Evaluating ChatGPT to Support Data Visualization Design

George M. Oliveira[1], **Simone D. J. Barbosa**[1]

[1]Departamento de Informática
Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio)
Rua Marquês de São Vicente, 225 / 4o andar RDC – 22451-900 – Rio de Janeiro – RJ – Brazil

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

Abstract. *Large language models (LLMs) can help retrieve information to answer questions, construct images and audio, and assist in complex activities such as data visualization design. The latter requires specific knowledge that can be found on the internet and therefore used to train LLMs. This work investigates the ability of ChatGPT to assist in data visualization design. We conduct a metrics-based evaluation of the model and plan to expand it to understand the views of users who create visualizations, whether they are experts or not.*

Resumo. *Grandes modelos de linguagem (LLMs) podem ajudar a recuperar as informações para responder perguntas, construir imagens e áudios, e auxiliar em atividades complexas como o design de visualização de dados. Este último requer conhecimentos específicos que podem estar disponíveis na internet e utilizados para treinar LLMs. Este trabalho investiga a capacidade do ChatGPT para auxiliar no design de visualização de dados. Conduzimos uma avaliação do modelo com base em métricas e planejamos expandi-la para entender a visão dos usuários criadores de visualização, sejam ou não especialistas.*

1. Motivação

O uso de grandes modelos de linguagem (LLMs) vem crescendo exponencialmente nos últimos anos. Modelos como o Gemini, Bing Chat e ChatGPT permitem a interação com o sistema por meio de um chat, interpretar a linguagem humana, em vários idiomas, e respondam da mesma forma. Esse processamento de linguagem natural (PLN) facilita o acesso da tecnologia a pessoas alfabetizadas e com algum letramento digital. Essas ferramentas evoluem para diminuir as barreiras existentes no uso: o GPT-4o compreende não apenas linguagem verbal escrita, mas também expressa por voz e imagem [Zhu et al. 2024]. O usuário pode recuperar as informações utilizadas para o treinamento do modelo; gerar textos, imagens e áudios; traduzir documentos; e realizar atividades repetitivas ou complexas utilizando o modelo. No entanto, algumas atividades complexas, como o design de visualizações de dados, ainda não foram muito exploradas. Se bem treinados, os LLMs podem auxiliar no design de visualização de dados, produzindo insumos, avaliando visualizações ou gerando o produto final.

Este trabalho objetiva investigar a capacidade do ChatGPT,[1] modelo produzido pela OpenAI, em apoiar o design de visualização de dados por pessoas especialistas ou não. Essa análise pode revelar limitações e oportunidades de melhoria em *prompts* e possíveis ajustes finos no modelo.

[1]https://openai.com/chatgpt/

Trata-se de um estudo interdisciplinar, que se baseia em conhecimentos das áreas de design, visualização e de LLMs. Essa interdisciplinaridade permite que avaliemos a interação desses usuários com o sistema por meios de técnicas pertinentes a área.

Este artigo apresenta a fundamentação teórica (seção 2), com os conceitos e trabalhos que encontramos na literatura para nortear a metodologia (seção 3), definindo os procedimentos éticos (seção 4) necessários para os resultados preliminares (seção 5) e finalizando com a proposta de cronograma para este trabalho (seção 6).

2. Fundamentação teórica e trabalhos relacionados

Buscamos trabalhos relacionados nas áreas de design, visualização de dados e LLMs, visando a responder duas perguntas principais: RQ1: "Como LLMs podem ajudar no design de visualização de dados?"; e RQ2: "Como analisar as respostas do modelo no processo de design?" Em geral, encontramos investigações sobre LLMs relacionadas à produção, avaliação e melhoria da informação produzida pelos modelos em diversas áreas, por exemplo, saúde [Wei et al. 2024], educação [Alexandra Farazouli and McGrath 2024] e empresarial [Chenfu et al. 2024].

Encontramos diversos trabalhos sobre o uso de LLMs e de PLN para criar visualizações resultando na implementação em código [Maddigan and Susnjak 2023, Sun et al. 2010, Narechania et al. 2021]. Nesses trabalhos, o usuário inseria textualmente o seu objetivo para a criação da visualização e o sistema retornava o código pronto para ser executado. No entanto, queremos investigar o *processo* de design de visualizações: identificação do problema, ideação, avaliação da ideia e entrega do produto final.

LLMs podem analisar textos e gerar uma avaliação, mesmo que não tão precisa quanto no caso de realização de um cálculo [Rodrigues Catalano and Rossi Lorenzi 2023]. Tais capacidades podem auxiliar no processo de design, ajudando a responder perguntas, avaliar opções e gerar resultados. Alguns estudos utilizam técnicas comuns da área de Interação Humano-Computador (IHC) para entender investigar a usabilidade [Mulia et al. 2023, Skjuve et al. 2023], como questionários mais abertos sobre os modelos [Chang et al. 2024]. Para o nosso trabalho, focaremos nas respostas do modelo, principalmente com relação ao conteúdo e sua qualidade, seguindo algumas métricas definidas na literatura, analisando também métricas com usuários especialistas ou não.

Kim et al. (2024) investigaram a qualidade do modelo para responder perguntas sobre visualização de dados [Kim et al. 2024] e para avaliar as respostas do modelo ChatGPT focado no processo de design de visualizações. Utilizaram perguntas do VisGuides[2] para que o modelo gerasse opções de resposta. Eles classificaram as respostas do modelo em seis grupos, utilizados como métricas para avaliar as respostas do modelo, são eles: *cobertura*: quão completa é a resposta em relação às partes da pergunta; *foco*: quão bem o modelo mantém o objetivo na resposta, em relação à pergunta; *amplitude*: capacidade do modelo em dar respostas além do necessário de maneira complementar; *clareza*: quão fácil é entender as respostas; *profundidade*: quão explicativa é a resposta sobre a escolha do tipo da visualização; e *aplicabilidade*: capacidade de aplicar a resposta no contexto informado na pergunta [Kim et al. 2024].

[2]Disponível em: `https://visguides.org/`. Acessado em: 22 de junho de 2024

O nosso trabalho se difere do de [Kim et al. 2024] ao expandir a avaliação do modelo para todo o processo de design de visualizações de dados. Aplicamos inicialmente as mesmas métricas para avaliar as respostas do modelo em todas as etapas do processo, mas iremos expandir a avaliação considerando outras bases teóricas.

3. Metodologia de pesquisa

Nosso trabalho consiste em avaliar as respostas do modelo sobre o design de visualização de dados. Isso envolve identificar as etapas do processo de design, usar o modelo para responder as demandas, identificar um método de avaliação para o modelo e avaliar as respostas geradas com usuários e com métricas existentes na literatura. Iteramos entre gerar informações e avaliá-las. Esse processo iterativo reflete um processo de design típico, não linear e mediante refinamentos sucessivos.

Até o momento, analisamos como o modelo responde perguntas iniciais e avalia as respostas que ele próprio gerou. Solicitamos que o modelo gerasse perguntas sobre visualização utilizando um formato de cenário de uso que incluía um perfil de usuário e uma demanda (denominada "desejo"), segundo o formato abaixo: *"Eu, como um [usuário], preciso [desejo], qual seria o melhor tipo de visualização para fazer isso?"*

Para explorar a capacidade do modelo em gerar diversas opções, para cada pergunta solicitamos 3 opções de resposta obedecendo a seguinte estrutura: *tipo* - nome do tipo da visualização; *descrição* - resumo da aplicação no contexto da pergunta; e *variáveis* - variáveis necessárias para construir a visualização com base na pergunta.

Na fase de avaliação automática das respostas, passamos como entrada para o modelo o que ele gerou: uma lista com 15 perguntas e 3 opções de resposta para cada, e pedimos que o modelo avaliasse cada uma das opções, retornando uma nota de 1 (não recomendada) a 7 (totalmente recomendada) e afirmando o motivo da nota (Figura 1).

Além disso, o autor principal deste trabalho realizou uma avaliação manual das respostas do modelo nas duas fases, utilizando as seis métricas encontradas no trabalho de [Kim et al. 2024], em escalas de 1 (muito ruim) a 5 (muito bom).

Pergunta	Respostas			Avaliação	
	Tipo de visualização	Descrição	Variáveis	Motivo	Nota
Eu, como um gerente de marketing, preciso entender as tendências de consumo dos clientes em diferentes regiões, qual seria o melhor tipo de visualização para fazer isso?	Mapa de calor	Utilizar um mapa de calor permite representar visualmente os dados de consumo por região, destacando áreas de maior e menor demanda. É eficaz para identificar padrões geográficos rapidamente e ajustar estratégias de marketing localizadas.	['Região como coordenadas geográficas', 'Consumo como valor numérico']	O mapa de calor recebe a maior nota devido à sua alta eficiência em mostrar padrões geográficos de consumo de uma forma visualmente intuitiva e imediata. Ele facilita a identificação de regiões críticas para tomada de decisões estratégicas, sendo mais eficaz que os gráficos de barras ou de bolhas dada a possibilidade de visualizar os dados em seu contexto geográfico.	7
	Gráfico de barras	Um gráfico de barras pode mostrar o consumo dos clientes em várias regiões de forma clara e direta. É útil para comparações rápidas entre regiões e para apresentar dados de forma organizada, facilitando a interpretação e o planejamento.	['Região como categoria na eixo X', 'Consumo como valores no eixo Y']	O gráfico de barras é uma boa opção para comparar o consumo entre diferentes regiões de forma clara e direta, mas não fornece a mesma imersão visual ou contexto geográfico que um mapa de calor oferece. Ele é menos efetivo para visualizar padrões espaciais, o que é crucial para um gerente de marketing que precisa entender tendências geográficas.	5
	Gráfico de bolhas	O gráfico de bolhas pode ser usado para representar três dimensões de dados: região, consumo e um terceiro fator, como tempo ou categoria de produto. Isso ajuda a visualizar complexidades nos dados de consumo que não são imediatamente aparentes com outros métodos.	['Região no eixo X', 'Consumo no eixo Y', 'Tamanho da bolha representando uma terceira variável, como tempo ou categoria']	O gráfico de bolhas adiciona uma terceira dimensão de análise, o que pode ser útil para insights mais complexos. No entanto, pode ser menos intuitivo e mais difícil de interpretar comparado ao mapa de calor, especialmente para um público não técnico. Além disso, a representação geográfica não é tão imersiva quanto no mapa de calor, o que limita sua eficácia para análise regional de consumo.	4

Figura 1. Formato de resposta após todo o processo

4. Procedimentos éticos

O trabalho ainda não envolveu usuários, mas no futuro avaliaremos partes do processo de design de visualização de dados com pessoas especialistas ou não na área. Esse projeto foi submetido e aprovado pela Câmara de Ética em Pesquisa da PUC-Rio, que considerou

adequados os procedimentos planejados, haja vista os riscos mínimos aos participantes e os benefícios potenciais para as comunidades de pesquisa e prática em visualização de dados. O uso de questionários para avaliar a resposta do modelo com especialistas na área e a observação de uso do modelo com pessoas não especialistas, técnicas comuns à área de IHC, podem ajudar a avaliar o avanço do estudo.

5. Resultados Preliminares

Pelo estudo realizado, identificamos que o ChatGPT é capaz de dar respostas textuais com clareza e foco quando recebe perguntas mais diretas e completas. O melhor desempenho do modelo ocorreu quando pedimos para ele avaliar uma lista de opções de visualizações. As respostas do modelo para as **perguntas iniciais** obtiveram os seguintes resultados: o *foco* foi a métrica com maior média (3.84). Consideramos que as respostas produzidas foram *claras* (3.77) e fáceis de serem entendidas por usuários não especialistas. As métricas de *cobertura* (3.6), *amplitude* (3.2) e *profundidade* (3.4) foram avaliadas de modo diferente, uma vez que no trabalho de [Kim et al. 2024] era avaliado o grupo de respostas para a pergunta e não cada resposta individualmente. A nota mais baixa foi para *profundidade* porque as respostas sobre como usar a visualização foram superficiais. Quando pedimos para o modelo **avaliar as respostas** para cada pergunta, julgamos que ele manteve o *foco* (4.56) e foi *claro* (4.47) ao explicar o motivo de cada nota no contexto, confirmando a noção de que LLMs geram bons textos. A *cobertura* (4.22), *profundidade* (4.18) e *amplitude* (4.09) também apresentaram notas altas.

Ao fim dessa avaliação preliminar, consideramos o modelo efetivo para gerar e avaliar respostas, conforme as métricas de avaliação de [Kim et al. 2024]. Essa efetividade foi aumentando ao decorrer do processo de pesquisa, à medida que refinamos os procedimentos e os comandos (*prompts*) fornecidos para o modelo. O resultado esperado deste trabalho são instrumentos para avaliar o quão bem um LLM (no caso, o ChatGPT) pode apoiar o design de visualizações de dados, identificando pontos que ajudem a usuários não especialistas a utilizar essas ferramentas. Esperamos ter um catálogo de comandos que ajudem a interagir de maneira eficaz com o modelo, como um guia de uso de LLMs para o design de visualizações.

6. Cronograma

Com o protótipo atual, podemos avaliar os primeiros passos de um usuário criador de visualização de dados durante o processo de designapós a identificação do perfil do usuário e do objetivo da visualização: (i) realizar perguntas sobre qual visualização utilizar no contexto e (ii) pedir ao modelo para julgar qual a melhor em meio a uma lista de opções. Os próximos passos previstos são:

ago/2024: identificar as etapas do processo de design de visualização de dados;
ago/2024: identificar os métodos que apoiam cada etapa;
set/2024: aplicar cada método com o apoio do modelo;
out/2024: analisar as respostas do modelo com as métricas;
nov/2024: analisar o uso por pessoas não especialistas; e
dez/2024: defesa da dissertação.

Referências

Alexandra Farazouli, Teresa Cerratto-Pargman, K. B.-L. and McGrath, C. (2024). Hello GPT! Goodbye home examination? An exploratory study of AI chatbots impact on university teachers' assessment practices. *Assessment & Evaluation in Higher Education*, 49(3):363–375.

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang, Y., Chang, Y., Yu, P. S., Yang, Q., and Xie, X. (2024). A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.*, 15(3).

Chenfu, S., Shoji, Y., Yamamoto, T., Tanaka, K., and Dürst, M. J. (2024). Generating experiential descriptions and estimating evidence using generative language model and user products reviews. In *2024 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 254–261.

Kim, N. W., Myers, G., and Bach, B. (2024). How good is ChatGPT in giving advice on your visualization design? *arXiv:2310.09617*.

Maddigan, P. and Susnjak, T. (2023). Chat2VIS: Generating data visualizations via natural language using ChatGPT, codex and GPT-3 large language models. *IEEE Access*, 11:45181–45193.

Mulia, A. P., Piri, P. R., and Tho, C. (2023). Usability analysis of text generation by ChatGPT OpenAI using system usability scale method. *Procedia Computer Science*, 227:381–388. 8th International Conference on Computer Science and Computational Intelligence (ICCSCI 2023).

Narechania, A., Srinivasan, A., and Stasko, J. (2021). NL4DV: A toolkit for generating analytic specifications for data visualization from natural language queries. *IEEE Transactions on Visualization & Computer Graphics*, 27(02):369–379.

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

Skjuve, M., Følstad, A., and Brandtzaeg, P. B. (2023). The user experience of chatgpt: Findings from a questionnaire study of early users. In *Proceedings of the 5th International Conference on Conversational User Interfaces*, CUI '23, page 1–10, New York, NY, USA. Association for Computing Machinery.

Sun, Y., Leigh, J., Johnson, A., and Lee, S. (2010). Articulate: A semi-automated model for translating natural language queries into meaningful visualizations. In Taylor, R., Boulanger, P., Krüger, A., and Olivier, P., editors, *Smart Graphics*, pages 184–195, Berlin, Heidelberg. Springer Berlin Heidelberg.

Wei, Q., Yao, Z., Cui, Y., Wei, B., Jin, Z., and Xu, X. (2024). Evaluation of chatgpt-generated medical responses: A systematic review and meta-analysis. *Journal of Biomedical Informatics*, 151:104620.

Zhu, N., Zhang, N., Shao, Q., Cheng, K., and Wu, H. (2024). OpenAI's GPT-4o in surgical oncology: Revolutionary advances in generative artificial intelligence. *European Journal of Cancer*, 206:114132.