-specialists to understand;
- C6: Recommends an appropriate visualization for the question.

These criteria were adapted as shown in Table 1.

Tabela 1. Adaptation of criteria from [Kim et al. 2025] and [Grice et al. 1975]

	Criterion	C1	C2	C3	C4	C5	C6
[Kim et al. 2025]	Coverage		X				X
	Focus	X					
	Breadth		X		X		
	Clarity			X		X	
	Depth		X		X		
	Applicability			X			X
[Grice et al. 1975]	Quantity			X		X	X
	Quality	X	X	X			
	Relevance	X		X	X		X
	Manner			X		X	

For each criterion, participants rated their agreement or disagreement using a 7-point Likert scale: "Strongly Disagree,Partially Disagree,Disagree,Neutral,Agree, Partially Agree,"and "Strongly Agree."

2.4. Comparative analysis between the answers given by human experts and LLMs

To evaluate how the investigated model supports the design process, we compared suggestions provided by data visualization experts[2] with those suggested by the LLM.

The primary objective of this comparison was to identify differences and similarities between responses and to gather specific knowledge for future research. This comparison was conducted with the assistance of experts who provided responses for the same use scenarios employed in the user observation study.

The comparison was carried out by the author, using two primary materials: the written material prepared by the expert and notes made during the observation of the construction of the expert's guide. These materials allowed for the identification of insights that written documents alone could not reveal, such as "think aloud" processes or real-time edits based on the logical progression of the work. This approach helped us find similarities in the written format of responses, but also highlighted significant differences in verbal communication during real-time observation, as detailed in Section 3.

2.5. Ethical Considerations

This project sought to analyze another form of design support for the exploratory data analysis through data visualization, as part of the ongoing research activities under an overarching project (Parecer 063/2020 (Protocol 97/2020) approved by the Research Ethics Committee PUC-Rio (Câmara de Ética em Pesquisa da PUC-Rio).

All research procedures involving human participants were conducted in strict adherence to ethical guidelines. Participants were recruited on a voluntary basis, ensuring they understood the study's objectives, the detailed procedures involved, and their fundamental right to withdraw from participation at any time without penalty.

Prior to their involvement, all individuals provided their informed consent. This process ensured participants fully comprehended the nature of the study, the types of data that would be collected (specifically screen and voice recordings), and the measures implemented to guarantee their confidentiality and anonymity throughout the research process. To protect their privacy during data analysis and presentation of results, personal identifying information was meticulously removed from all collected data, and participants were systematically assigned pseudonyms (D1-D6 for non-specialist designers and E1-E6 for data visualization experts). Furthermore, all collected data were stored securely and accessed exclusively by the authorized research team for the sole purpose of this study. Participants were also consistently provided with clear contact information for the research team, ensuring they could address any questions or concerns before, during, or after their participation.

3. Results and Discussion

This section synthesizes the main findings from our study, presenting the observed patterns in LLM (ChatGPT) usage by non-specialist designers and comparing them with insights from human experts, and describes the results of the prompts to perform the activities.

[2]In this work, experts were considered participants with a Master's degree or higher and published works in data visualization.

3.1. Prompt Construction

We specified the type of textual output by using the command: "Construct a guide that describe the steps, with guidance, that the designer should follow to build or create...". This was supplemented with different instructions for the two activities: "..., but without generating the final questions" for the interview task, and "..., if necessary, generate questions to help the designer to think about visualization" for the wireframe task. A notable issue observed by participants was the model's tendency to disregard the instruction not to generate final questions, instead providing complete answers rather than directional guidance.

Finally, we defined the output length with a minimum size using: "Generate a guide with at least 10 items to support the designer in building a..." The model consistently adhered to this minimum, generating a minimum of 10 items in most cases, with one instance producing 15.

The final prompt for the interview script construction task was:

> You are a designer assistant who should directly guide a designer who wants to conduct an interview with technology and information analysts at the Central Bank to perform a data visualization design.
> Construct a guide that outlines the steps, with guidance, that the designer should follow to build the interview script, but without generating the final questions for the interview.
> Generate a guide with at least 10 items to support the designer in building a script for an interview with users to identify data analysis needs regarding the use of PIX in Brazil.

And for the wireframe construction task:

> You are a designer assistant who should directly guide a designer who wants to build a wireframe in a data visualization design process.
> Construct a guide that outlines the steps, with guidance, that the designer should follow to build the wireframe; if necessary generate, questions that challenge the designer.
> Generate a guide with at least 10 items to support the designer in building a wireframe for an interactive data visualization dashboard to have a preliminary visual representation of the interface that allows for validation of the structure, information hierarchy, and planned interactions for a data visualization about the use of PIX as a payment method in Brazil.

3.2. Using ChatGPT with Non-Specialist Designers

In a data visualization design process, various tools can assist in identifying requirements for the visualization, such as its content, display location, interaction, or accessibility. With these requirements defined, designers can create sketches to test the solution before development. These sketches can range from less detailed wireframes to more detailed high-fidelity prototypes.

We chose to focus on two common design activities – identifying requirements through interviews and creating wireframes – with non-specialist designers who had some prior knowledge of design.

The activities were conducted remotely and synchronously via Google Meet, allowing participants to use their own devices while the evaluator recorded the screen and audio. The ChatGPT model, a freely available chat interface accessible on various devices, was an ideal choice as it requires no programming knowledge, making it a highly accessible tool for designers.

3.3. LLM Performance with Non-Specialist Designers

We conducted six observation sessions with non-specialist designers (D1-D6), divided between two core data visualization design activities: interview script construction and wireframe creation. Participants were selected for their basic design knowledge, not necessarily data visualization expertise, and their reported practical experience varied (Table 2). All sessions were conducted remotely and synchronously, allowing for screen and voice recording for subsequent analysis.

Tabela 2. Table of participants in the user observation

Designer	Activity	Experience	Area	Time
D1	Interview	1	Graphic Design	< 5 years
D2	Interview	2	Experience Design	> 3 years
D3	Interview	3	Digital Product Design	< 5 years
D4	Wireframe	2	Digital Product Design	> 5 years
D5	Wireframe	3	Digital Product Design	< 4 years
D6	Wireframe	3	Digital Product Manager	< 5 years

3.3.1. Interview Script Construction

For the interview script construction, the ChatGPT model generally provided good initial guidance on the overall conceptualization of the script, offering generic yet helpful starting points. For instance, it suggested “Define clearly the objective of the interview: [...]. This objective will be the basis for directing all questions and the script structure.” This behavior of providing generic initial directions was consistent across all participants, indicating a similar response structure from the model.

Following these initial conceptual directions, ChatGPT attempted to be more specific to the context provided in the prompt, detailing potential data, interactions, and even quality criteria that could be elicited during the interview, such as in: “Investigate where PIX data comes from and how analysts access this data. Are there limitations in data access or format? Does the visualization need to reflect data from different sources or systems?” These detailed directions, while phrased differently, presented largely similar content across all three participants. Key thematic areas covered by the model included:

- Identification of the target audience: “Identify the characteristics of the technology and information analysts at the Central Bank.”

- Context and workflow: "Before diving into more technical questions, seek to understand how PIX has been used by the Central Bank."
- Data origin: "Investigate where the data on PIX comes from and how analysts access this data."
- Metrics and indicators: "Ask about the most relevant types of data (transactions, fraud, regional adoption, volume of money moved, etc.)."
- Common formats and visualization: "Investigate which data presentation formats are most effective (charts, tables, maps, interactive dashboards, etc.)."
- Common tools and systems: "Ask about the tools or platforms that analysts already use for their analyses."
- Information accessibility: "Identify difficulties users face in navigating and understanding current dashboards or reports."

For each point, the model offered justifications, examples of responses, or even direct questions. These examples proved particularly valuable for designers D1 and D2, who had less practical experience in interview script construction. D1, struggling to formulate initial questions, found immediate utility, directly copying suggested content for their script. D2 viewed the model's examples as a "source" that needed only "minimal" editing. This highlights ChatGPT's effectiveness as a rapid idea generator, helping less experienced users overcome the "blank page syndrome" by providing a concrete starting point for refinement.

However, D3, an experienced designer, expressed discomfort when the model started generating complete questions, which contradicted the prompt's instruction to guide rather than generate final questions. D3 noted that, while the suggestions were good, they "did not respect the initial command" and "ended up inducing me towards the questions formulated by the model." D3 preferred a more abstract guidance, akin to constructing a user journey, by mapping "step-by-step," understanding "sentiments in relation to all steps," collecting "resources and support tools during the journey," and identifying "improvement opportunities." This nuanced feedback reveals a tension: while explicit, ready-made answers can provide a good start to novice users, they can inhibit the creative and critical thinking of experienced designers who expect a more open-ended, meta-level guidance. D2 also echoed this sentiment, suggesting the model should start with generic ideas and become more specific only upon further interaction, acting as a flexible idea base rather than a direct answer generator.

3.3.2. Wireframe Construction

The wireframe construction activity presented a different dynamic, as it inherently requires a visual output. Participants were provided with a prompt and a list of five key system requirements for an interactive dashboard. These requirements aimed to guide their thinking about visualizations, interactions, and data access: (1) Explore trends in the data, (2) Compare metrics between categories, (3) Navigate quickly between levels of detail, (4) Share and communicate insights, and (5) Access information inclusively.

Unlike the interview activity, all designers in this group (D4, D5, D6) engaged in multiple interactions with ChatGPT to refine their requests and achieve satisfactory outputs.

The model's initial responses for wireframe creation were primarily conceptual, focusing on general design considerations rather than explicit visual construction steps. It provided indications related to:

- Classifying information importance (*e.g.*, "What is the main question the dashboard should answer?");
- Defining key data and sources (*e.g.*, "Are there geographic or demographic data that need to be included?");
- Information hierarchy (*e.g.*, "Which information should be most highlighted?");
- Selecting visualization types (*e.g.*, "Line graph for trends over time");
- Defining panel interactivity (*e.g.*, "Will there be dynamic filters (by date, region, sector)?");
- Defining information accessibility (*e.g.*, "Collect feedback on usability and understanding of the data").

These directions were interpreted and used differently based on each participant's experience. D4, who reported having average practical experience but was not actively working in the field, found the initial responses generic and lacking direct guidance for practical needs like specific data or interaction types. D4's frustration mounted when the model, after receiving responses to its own guiding questions, attempted to "correct" the initial command instead of generating new, more refined suggestions. As D4 stated, "It [ChatGPT] failed in its second response, in my view. It changed the initial request, but I didn't want that." However, after persistent interaction, the model did eventually provide more tailored responses based on the user's detailed input, such as defining data priority lists (*e.g.*, "Total transaction volume - High priority").

D5, who actively works in the area, also found the initial responses generic, necessitating requests for specific guidance on data usage, filter types, and visual presentation – more practical, execution-oriented questions. ChatGPT adapted more easily to these direct requests, providing more practical directions. D5 subsequently disregarded the initial instructions, relying solely on the model's suggestions. A request for example data, aiming to align with the task's objective and desired years, resulted in a text that was difficult for the participant to interpret, requiring further interactions.

D6 demonstrated high proficiency in LLM interaction, quickly skimming initial responses and initiating new, shorter, and more direct prompts. This led to a more efficient iterative process, yielding focused responses. D6 also leveraged ChatGPT's multimodal capabilities, uploading a screenshot of their initial wireframe for the model to analyze based on its previous recommendations. This resulted in a list of unmet items requiring attention, initiated by a prompt like: "Now, perform an analysis [visual analysis of the prototype], based on your suggestions, of this wireframe for a data analysis dashboard. Give tips on how it can be improved [...]." This highlights that sophisticated prompt engineering, including multimodal input, significantly enhances LLM utility for visual design tasks.

Building the wireframe was more difficult than the interview, as the outcome of the activity was not clearly defined in the prompt. Each user used the chat differently, sometimes requesting text assistance and sometimes visual results.

The difficulty lay in following a linear flow of activity execution to achieve a result within the defined timeframe, even when observing usage.

3.3.3. Post-Observation Questionnaire Results

Upon completing each activity, participants evaluated the model's responses using a questionnaire based on six criteria (C1-C6) adapted from [Kim et al. 2025] and Grice's communication maxims [Grice et al. 1975] (Table 1 in Section 2 – Methodology). The criteria were adjusted to align with the specific activity (*e.g.*, "sufficient information to (build the interview script|build the wireframe)").

Participants generally provided positive evaluations for C1 ("Is relevant and focuses on what the question asks"), confirming the model's relevance. For C2 ("Mentions data not provided in the question"), most indicated the model did not introduce extraneous information, perceiving its role as an idea generator rather than a source of unasked data. This was notable given some initial user confusion about the phrasing of this question (*e.g.*, "I think it brought information not provided in the question, but in a positive way.", indicating a positive interpretation of idea generation). High agreement was found for C3 ("Provides sufficient information"), though some participants in the wireframe activity noted initial shortcomings, requiring further interaction. C4 ("Provides more information than necessary") yielded varied results, reflecting the generic nature of some model responses, which could be perceived as "more information than necessary" if not directly actionable.

Regarding ease of understanding for non-specialists (C5 – "Is easy to understand for non-specialists"), the participants showed high agreement, attributing this to the model's more practical directions for their task. In contrast, participants in the wireframe activity largely disagreed, reflecting the increased need for iterative prompting and clarification. Finally, for C6 ("Recommends an appropriate visualization"), participants most strongly agreed that the model provided good recommendations for the requested design activities, indicating its overall capability in guiding design choices.

In addition to the specific criteria, most participants felt supported by the model's guidance. However, they also suggested improvements: more practical, action-oriented responses in the wireframe activity (*e.g.*, asking the model to "list information and data in order of priority that would be good to include in a dashboard") and, for the interview activity, a preference for initially generic responses that deepen contextually upon further demand rather than immediate detailed questions. This aligns with the experienced designers' desire for meta-guidance.

3.4. Expert-Guided Design Guide Construction

To evaluate ChatGPT's performance, we hired six data visualization experts (E1-E6) to construct design guides for the same interview and wireframe construction activities. Experts, all with postgraduate degrees in Computer Science from PUC-Rio, published works in data visualization, and with more than 5 years of experience, participated in synchronous, remote sessions, allowing us to capture their "think-aloud" processes [Van Someren et al. 1994] alongside their written output. This provided crucial insights into their reasoning, examples, experiences, and challenges, often not fully captured in the final written document.

3.4.1. Interview Guide Construction by Experts

Experts provided comprehensive recommendations for interview script construction, often aligning with a basic structure for engaging target audiences and identifying system/data requirements. Table 3 presents the recommendations most cited by experts for producing an interview script.

Tabela 3. Expert recommendations for interview script construction.

Recommendation	E1	E2	E3
Data-driven decisions	X	X	X
Desired devices/platforms	X	X	X
Visualization tools	X	X	X
Data importance	X	X	X
System navigation	X	X	X
Other uses of visualizations	X	X	X
System personalization	X	X	X
Desired visualizations	X	X	X
Accessibility	X		X
Necessary data	X	X	
Expected interactions	X	X	
Data sensibility	X	X	

Experts' thought processes often exhibited continuous, iterative reasoning, where a question depended directly on previous answers. For instance, E1 justified asking about familiar visualization tools by noting it helps "define which tool will be used to build interactive visualizations" and allows exploring "what they think of each tool." This dynamic, often described as "If X, then Y, otherwise Z," was particularly evident when discussing the elicitation of visualization tools and current user visualizations, frequently leading to inquiries about their pros and cons.

Experts also consistently justified their recommendations with practical applications in later project phases. This emphasis on contingent reasoning and contextual justification was a hallmark of the expert-generated guides, providing a more adaptive and nuanced pathway for the designer. Furthermore, during the "think aloud" process, experts spontaneously mentioned "bad cases" or common mistakes from their experiences, offering warnings and alternative strategies that were not explicitly requested or typically found in LLM-generated guides. This implicit knowledge and adaptive problem-solving, communicated through their verbalization, highlighted a qualitative difference in guidance.

3.4.2. Wireframe Guide Construction by Experts

For the wireframe guide construction, specialists provided recommendations that focused on the conceptual and structural aspects of wireframing for data visualization dashboards, aligning with the system requirements outlined in the study. Their guidance emphasized systematic approaches to translating functional requirements into visual layouts. The key recommendations observed across experts (E4, E5, E6) are summarized in Table 4.

Tabela 4. Expert recommendations for wireframe construction.

Recommendation	E4	E5	E6
Defining User Goals	X	X	X
Information Hierarchy	X	X	X
Layout and Structure	X	X	X
Interaction Design	X	X	X
Data Representation Choice	X	X	X
Filtering and Navigation	X	X	X
Accessibility Considerations	X		X
Feedback and Iteration	X	X	
Scalability	X		
Integration with other systems		X	

Similar to the interview guide construction, experts in the wireframe activity often demonstrated a highly structured yet flexible thought process. They emphasized starting with user needs and data available before moving to visual elements. For instance, discussions often revolved around prioritizing information based on user goals (linking to "Defining User Goals") and then selecting appropriate visualization types ("Data Representation Choice") that also facilitate common user interactions like filtering and navigation.

The "think-aloud" protocol revealed how experts considered not just what to include in a wireframe, but why and how different elements would support user tasks and information consumption, often referencing principles of visual perception and usability. This included considering "bad cases" of cluttered or confusing layouts, and how to simplify them. The emphasis on iterative design and user feedback was also prominent, suggesting that wireframing is not a one-time task but a continuous refinement process.

3.5. Discussion: Comparing LLM and Expert Support

Our findings reveal key differences and similarities between LLM and human expert support in data visualization design, addressing the problem of "how LLMs can support the data visualization design process." While LLMs (specifically ChatGPT) excelled at rapid generation of broad ideas and structured initial guidance, particularly beneficial for non-experts struggling with initial conceptualization, human experts provided a deeper, more contextualized, and adaptive form of support.

Strengths of LLM Support:

- Rapid Idea Generation: For novice designers, the LLM effectively served as a "brainstorming machine," quickly providing a starting point and overcoming the "blank page" problem. This was particularly evident in the interview script activity where D1 and D2 directly leveraged the model's suggestions.
- Adaptability to Direct Prompts: When users, especially more experienced ones like D6, provided highly specific and iterative prompts, the LLM demonstrated considerable adaptability, refining its output to better meet complex needs, even handling multimodal input (image analysis of wireframes).

Limitations of LLM Support:

- Failure to Adhere to Constraints: A significant limitation was the LLM's occasional failure to follow instructions (*e.g.*, "not generating final questions"). This produced straightforward answers that, while potentially useful for beginners, could hinder critical thinking and independent problem-solving for more experienced users.
- Generic initial responses for complex tasks: For wireframes, LLM's initial responses were often too conceptual or generic, requiring extensive subsequent guidance to gain practical guidance.

Strengths of Expert Support:

- Contextualized Guidance: Experts provided adaptive, "if-then-else" reasoning, offering guidance that evolved based on hypothetical scenarios and potential user responses. This level of dynamic, contextual support is difficult for current LLMs to replicate.
- Implicit Knowledge and Experiential Alerts: Experts revealed valuable implicit knowledge, such as common mistakes or alternative strategies, arising from their extensive practical experience.

Limitations of Expert Support:

- Scalability and Accessibility: Human experts are a scarce resource. Their availability is limited, making large-scale deployment or on-demand support impractical for many organizations or individual designers.
- Subjectivity and Consistency: The quality and type of guidance can vary between different experts, and even from the same expert on different occasions, due to the inherent subjectivity of human interpretation and problem-solving. This can make consistent support challenging. However, it is not possible to say that the model would be more consistent than the user, since with just one word or a little more context it changes the direction of the response.

In essence, while LLMs are powerful idea generators and initial conceptual guides, particularly for novice users, they are not yet substitutes for the adaptive, context-rich, and careful evaluative support offered by human experts. The effectiveness of LLM support is heavily contingent on the user's domain knowledge and their ability to craft precise prompts and critically assess the model's output.

For complex, iterative design tasks, human expertise provides an invaluable layer of nuanced, experiential, and adaptive guidance. Future tools could benefit from combining the LLM's rapid generation capabilities with human expert oversight or more sophisticated mechanisms for contextual adaptation and the integration of implicit knowledge. This would move towards a collaborative intelligence model where LLMs augment human designers rather than merely providing rote answers.

4. Contributions to HCI

This research offers several significant contributions to the field of Human-Computer Interaction (HCI), particularly concerning the evolving role of Large Language Models (LLMs) in supporting design processes. Firstly, by providing an empirical comparison of LLM-driven support versus human expert guidance in data visualization design, this study sheds light on their respective strengths and limitations. It clarifies that, while

LLMs excel at rapid idea generation and structured initial guidance, especially for non-expert designers facing common design challenges, human experts remain invaluable for nuanced, contextualized, and adaptive support, often leveraging implicit knowledge and experiential insights that current LLMs lack. This comparative perspective contributes to a more informed understanding of how AI can augment, rather than replace, human intelligence in design.

Secondly, our work offers practical insights into the specific characteristics of effective prompts for design-oriented LLM interactions. By detailing the prompt construction process and observing user interactions, we highlight the need for careful consideration of role definition, desired output specificity, and the management of negative constraints. The observed tension between the LLM's tendency to provide direct answers and experienced designers' preference for meta-level guidance underscores crucial considerations for designing future human-AI co-creative systems.

Thirdly, a notable contribution lies in the development and adaptation of a comprehensive set of evaluation criteria for assessing AI-generated design guidance. By drawing upon established principles from communication theory (Grice's maxims) and human-computer interaction (criteria from [Kim et al. 2025]), and adapting them to the specific context of data visualization design, this study provides a valuable methodological framework. These refined criteria can serve as a robust tool for future research aiming to systematically evaluate the quality and utility of AI-powered design assistants across various domains, enhancing the rigor of HCI evaluations.

Finally, by focusing on data visualization design, a domain that requires both creativity and adherence to specific principles, our research provides a concrete case study for understanding AI's applicability in complex, knowledge-intensive design tasks. The findings extend beyond generic design assistance, offering specific implications for how LLMs can be integrated into specialized design workflows, potentially fostering more inclusive design practices by lowering the barrier to entry for complex visualization creation. These contributions collectively advance our understanding of effective human-AI collaboration in creative domains, paving the way for more intelligent and user-centered design tools.

5. Conclusion

At the outset of this study, we outlined four research questions — a main one and three subquestions — that guided our methodological process, focusing on how LLMs could support the data visualization design process, identify specific points of assistance, and aid non-specialist designers.

To answer these questions, we first investigated design activities commonly used in the data visualization process, characterizing them by their stages and whether they involved development or evaluation, as highlighted by works such as [McKenna et al. 2014] and [Parsons 2022]. This led to the selection of two familiar and frequently used activities: interview script construction and wireframe creation. The instructions for these activities were standardized for both ChatGPT and human expert guidance, allowing a direct comparison of the support provided by each group.

For the (textual) interview script construction, where the output was textual, participants, especially those with less design experience, reported good performance

from the model in terms of structure and question suggestions. This contrasted with experts, who provided more direct guidance on how to build the script rather than generating questions, often leading to a clearer fulfillment of prompt instructions (*e.g.*, avoiding direct answers).

For creating the (visual) wireframes, initial interactions with ChatGPT were less successful, often requiring multiple prompts for optimal results, even without asking him to build the visual result for the requests in the initial prompt of the planned interaction.

And when users asked ChatGPT to create a visual example for visually organizing information, the model did not respond well, first trying to describe it in text, which took a few more commands to generate a draft.

A notable difference here was the experts' ability to provide numerous practical examples and contextual insights, which were less prevalent in the model's textual responses. This observation suggests that, for oral interactions between non-experts and experts, the results may outperform those of written materials, especially in real-time two-way conversations.

Perhaps by sending more information to the chat, the response could be more helpful in the zero-shot prompt. In this study, the expert received more context than the chat because, in the free version, ChatGPT has a message size limit.

Another possible limitation for chat responses could be the prompt. However, as described in Section 3.1, the prompt is designed to assist the user as an assistant, not to provide a direct answer or product to the activity.

In conclusion, while LLMs are powerful tools for augmenting design processes, particularly in the initial phases, they are not yet a silver bullet for complex design challenges, nor a replacement for human expertise. The most promising future lies in synergistic human-AI collaboration, where LLMs handle generative tasks and provide structured inputs, while human designers exercise critical judgment, refine outputs, and provide the deep contextual understanding and ethical oversight necessary for high-quality, responsible design. Future work should explore hybrid systems that integrate AI assistance with embedded validation mechanisms and educational scaffolds to cultivate a more discerning and effective user base.

5.1. Limitations and Threats to Validity

This work has certain limitations that must be acknowledged. Firstly, focusing on only two design activities and a limited number of participants restricts the generalizability of findings regarding which activities are best supported, the recommended experience level for designers, and the optimal LLM for support.

Secondly, a significant concern pertains to the origin and verifiability of the information provided by LLMs. As the model draws from its training data with unrestricted sources, it can generate convincing yet erroneous content. Recommending a tool with an unknown information provenance to users lacking specific domain knowledge, who cannot discern between correct and incorrect information, is therefore not advisable.

5.2. Future Work

This study opens avenues for future research. Expanding investigations to include more diverse design activities, a larger and more varied participant pool (categorized by experience levels), and comparative analyses across different LLMs would provide a more comprehensive understanding of AI's role in supporting design. Furthermore, evaluating mixed-origin (LLM and expert) support and real-world, sustained interaction scenarios would offer deeper insights into practical integration and long-term impact.

6. Acknowledgments

The author gratefully acknowledges the assistance provided by Gemini (Google's AI model) for its support in translating portions of this text from Portuguese to English, contributing to the clarity and accessibility of the manuscript.

Referências

Alves, R. e Jardim Nunes, N. (2013). Towards a taxonomy of service design methods and tools. In Falcão e Cunha, J., Snene, M., e Nóvoa, H., editors, *Exploring Services Science*, page 215–229, Berlin, Heidelberg. Springer.

Bischof, N., Comi, A., e Eppler, M. J. (2011). Knowledge visualization in qualitative methods – or how can i see what i say? In *2011 15th International Conference on Information Visualisation*, page 371–376.

Carvalho, F. A. N. d. (2022). Compreende: um framework para a seleção de ferramentas no desenvolvimento de projetos de produtos digitais. *Repositório da Universidade Federal do Ceará*. Accepted: 2022-09-12T15:16:36Z.

Chen, N., Zhang, Y., Xu, J., Ren, K., e Yang, Y. (2025). Viseval: A benchmark for data visualization in the era of large language models. *IEEE Transactions on Visualization and Computer Graphics*, 31(1):1301–1311.

Grice, Kimball, J. P., Morgan, J. L., e Cole, P. (1975). *Syntax and semantics*. Academic press, Harcourt Brace Jovanovich, New York San Francisco London.

Kim, N. W., Ahn, Y., Myers, G., e Bach, B. (2025). How good is chatgpt in giving advice on your visualization design? *ACM Trans. Comput.-Hum. Interact.*

Kirk, A. A. (2016). *Data Visualisation: A Handbook for Data Driven Design*. Sage Publications, Los Angeles London New Delhi Singapore Washington DC Melbourne, 1ª edição edition.

Knaflic, C. N. (2015). *Storytelling with data: a data visualization guide for business professionals*. Wiley, Hoboken, New Jersey.

Kumar, V. (2013). *101 design methods: a structured approach for driving innovation in your organization*. Wiley, Hoboken, N.J.

McKenna, S., Mazur, D., Agutter, J., e Meyer, M. (2014). Design activity framework for visualization design. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2191–2200.

Munzner, T. (2014). *Visualization Analysis and Design*. AK Peters Visualization Series. CRC Press.

Nguyen, T. H., Waizenegger, L., e Techatassanasoontorn, A. A. (2022). "don't neglect the user!"–identifying types of human-chatbot interactions and their associated characteristics. *Information Systems Frontiers*, 24(3):797–838.

Parsons, P. (2022). Understanding data visualization design practice. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):665–675.

Rodrigues Catalano, J. V. e Rossi Lorenzi, B. (2023). Sem referências: o ChatGPT sob a perspectiva latouriana do duplo clique. *Revista Faz Ciência*, 25(41).

Stickdorn, M., Hormess, M., Lawrence, A., e Schneider, J., editors (2021). *This is service design doing: applying service design thinking in the real world ; a practitioners' handbook*. O'Reilly Media, Sebastopol, CA, 10. nachdr edition.

Van Someren, M., Barnard, Y. F., e Sandberg, J. (1994). The think aloud method: a practical approach to modelling cognitive. *London: AcademicPress*, 11(6).