

USO DE INTELIGÊNCIA GENERATIVA NA AVALIAÇÃO HEURÍSTICA DE USABILIDADE DE UMA PÁGINA WEB: potencialidades e limitações

THE USE OF GENERATIVE ARTIFICIAL INTELLIGENCE IN THE HEURISTIC USABILITY EVALUATION OF A WEB PAGE: capabilities and limitations

Ranna Redlley Silva Oliveira | Sandra de Albuquerque Siebra

<https://doi.org/10.21747/21836671/pag25a1>

Resumo: Este estudo analisa o uso da Inteligência Artificial Generativa como ferramenta de apoio à avaliação de usabilidade de uma página *web*, com base nas heurísticas de Nielsen. Adotou-se uma abordagem quali-quantitativa, de caráter descritivo e exploratório, que utilizou revisão bibliográfica e a implementação de um teste empírico, no qual foi desenvolvido um *prompt* avaliativo para avaliação heurística da página inicial de um repositório digital. O *prompt* foi submetido às ferramentas ChatGPT, Copilot e Gemini e foram verificadas as convergências e divergências dos resultados apresentados. Os resultados evidenciam que todas as GenAI foram capazes de realizar avaliações heurísticas da interface, porém apresentaram diferenças no detalhamento descritivo da avaliação e em sua precisão, na consideração ou não do perfil do usuário descrito no *prompt* e na atribuição do grau de severidade dos problemas encontrados. O ChatGPT destacou-se por apresentar análises mais completas, contextualizadas e próximas de uma avaliação humana especializada. O Gemini apresentou resultados com menos detalhamento e contextualização que o ChatGPT, focando em aspectos estruturais e visuais da interface. E o Copilot demonstrou inconsistências interpretativas e equívocos na avaliação, que foi apresentada de forma objetiva, sem muitas explicações. Conclui-se que as GenAI apresentam potencial como ferramenta de pré-diagnóstico em avaliações heurísticas de usabilidade, desde que utilizadas de forma crítica e supervisionada.

Palavras-chave: Heurísticas de Nielsen; Inteligência Artificial Generativa; *Prompt*; Usabilidade.

Abstract: This study analyzes the use of Generative Artificial Intelligence as a support tool for evaluating the usability of a web page based on Nielsen's heuristics. A qualitative-quantitative, descriptive, and exploratory approach was adopted, using a literature review and the implementation of an empirical test in which an evaluative prompt was developed to conduct a heuristic evaluation of a digital repository's homepage. The prompt was submitted to the ChatGPT, Copilot, and Gemini tools, and the convergences and divergences among the results obtained were analyzed. The findings show that all GenAIs could perform heuristic evaluations of the interface; however, they differed in the descriptive detail and precision of the evaluations, in whether or not they considered the user profile described in the prompt, and in the assignment of severity ratings for the identified problems. ChatGPT stood out for providing more complete, contextualized analyses that were closer to a specialized human evaluation. Gemini produced results with less detail and contextualization than ChatGPT, focusing mainly on the structural and visual aspects of the interface. Copilot, in turn, demonstrated interpretative inconsistencies and evaluation errors, presenting the analysis in a more objective manner and with fewer explanations. It is concluded that GenAIs have potential as a pre-diagnostic tool in heuristic usability evaluations, since they are used critically and under supervision.

Keywords: Nielsen's heuristics; Generative Artificial Intelligence; Prompt; Usability.

Introdução

Segundo Nielsen (1993), a usabilidade relaciona-se à facilidade de uso de uma interface; ao quão fácil é para os usuários realizarem tarefas básicas na primeira vez que interagem com um sistema; à rapidez com que os usuários executam suas tarefas (eficiência); à facilidade de memorização; a uma baixa taxa de erros cometidos e à facilidade de se recuperar deles;

além da satisfação durante o uso da interface. Ela configura-se como um atributo central de qualidade, ao contribuir para o desenvolvimento de interfaces mais simples, funcionais e intuitivas, impactando diretamente a experiência do usuário. (GAO, JIA e YIN, 2024).

Nesse sentido, questões de usabilidade precisam ser consideradas no desenvolvimento e na manutenção de ambientes digitais (MAYHEW, 1999), tornando-se relevante avaliá-los de maneira sistemática, correta e coerente. Neste estudo, compreende-se o ambiente digital, com base em Forrest e Wexler (2023), como espaços virtuais acessíveis por meio da conexão à Internet, onde ocorrem transações, comunicações e interações mediadas por tecnologias digitais, e que podem ser acessados por dispositivos como computadores, consoles e *smartphones*. A avaliação de usabilidade permite detectar problemas de interação nesses ambientes (como sistemas digitais, *sites*, aplicativos, redes sociais, entre outros) e propor melhorias orientadas à experiência do usuário, de modo a garantir que o *design* atenda às necessidades cognitivas, perceptivas e operacionais do público-alvo (NIELSEN, 1993; BARBOSA e SILVA, 2010).

Entre os métodos de avaliação de usabilidade, encontram-se as heurísticas de Nielsen (1993; 2005; 2024), amplamente utilizadas por sua aplicabilidade prática e pela capacidade de identificar problemas recorrentes de interação (CYBIS, BETIOL e FAUST, 2017). De fato, embora tenham sido definidas em 1993 e várias propostas de novas heurísticas e de variações das heurísticas de Nielsen tenham surgido ao longo do tempo, o estudo realizado por Krawiec e Dudycz (2020) mostrou que, apesar do passar dos anos, as heurísticas clássicas de Nielsen continuam relevantes, universais e as mais utilizadas por aqueles que primam pela qualidade das interfaces.

Em geral, a avaliação heurística de usabilidade é realizada manualmente por especialistas, o que pode demandar tempo, conhecimento técnico e recursos específicos. Com o avanço da Inteligência Artificial Generativa (GenAI), surge a possibilidade de utilizá-la para apoiar a identificação de problemas de usabilidade, ampliando a agilidade e a abrangência do processo avaliativo. De fato, ferramentas de GenAI, tais como ChatGPT, Copilot e Gemini, são capazes de interpretar e produzir linguagem natural com alto nível de sofisticação e podem ser orientadas por prompts estruturados para gerar análises com base de critérios previamente estabelecidos, como as heurísticas de Nielsen (2024).

Tal perspectiva inaugura um campo de investigação ainda pouco explorado, que envolve analisar se a GenAI pode contribuir para a identificação de problemas de usabilidade de forma ágil e coerente (ZHANG e WANG, 2024). Embora a literatura recente aponte para o uso da IA em tarefas de *design* e interação, ainda são escassas as investigações que examinam o uso da GenAI como ferramenta de apoio na avaliação heurística de interfaces (DUAN *et al.*, 2024).

Diante disso, este estudo tem como objetivo analisar o uso da Inteligência Artificial Generativa como ferramenta de apoio à avaliação de usabilidade de uma página *web*, com base nas heurísticas de Nielsen. A ideia-chave é ampliar as possibilidades de identificar e corrigir falhas que comprometem a interação humano-computador, contribuindo para a construção de ambientes digitais mais intuitivos e eficazes.

Inteligência Artificial Generativa e a criação de prompts

A Inteligência Artificial (IA) pode ser compreendida como um campo científico voltado ao desenvolvimento de sistemas capazes de executar tarefas que, tradicionalmente, exigiriam inteligência humana, como raciocínio, aprendizado e tomada de decisão (MCCARTHY, 2007). Ao longo das décadas, a IA evoluiu de modelos baseados em regras para abordagens fundamentadas em dados e em aprendizado de máquina, consolidando-se como uma tecnologia central na sociedade contemporânea (RUSSELL e NORVIG, 2020).

Nesse contexto, destaca-se a GenAI, considerada uma evolução da IA tradicional, por sua capacidade não apenas de analisar ou classificar informações, mas também de gerar novos conteúdos, como textos, imagens e códigos (HESSEL e LEMES, 2023). Diferentemente dos modelos convencionais, a GenAI opera com base em grandes volumes de dados e em modelos probabilísticos, o que permite a geração de respostas dinâmicas e contextualizadas.

O funcionamento dessas tecnologias está diretamente relacionado aos avanços no Processamento de Linguagem Natural (PLN) e no uso de modelos de linguagem de grande escala (LLMs), que aprendem padrões a partir de extensos conjuntos de dados e passam a prever sequências linguísticas de forma coerente (COUTINHO, 2022; VASWANI *et al.*, 2017). Esse processo envolve etapas como pré-treinamento, ajuste fino e inferência, o que possibilita que sistemas como o ChatGPT, o Copilot e o Gemini respondam a comandos em linguagem natural.

Além de seu caráter tecnológico, a GenAI tem suscitado reflexões sobre criatividade, autoria e produção do conhecimento. Conforme discutido por Santaella e Kaufman (2024), essas ferramentas não apenas reproduzem informações, mas também reorganizam conteúdos existentes, criando novas possibilidades de sentido. Nesse cenário, a interação entre o humano e a máquina passa a configurar um modelo híbrido de produção intelectual, no qual a IA atua como suporte ao processo cognitivo.

No âmbito acadêmico e profissional, a GenAI vem sendo aplicada em diversas atividades, como a produção de textos, a análise de dados, a geração de imagens e o apoio à tomada de decisão, o que exige competências específicas dos usuários para seu uso crítico e responsável (TRINDADE e OLIVEIRA, 2024). Entre essas competências, destaca-se a capacidade de formular instruções adequadas para orientar uma GenAI na criação de conteúdo, os chamados *prompts*, que orientam o comportamento dos modelos e influenciam diretamente a qualidade das respostas geradas.

Nesse sentido, o objetivo do *prompt* é fornecer à GenAI as informações necessárias e suficientes para produzir um resultado pertinente à solicitação (TECHTARGET, 2025). O *prompt* vai definir o tipo, o formato e o nível de detalhamento das respostas. Dessa forma, a formulação adequada de *prompts* torna-se um fator determinante para a eficácia no uso dessas tecnologias e para maximizar a utilidade e a precisão das respostas (BARBOSA FILHO, 2025).

A eficácia do *prompt* não reside apenas no comando em si, mas também na capacidade do interlocutor humano de fornecer contextos ricos e mitigar ambiguidades, reduzindo, assim,

a possibilidade de alucinações¹ ou de respostas genéricas por parte do modelo (DUARTE, 2025). *Prompts* vagos ou pouco estruturados tendem a gerar respostas genéricas, imprecisas ou incompletas, enquanto instruções claras, específicas e contextualizadas aumentam significativamente a qualidade, a relevância e a utilidade das respostas produzidas pela GenAI. Nesse sentido, a chamada Engenharia de *Prompts* trata do processo de elaborar e refinar um *prompt* específico e detalhado (GEORGIA TECH, 2024).

A literatura científica apresenta diferentes tipos de *frameworks* de *prompt*, tais como o GPEI (*Goal* → *Prompt* → *Evaluation* → *Iteration*) (DORTA HERNÁNDEZ, 2025), o *Prompt Pattern Catalog* (MENEZES et al., 2024; YAO et al., 2024) e o *Zero-shot Learning* (SHORT e SHORT, 2023). Estas técnicas impactam a qualidade das respostas, atuando como *frameworks* operacionais na engenharia de *prompts* (GEORGIA TECH, 2024). Além deles, Greg Brockman desenvolveu um *framework* de *prompts* para otimizar a comunicação com modelos de GenAI (DUARTE, 2025).

O modelo de Brockman não é uma metodologia formal, e sim um conjunto de diretrizes práticas destinadas a melhorar a eficiência imediata das interações com GenAI. Assim, Brockman propõe um modelo voltado à praticidade e ao uso aplicado, menos sistemático, mas ainda assim eficiente para fins operacionais (DUARTE, 2025). Esse modelo baseia-se em quatro elementos principais, conforme se vê na Fig. 1.

Fig. 1 – Modelo de *prompt* de Greg Brockman



Fonte: Adaptado de Oliveira (2025), com uso do ChatGPT.

¹ Alucinação, no contexto da GenAI, refere-se à produção de respostas aparentemente coerentes, mas factualmente incorretas, inventadas ou que não derivam de evidências reais nos dados de treinamento, surgindo quando o modelo “preenche lacunas” com informações plausíveis, porém falsas. Essa falha decorre de limitações estatísticas dos modelos de linguagem, que operam por predição probabilística de palavras e não por compreensão semântica, levando a inferências equivocadas quando confrontados com dados ambíguos, insuficientes ou fora do seu escopo (HUANG et al., 2024).

O primeiro elemento, considerado o núcleo da estrutura, é a definição clara e específica do objetivo, que orienta a tarefa a ser executada. Ele orienta e direciona o comportamento do modelo de acordo com a intenção do usuário, garantindo que a resposta gerada mantenha coerência com a finalidade proposta. O segundo elemento é a especificação do formato de retorno, que define como a resposta deve ser apresentada. Esse componente atua como um modelo de referência (*blueprint*), garantindo que o conteúdo produzido pela IA seja organizado, coerente, funcional e de fácil interpretação (DUARTE, 2025).

O terceiro elemento engloba avisos ou restrições que orientarão o modelo quanto a comportamentos e conteúdos a serem evitados e, assim, contribuirão para a produção de respostas mais adequadas, precisas e dentro do escopo definido. E, por fim, o quarto elemento é o fornecimento de contexto (informações de fundo e dados complementares), para ampliar a capacidade do modelo de gerar respostas mais relevantes e alinhadas às necessidades do usuário (DUARTE, 2025).

A adoção dessa estrutura contribui para reduzir ambiguidades, minimizar erros e elevar a qualidade das respostas, tornando a interação com a GenAI mais eficiente e controlada. Além disso, ao estruturar adequadamente o *prompt*, o usuário passa a exercer um papel mais ativo no processo de geração de conteúdo, atuando como mediador entre a capacidade técnica do modelo e o objetivo da tarefa.

É importante destacar que, embora essa estrutura torne o *framework* uma ferramenta com potencial de precisão, sua aplicação é sempre contextual e deve ser adaptada às particularidades de cada ambiente de GenAI, já que nem todos requerem o mesmo nível de detalhamento ou padronização na construção de *prompts*.

Verifica-se, a partir do que foi apresentado, que a GenAI mostra-se uma tecnologia com potencial de aplicação em diferentes contextos, incluindo a avaliação de usabilidade, pois pode possibilitar a geração de análises automatizadas com base em critérios previamente definidos, como as heurísticas de Nielsen (2024).

Avaliação de usabilidade por meio das heurísticas de Nielsen

De acordo com a norma ISO 9.241-11:2018, a usabilidade é definida como a medida em que um produto pode ser utilizado por usuários específicos para alcançar objetivos, em um contexto de uso definido (INTERNATIONAL..., 2018). Configura-se, portanto, como um fator essencial para a eficácia e a eficiência no uso de *websites*, aplicativos e sistemas digitais, pois influencia diretamente a forma como os usuários interagem com as funcionalidades disponíveis. Para Steve Krug (2008), quando os elementos de uma interface são bem organizados e a navegação é simples, o usuário consegue entender rapidamente o que deve fazer, sem precisar parar para pensar ou interpretar excessivamente as opções disponíveis.

Para verificar se questões de usabilidade estão sendo consideradas adequadamente em um ambiente digital, deve-se realizar uma avaliação de usabilidade, que consiste em um processo sistemático voltado à identificação de problemas que possam comprometer a interação entre o usuário e a interface, a fim de analisar em que medida um sistema pode ser utilizado com eficácia, eficiência e satisfação em um determinado contexto (INTERNATIONAL..., 2018). A norma ISO 9.241-11:2018 aponta duas formas principais

de avaliação de usabilidade: a avaliação baseada em desempenho e resultados, envolvendo amostras de usuários reais, como os testes de usabilidade; e a avaliação baseada em atributos, que envolve a inspeção de conformidade com atributos ou características, tal como a avaliação heurística.

Entre os métodos de avaliação heurística de usabilidade mais consolidados estão as heurísticas de usabilidade propostas por Jakob Nielsen (CHO *et al.*, 2024). As chamadas heurísticas de Nielsen (1994; 2005; 2024) consistem em princípios gerais que orientam a análise da interação entre o usuário e o sistema, auxiliando na identificação de falhas e oportunidades de melhoria. No Quadro 1, encontra-se um resumo das heurísticas e, com base nas publicações do Nielsen Norman Group (2025) e de Nielsen (2024), algumas perguntas que podem servir de guia para sua aplicação, o que é útil na configuração de um *prompt* para uma GenAI.

Quadro 1 – Questionamentos referentes às heurísticas de Nielsen

Heurísticas	Guia de questionamentos
1-Visibilidade do estado do sistema	Após realizar uma ação, é perceptível ao usuário a mudança de estado ou <i>status</i> do sistema? Após as ações do usuário, o sistema apresenta um <i>feedback</i>? São fornecidos ao usuário indicadores de estados ativos/inativos e de mudança de estado? O sistema informa claramente quando uma tarefa está em andamento? Os elementos ativos/selecionados estão visualmente destacados? Quando há uma espera longa, o sistema comunica adequadamente o tempo estimado ou o progresso da ação em andamento?
2-Compatibilidade entre o sistema e o mundo real	O vocabulário utilizado na interface é adequado ao contexto do usuário? Os ícones, imagens e rótulos dos elementos da interface são adequados ao contexto do usuário? Os ícones e símbolos indicam claramente sua função, mesmo sem legenda? A aparência dos elementos sugere corretamente como interagir com eles? (ex.: parecer clicável, arrastável, preenchível) Ou seja, a <i>affordance</i> é aparente? As informações estão organizadas e estruturadas de forma natural, em uma ordem coerente?
3-Controle e liberdade do usuário	O usuário consegue transitar livremente (ir e voltar) entre diferentes campos de formulários ou <i>itens</i> de menu? O <i>design</i> permite que os usuários voltem uma etapa no processo? Os <i>links</i> são facilmente detectáveis? Os usuários podem cancelar uma ação facilmente? Há suporte para desfazer e refazer ações? Os usuários conseguem cancelar uma operação em andamento? É claro e fácil localizar a opção de fechar janelas? O usuário pode pausar e retomar tarefas longas?

	O sistema permite personalizar a interface? Há alguma forma de confirmar ações que possam levar à perda de dados?
4-Consistência e padrões	O sistema utiliza símbolos e padrões de navegação comuns à Internet, de modo que o usuário não precise aprender novos significados para coisas que já conhece? É mantido um padrão de <i>design</i> consistente? Os elementos de navegação e os botões de ação mantêm uma localização padronizada entre as páginas do sistema? Os campos de formulário utilizam máscaras de formatação automática para facilitar o preenchimento de dados, como datas, CPF, CEP e telefone? O tamanho visual dos campos de entrada de dados condiz com a quantidade de caracteres que o usuário precisa digitar?
5-Prevenção de erros	Os campos de entrada de dados expressam como devem ser preenchidos (ex.: uso de máscaras de formatação, indicação de obrigatoriedade e tamanho adequado do campo)? Há alguma forma de confirmar ações que possam levar à perda de dados? Há mecanismos para recuperar dados esquecidos (ex.: "Esqueci minha senha")? O sistema possui salvamento automático ou exibe uma mensagem se o usuário tentar sair sem salvar?
6-Reconhecimentos em vez de memorização	O <i>design</i> mantém informações importantes visíveis para que os usuários não precisem memorizá-las? O sistema fornece pré-visualização de escolhas do usuário, quando pertinente (ex.: fontes, cores, imagens e temas)? O caminho de navegação percorrido pelo usuário é exibido como um <i>breadcrumb</i>? O usuário consegue acessar facilmente um histórico de ações ou de documentos recentes? O usuário precisa memorizar os dados de uma tela para usar na próxima? As simbologias usadas são claras o suficiente para serem reconhecidas sem esforço? Os controles relacionados a uma tarefa estão agrupados e refletem a sequência de ações da tarefa?
7-Eficiência e flexibilidade de uso	O sistema oferece teclas de atalho? Existem múltiplas formas de realizar a mesma ação (ex.: menu ou botão rápido, via teclado e com o mouse, por opção de menu ou por tecla de atalho)? O sistema permite configurar ações para acesso rápido? Há recursos de preenchimento automático ou sugestões inteligentes (<i>autocomplete</i>)? A interface permite alguma forma de personalização (ex.: escolha de tema, de fonte, de funcionalidades para compor um menu de acesso rápido, entre outras)? O sistema é disponibilizado em mais de um idioma?

8 -Estética e <i>design</i> minimalista	Há um número excessivo de elementos visuais no <i>design</i> da página? Há "ruído visual", elementos em movimento ou informações irrelevantes que distraem o usuário da tarefa principal? O <i>layout</i> utiliza espaços em branco e agrupamentos para facilitar a leitura, evitando a sobrecarga cognitiva?
9 -Ajuda aos usuários a reconhecerem, diagnosticarem e recuperar-se de erros	As mensagens de erro são escritas em linguagem simples, sem códigos técnicos? O texto utiliza frases afirmativas e diretas, evitando construções escritas na negativa ou ambíguas (ex.: "não pode não ser...")? A mensagem oferece uma solução construtiva para o problema encontrado ou um <i>link</i> direto para resolver o problema? O design visual (ex.: cor vermelha, ícones de alerta) deixa claro que ocorreu um erro e onde ele ocorreu?
10 -Ajuda e documentação	A documentação de ajuda é facilmente localizada? A ajuda é fornecida no contexto adequado, no momento em que o usuário a solicita? A ajuda é focada nas tarefas do usuário ("como fazer")? Existem formatos variados de ajuda, como tutoriais, vídeos, FAQs ou <i>tours</i> guiados?

Fonte: Baseado em Nielsen Norman Group (2025).

Na avaliação heurística, após a identificação de um problema, recomenda-se atribuir um grau de severidade, pois isso permite priorizar as correções e direcionar esforços aos problemas mais críticos (NIELSEN..., 2025). Este grau varia de 0 a 4, sendo: 0 - sem relevância, não é um problema de usabilidade; 1 - problema cosmético, quando não compromete o uso e pode ser corrigido apenas se houver tempo; 2 - problema simples, que causa algum incômodo e tem baixa prioridade de correção; 3 - problema grave, que prejudica significativamente a interação e tem alta prioridade de correção; e 4 - problema catastrófico, quando impede o uso e demanda correção imediata.

O grau de severidade é calculado com base em três fatores: frequência (se o problema é comum ou raro), impacto (quão fácil é para o usuário superar o problema) e persistência (se o problema é momentâneo, pois os usuários podem aprender a superá-lo) (NIELSEN, 1994). Porém, Nielsen (1994) e o Nielsen Norman Group (2025) não apresentam um método exato para calcular o grau de severidade de um determinado problema, deixando a atribuição a cargo do julgamento do avaliador, o que a torna, de certa forma, subjetiva. Por fim, ao final do processo avaliativo, também podem ser sugeridas melhorias com base nos problemas identificados.

Metodologia

Esta pesquisa caracteriza-se por uma abordagem quali-quantitativa, de natureza descritiva e exploratória (GIL, 2002). Além disso, foram adotados procedimentos não empíricos e empíricos. Como procedimentos não empíricos, foram utilizados a revisão de literatura e a análise documental, com o objetivo de construir o problema de pesquisa, elaborar o corpus teórico e estabelecer critérios para a análise posterior realizada com as ferramentas de

GenAI. A revisão de literatura concentrou-se em livros e artigos científicos que abordavam os temas centrais desta investigação: avaliação de usabilidade, GenAI e princípios de engenharia de *prompts*. Na análise documental, foram examinadas normas e materiais técnicos, com destaque para a ISO 9.241-11:2018, que padroniza conceitos de usabilidade, seus componentes e formas de avaliação de usabilidade. Além de documentos institucionais referentes ao Repositório Digital ATTENA² da Universidade Federal de Pernambuco (UFPE), Brasil, no qual foram aplicadas as heurísticas de Nielsen. Esta consulta permitiu compreender o propósito institucional do repositório, suas funcionalidades e seu contexto de uso, oferecendo elementos para a etapa empírica. Escolheu-se analisar a página inicial de um repositório, visto que se considera que os repositórios institucionais são ambientes digitais relevantes em instituições acadêmicas de destaque em nível nacional, como a UFPE.

A etapa empírica envolveu a criação e a submissão de um prompt avaliativo (ver Quadro 2) a três ferramentas de GenAI (ChatGPT - GPT-5.1, Copilot e Gemini - Flash 2.5), com o intuito de avaliar sua capacidade de identificar problemas de usabilidade em uma página Web. Por se tratar de um experimento preliminar, que se pretende expandir posteriormente e que visava à construção do prompt, bem como à experimentação dele para verificar como seria interpretado pelas GenAI, foi selecionada como objeto de análise apenas a página inicial da versão desktop do repositório digital ATTENA-UFPE.

As heurísticas de Nielsen foram escolhidas para uso no processo avaliativo devido à sua ampla adoção e à eficácia na identificação de problemas de interface. Dessa forma, o *prompt* avaliativo foi elaborado de acordo com o *framework* de Greg Brockman, estruturado segundo as heurísticas de Nielsen (1993; 2005, 2024) e as diretrizes do Nielsen Norman Group (2025), tomando como base os questionamentos explicitados no Quadro 1. Para contribuir para uma maior padronização na atribuição do grau de severidade pelas GenAI, uma vez que não há um método exato recomendado pelos autores supracitados, considerou-se o grau de severidade como a média aritmética dos três fatores a ele relacionados (frequência, impacto e persistência), e cada fator receberia um valor a partir de uma escala Likert de 0 a 4.

Ressalta-se que o *prompt* inicialmente traçado passou por uma série de refinamentos e ajustes, a partir dos testes com as GenAI que foram sendo realizados, em busca de maior completude dos resultados, até se chegar ao *prompt* especificado no Quadro 2, que também pode ser considerado uma contribuição deste trabalho. A criação e submissão do *prompt* foram realizadas em novembro de 2025.

² Criado em 2014 e consolidado em 2019 em parceria com o Sistema Integrado de Bibliotecas (SiB), o repositório ATTENA tem como objetivo reunir, preservar e disponibilizar a produção acadêmica e científica da Universidade Federal de Pernambuco em um único ambiente digital (Repositório Digital da UFPE, 2022).

Quadro 2 – Prompt Avaliativo Utilizado

1) Persona (Papel)

Atue como especialista sênior em Interação Humano-Computador (IHC), Usabilidade e Avaliação Heurística, com conhecimento na condução de auditorias de experiência do usuário (UX).

2) Tarefa Principal (Objetivo)

Realize uma avaliação heurística de usabilidade completa da página inicial do Repositório ATTENA/UFPE (<https://attena.ufpe.br/>), aplicando rigorosamente as Dez Heurísticas de Nielsen (1994; 2005), considerando o contexto da interação, o fluxo de navegação, os elementos de design, o vocabulário utilizado no repositório, as imagens e os elementos da interface. Avalie apenas a parte da página inicial visível aos usuários, sem inferir funcionalidades ou elementos gráficos e de design que não estejam presentes nesta página. Nenhuma hipótese sobre comportamento do sistema deve ser assumida sem respaldo visual explícito.

3) Formato de Retorno

Os resultados obtidos devem estar organizados em um quadro, com as seguintes cinco colunas:

1. Nome da heurística analisada

2. Flag indicador de se a heurística foi atendida, utilizando os valores “Sim”, quando a heurística for atendida, e “Não”, quando a heurística não for atendida.)

3. Descrição do problema identificado, caso exista (flag indicador = “Sim”)

4. Indicação do grau de severidade, utilizando uma escala de 0 a 4, sendo 0 um problema sem importância ou não relacionado à usabilidade e 4 um problema crítico de usabilidade. Essa classificação deve basear-se na combinação de três fatores principais: frequência (se o problema ocorre com frequência ou raramente), impacto (grau de dificuldade para o usuário superar o problema) e persistência (se o problema é momentâneo e o usuário pode aprender a superá-lo ou se é contínuo ao longo do uso). Com relação aos graus: o grau 0 indica que não há, de fato, um problema de usabilidade. O grau 1 indica questões apenas cosméticas, que não exigem correção imediata e podem ser ajustadas quando houver tempo disponível. O grau 2 corresponde a problemas simples, de baixa prioridade. O grau 3 representa problemas graves, que demandam alta prioridade de correção por causarem frustração ou dificuldade significativa ao usuário. Por fim, o grau 4 caracteriza problemas catastróficos que devem ser corrigidos com brevidade. Para calcular este grau de severidade, utilize a fórmula a seguir:

Frequência (F) refere-se a com que frequência o problema ocorre: 0 = nunca, 1 = raramente, 2 = às vezes, 3 = frequentemente, 4 = sempre.

Impacto no usuário/Tarefa (I) - refere-se ao grau de dificuldade o usuário superar o problema: 0 = nenhum, 1 = baixo, 2 = moderado, 3 = alto, 4 = crítico (impede a conclusão) e

Persistência (P): custo de contornar/superar o problema: 0 = inexistente (não afeta), 1 = fácil, 2 = possível com esforço, 3 = difícil, 4 = sem solução. Fórmula final calculada pela média aritmética simples arredondada:

Severidade $S = \text{round}((F + I + P) / 3)$.

5. Indicação de sugestões de melhoria para o problema identificado.

4) Avisos (Regras e Restrições)

Siga as regras e restrições definidas a seguir:

- A análise deve considerar exclusivamente os elementos presentes na página inicial do Repositório ATTENA
- Não gere interpretações subjetivas ou comportamentos não observáveis (ex.: “provavelmente há um menu oculto”).
- Use linguagem técnica, clara e objetiva, como se estivesse elaborando um relatório técnico profissional.
- Utilize terminologia profissional de UX e IHC.
- Analise os elementos com foco nas heurísticas de usabilidade de Nielsen, primando pela clareza comunicacional, pelas affordances e pela consistência visual e semântica.
- Não utilize julgamentos estéticos vagos; foque no impacto funcional.
- A análise deve ser inteiramente baseada na parte visível ao usuário do repositório.
- Se a heurística não puder ser avaliada na página inicial, declare explicitamente esse fato.
- Utilize o conceito original de heurísticas, conforme o Nielsen Norman Group.
- Atribua o grau de severidade de cada problema encontrado considerando a frequência com que o problema ocorre, o impacto do problema na atividade dos usuários e na persistência/reincidência do problema, conforme a fórmula descrita para preenchimento do quadro de resultados.
- Na avaliação, leve em consideração o perfil do usuário, conforme explicitado no contexto relevante.

5) Contexto relevante

Considere que o perfil do usuário é composto majoritariamente por estudantes de graduação e pós-graduação de universidades públicas brasileiras, bem como por pesquisadores, docentes e técnicos administrativos. Trata-se de um público heterogêneo em termos de letramento digital, abrangendo tanto usuários com baixa familiaridade com a tecnologia quanto aqueles com competências avançadas. Além disso, há variação quanto à origem geográfica, incluindo indivíduos de grandes centros urbanos e usuários de regiões interioranas, que podem ter acesso mais restrito a dispositivos e recursos tecnológicos.

E, para cada uma das heurísticas usadas na avaliação, analise-as considerando as seguintes questões:

Heurística 1. Visibilidade do status do sistema.

Após realizar uma ação, é perceptível ao usuário a mudança de estado ou status do sistema?

Após as ações do usuário, o sistema apresenta um feedback?

São fornecidos ao usuário indicadores de estados ativos/inativos e de mudança de estado?

O sistema informa claramente quando uma tarefa está em andamento?

Os elementos ativos/selecionados estão visualmente destacados? Quando há uma espera longa, o sistema comunica adequadamente o tempo estimado ou o progresso da ação em andamento?

Heurística 2. Correspondência entre o sistema e o mundo real.

O vocabulário utilizado na interface é adequado ao contexto do usuário?

Os ícones, imagens e rótulos dos elementos da interface são adequados ao contexto do usuário?

Os ícones e símbolos indicam claramente sua função, mesmo sem legenda?

A aparência dos elementos sugere corretamente como interagir com eles? (ex.: parecer clicável, arrastável, preenchível) Ou seja, a affordance é aparente?

As informações estão organizadas e estruturadas de forma natural, em uma ordem coerente?

Heurística 3. Controle e liberdade do usuário.

O usuário consegue transitar livremente (ir e voltar) entre diferentes campos de formulários ou itens de menu?

O design permite que os usuários voltem uma etapa no processo? Os links são facilmente detectáveis?

Os usuários podem cancelar uma ação facilmente?

Há suporte para desfazer e refazer ações?

Os usuários conseguem cancelar uma operação em andamento?

É claro e fácil localizar a opção de fechar janelas?

O usuário pode pausar e retomar tarefas longas?

O sistema permite personalizar a interface?

Há alguma forma de confirmar ações que possam levar à perda de dados?

Heurística 4. Consistência e padrões.

O sistema utiliza símbolos e padrões de navegação comuns à internet, de modo que o usuário não precise aprender novos significados para coisas que já conhece?

É mantido um padrão de design consistente?

Os elementos de navegação e os botões de ação mantêm uma localização padronizada entre as páginas do sistema?

Os campos de formulário utilizam máscaras de formatação automática para facilitar o preenchimento de dados, como datas, CPF, CEP e telefone?

O tamanho visual dos campos de entrada de dados condiz com a quantidade de caracteres que o usuário precisa digitar?

Heurística 5. Prevenção de erros.

Os campos de entrada de dados expressam como devem ser preenchidos (ex.: uso de máscaras de formatação, indicação de obrigatoriedade e tamanho adequado do campo)?

Há alguma forma de confirmar ações que possam levar à perda de dados?

Há mecanismos para recuperar dados esquecidos (ex.: "Esqueci minha senha")? O sistema possui salvamento automático ou exibe uma mensagem se o usuário tentar sair sem salvar?

Heurística 6. Reconhecimento em vez de memorização.

O design mantém informações importantes visíveis para que os usuários não precisem memorizá-las?

O sistema fornece pré-visualização de escolhas do usuário, quando pertinente (ex.: fontes, cores, imagens e temas)?

O caminho de navegação percorrido pelo usuário é exibido como um breadcrumb?

O usuário consegue acessar facilmente um histórico de ações ou de documentos recentes?

O usuário precisa memorizar os dados de uma tela para usar na próxima?
As simbologias usadas são claras o suficiente para serem reconhecidas sem esforço?
Os controles relacionados a uma tarefa estão agrupados e refletem a sequência de ações da tarefa?

Heurística 7. Flexibilidade e eficiência de uso.

O sistema oferece teclas de atalho?
Existem múltiplas formas de realizar a mesma ação (ex.: menu ou botão rápido, via teclado e com o mouse, por opção de menu ou por tecla de atalho)?
O sistema permite configurar ações para acesso rápido?
Há recursos de preenchimento automático ou sugestões inteligentes (autocomplete)?
A interface permite alguma forma de personalização (ex.: escolha de tema, de fonte, de funcionalidades para compor um menu de acesso rápido, entre outras)?
O sistema é disponibilizado em mais de um idioma?

Heurística 8. Estética e design minimalista.

Há um número excessivo de elementos visuais no design da página?
Há "ruído visual", elementos em movimento ou informações irrelevantes que distraem o usuário da tarefa principal?
O layout utiliza espaços em branco e agrupamentos para facilitar a leitura, evitando a sobrecarga cognitiva?

Heurística 9. Ajudar a reconhecer, diagnosticar e recuperar erros.

As mensagens de erro são escritas em linguagem simples, sem códigos técnicos?
O texto utiliza frases afirmativas e diretas, evitando construções escritas na negativa ou ambíguas (ex.: "não pode não ser...")?
A mensagem oferece uma solução construtiva para o problema encontrado ou um link direto para resolver o problema?
O design visual (ex.: cor vermelha, ícones de alerta) deixa claro que ocorreu um erro e onde ele ocorreu?

Heurística 10. Ajuda e documentação.

A documentação de ajuda é facilmente localizada?
A ajuda é fornecida no contexto adequado, no momento em que o usuário a solicita?
A ajuda é focada nas tarefas do usuário ("como fazer")?
Existem formatos variados de ajuda, como tutoriais, vídeos, FAQs ou tours guiados?

6) Finalização

Após o quadro de resultados, apresente um texto analítico explicando: os principais problemas identificados; quais são as heurísticas mais críticas (maior grau de severidade); e os pontos em que houve limitações para avaliar a usabilidade por meio das heurísticas de Nielsen.

Fonte: Elaboração própria (2026).

Os dados obtidos com a submissão do *prompt* foram coletados organizados, incluindo a identificação de problemas, a verificação de heurísticas atendidas ou não atendidas, o grau de severidade e as sugestões de melhoria, além de sínteses produzidas pelas IA sobre a

avaliação. Depois, eles foram comparados considerando os seguintes critérios: Nível de Detalhamento do Resultado da avaliação fornecido pelas GenAI; Heurísticas com maior severidade; Corretude da Avaliação (em relação à avaliação previamente realizada pelos autores); Completude da Avaliação e Heurísticas Consideradas Plenamente Atendidas. Também foram pontuadas as convergências e divergências entre os resultados obtidos pelas ferramentas utilizadas.

Por fim, verificou-se a similaridade entre as análises de GenAI e uma avaliação humana realizada no repositório ATTENA-UFPE por Paiva (2022), cujo trabalho foi premiado em nível nacional, considerando o nível de detalhamento e organização dos resultados e se havia semelhança ou não na atribuição dos graus de severidade (em consonância com o problema encontrado), pois Paiva atribuiu grau de severidade a seu critério, de forma subjetiva. Destaca-se que a avaliação de Paiva (2022) não se limitou à página inicial, mas para este estudo, apenas considerou-se a parte da avaliação feita a ela relacionada.

Resultados e discussões

Conforme descrito na metodologia, o *prompt* inicialmente elaborado foi submetido às ferramentas de GenAI (ChatGPT, Copilot e Gemini) diversas vezes; assim, foram identificadas melhorias que poderiam ser nele implementadas para aprimorar os resultados obtidos. Até que se chegou ao *prompt* do Quadro 2 e, após sua submissão, todas as ferramentas produziram um quadro com os resultados, seguido de um texto analítico que resumia a avaliação, conforme especificado no próprio *prompt*.

Todas as GenAI utilizadas foram capazes de realizar a avaliação heurística da página inicial do Repositório ATTENA/UFPE. A comparação dos resultados das avaliações, segundo os critérios especificados na metodologia, pode ser observada no Quadro 3.

Quadro 3 – Resultados obtidos com a submissão do *prompt* às três ferramentas de GenAI

Aspecto Avaliado	ChatGPT	Copilot	Gemini
Nível de detalhamento do resultado da avaliação	Detalhado, explicativo e centrado no impacto da usabilidade na jornada do usuário na primeira página do repositório. Esta foi a avaliação que mais considerou o perfil do usuário apresentado no contexto do <i>prompt</i> .	Sintético, objetivo, descrição superficial (especialmente comparada às outras GenAI) e, com forte ênfase em erros críticos.	Estruturado, técnico e focado na arquitetura visual e informacional da página do repositório.
Heurísticas não atendidas com maior severidade (segundo a GenAI)	H1 (Visibilidade), H6 (Reconhecimento), H10 (Ajuda) – severidade 3	H1 (Visibilidade), H5 (Prevenção), H7 (Flexibilidade), H9 (Recuperação de erros) – severidade 3	H6 (Reconhecimento), H10 (Ajuda) – severidade 3

Corretude da Avaliação	Não foi percebido problema <i>a priori</i>	Houve equívocos da ferramenta nas heurísticas 7 e 10.	Não foi percebido problema <i>a priori</i>
Completude da avaliação	Não conseguiu verificar se a heurística 9 é atendida ou não, identificando a mesma como “Não avaliável”, pois depende da ocorrência de erros, não visíveis na página inicial. Trouxe uma sugestão futura	Mesmo sem reconhecer corretamente todos os elementos da interface, avaliou todas as heurísticas.	Não conseguiu verificar se a heurística 9 é atendida ou não, identificando a mesma como “N/A”, pois depende da ocorrência de erros, não visíveis na página inicial.
Heurísticas que foram consideradas atendidas plenamente	Considerou as heurísticas 2, 4 plenamente atendidas e a heurística 8 parcialmente atendida.	Considerou as heurísticas 4 e 8 plenamente atendidas.	Considerou as heurísticas 1, 2, 4 e 8 plenamente atendidas

Fonte: Elaboração própria (2026).

O ChatGPT apresentou o desempenho mais consistente, com análises mais completas e estruturadas, e uma avaliação coerente com as heurísticas propostas por Nielsen (1994). Ele descreveu os problemas com profundidade técnica; justificou o grau de severidade com base no impacto na experiência do usuário, considerando o perfil delineado no *prompt*; relacionou as falhas visuais e de navegação à usabilidade prática da interface; além de explicitar as limitações metodológicas decorrentes da análise restrita à página inicial, como a ausência de observação de fluxos dinâmicos. Além disso, propôs melhorias alinhadas aos problemas identificados. O ChatGPT não realizou a avaliação da heurística 9, como observado no Quadro 3, uma vez que a página inicial não apresenta mensagens de erro, porém, destaca-se que a ferramenta reconheceu esta limitação. Esse comportamento está bem condizente com o estudo de Duan *et al.* (2024), que destaca o potencial dos grandes modelos de linguagem realizarem avaliações heurísticas com elevado grau de precisão inferencial, sobretudo quando o *prompt* fornece um contexto bem estruturado.

O Copilot realizou uma análise mais curta, objetiva, sem muitas justificativas e orientada a ações operacionais, característica compatível com sua arquitetura voltada a assistentes de produtividade. Esta foi a ferramenta de menor precisão, que apresentou inconsistência na identificação de problemas e interpretou incorretamente alguns elementos da interface. Por exemplo, na análise da heurística 10, o Copilot apontou ausência de acesso à ajuda na página inicial, quando na verdade, a opção de ajuda está presente no menu do repositório e, na análise da heurística 7, apontou a falta de suporte multilíngue visível, porém, o repositório possui uma opção de menu para troca do idioma da página, o que não foi registrado pelo Copilot, mas foi pelo ChatGPT. De fato, foi a única ferramenta que apresentou equívocos na análise das heurísticas e que não reconheceu alguns elementos da página avaliada. Além disso, o Copilot atribuiu severidades, sem apresentar justificativas para tais classificações com base na literatura de usabilidade, como foi feito pelo ChatGPT. Esse é um comportamento que Bender *et al.* (2021) caracterizam como tendência a interpretações superficiais e generalizações excessivas, que ocorrem quando o modelo não mobiliza contexto suficiente.

Adicionalmente, os problemas reais identificados pelo Copilot foram descritos de forma superficial e não contextualizados teoricamente, sem considerar noções fundamentais, como a carga cognitiva, o perfil do usuário (fornecido no “despejo do contexto” do *prompt*) e as *affordances*, discutidas por Krug (2008) e por Barbosa e Silva (2010). Essa ausência de rigor conceitual e os equívocos na análise remetem ao fenômeno identificado por Huang *et al.* (2024) como *hallucination by oversimplification*, no qual os modelos de GenIA tendem a preencher lacunas com respostas genéricas e pouco fundamentadas. Dessa forma, observa-se que o Copilot identifica alguns problemas relevantes na interface avaliada; entretanto, sua análise carece de estrutura crítica, maior precisão e melhor fundamentação teórica.

O Gemini apresentou um desempenho intermediário entre a profundidade conceitual do ChatGPT e a incompletude do pragmatismo operacional do Copilot. A análise realizada foi apresentada de forma estruturada, detalhando problemas e apresentando sugestões de melhoria sem superficialidade, quando comparado ao Copilot. Sua análise focou em aspectos visuais e estruturais da interface, como a organização da informação, a densidade informacional, a distribuição espacial dos componentes e a hierarquia visual. A ferramenta fundamentou cada grau de severidade atribuído com justificativas coerentes e precisão técnica, garantindo um diagnóstico sintético e objetivo de todos os elementos visíveis da interface. Não conseguiu analisar se a heurística 9 foi atendida. Essa ênfase em aspectos perceptivos e estruturais da interface é coerente com estudos recentes sobre *design* em sistemas de GenAI, como os de Zhang e Wang (2024), que destacam a importância da clareza visual e da organização espacial para a efetividade da interação em ambientes digitais cada vez mais complexos. A avaliação realizada não apresenta justificativas para as severidades atribuídas aos problemas, tendo, inclusive, deixado alguns problemas identificados sem apresentação do grau de severidade.

Vale mencionar que apenas a heurística 4 (Consistência e padrões) foi considerada atendida plenamente por todas as ferramentas, havendo concordância sobre o porquê de ter sido atendida, sendo pontuada a padronização de menus, tipografia e organização visual conforme convenções da *web*, além da consistência entre páginas e alinhamento com outros sistemas da UFPE (instituição a qual pertence o repositório), facilitando a navegação do usuário. Outro ponto que vale destacar é que a heurística 8 (Estética e design minimalista) foi considerada plenamente atendida pelo Copilot e pelo Gemini, que destacaram o layout limpo, o uso adequado de espaços em branco e a ausência de excesso de ruído visual, pontuando o layout atual como adequado. Porém, o ChatGPT apontou que tal heurística foi atendida parcialmente, pois identificou uma leve sobrecarga informacional na página inicial, sugerindo melhorar a hierarquia visual e priorização das informações.

Dessa forma, os resultados indicam que, embora todas as ferramentas sejam capazes de identificar problemas de usabilidade, há diferenças principalmente: 1) no detalhamento descritivo da avaliação, sendo a do ChatGPT a avaliação mais completa, com justificativa para as avaliações realizadas; 2) na precisão da avaliação, tendo o Copilot apresentado uma avaliação imprecisa e com algumas incoerências entre a heurística e o problema identificado; 3) na consideração do perfil de usuário descrito, sendo o ChatGPT o que demonstrou ter feito uso do perfil descrito no *prompt* na descrição de sua análise; 4) na atribuição do grau de severidade, pois houve uma convergência técnica, com coincidência exata entre o Gemini, o Copilot e o ChatGPT em 50% das heurísticas (H3, H4, H5, H6 e H10), mas nas outras heurísticas houve divergências na atribuição do grau de severidade,

tendo o Copilot uma tendência a atribuir severidades maiores. Um ponto a ser ressaltado é que todas as ferramentas apontaram a limitação algumas heurísticas não conseguirem ser avaliadas, devido a inexistência de fluxos dinâmicos na página inicial do repositório. Nesse sentido, remete-se ao que defendem Barbosa e Silva (2010), de que uma avaliação heurística eficaz requer a integração dos aspectos cognitivos, estruturais e perceptivos da interface. E ao afirmado por Nielsen (1993;1994) e Mayhew (1999), de que a avaliação heurística exige interpretação contextualizada e julgamento qualitativo, algo que ainda é reproduzido de forma desigual pelas GenAI, especialmente quando dependem da informação textual presente no *prompt*.

Goodfellow, Bengio e Courville (2016) argumentam que modelos treinados com maior diversidade textual apresentam maior capacidade de generalizar e de aplicar conceitos abstratos. Isso se materializa na forma como o ChatGPT articula as heurísticas de Nielsen, a ergonomia cognitiva e a arquitetura da informação com os achados empíricos. O que coaduna com as afirmações de Anjos (2024) e de Duan *et al.* (2024), segundo os quais as GenAI podem assumir funções avaliativas, inclusive em tarefas tradicionalmente humanas, como a análise heurística, desde que os modelos apresentem capacidade interpretativa adequada. E que a precisão dessas avaliações depende diretamente da arquitetura da GenAI, do *corpus* de treinamento e do tipo de tarefa solicitada. Dessa forma, a análise da execução do *prompt* nas três GenAI dá indícios que quanto maiores a capacidade inferencial e o volume de treinamento linguístico, como é o caso do ChatGPT, melhor o resultado da avaliação heurística, pois percebeu-se uma maior aderência às heurísticas de usabilidade e ao arcabouço teórico da IHC, sendo o ChatGPT o modelo que mais se aproxima de uma avaliação heurística fundamentada.

Por fim, como forma de verificar a similaridade entre as análises automatizadas e uma avaliação feita por um especialista, se comparou as avaliações heurísticas realizadas pela GenAI e a avaliação de Paiva (2022), que também utilizou as heurísticas de Nielsen para avaliar o mesmo repositório e apresentou como resultado um quadro equivalente ao solicitado no *prompt* e um texto descritivo sobre a avaliação. De modo geral, a avaliação heurística realizada por Paiva (2022) demonstrou maior detalhamento na análise dos problemas encontrados, articulando os problemas com fundamentos teóricos da usabilidade e da ergonomia cognitiva, seguindo de perto as recomendações de Nielsen (1994; 2024), além de considerar a experiência dos usuários em toda a análise. Entre as GenAI, o ChatGPT foi a ferramenta que mais se aproximou desta avaliação humana, ao identificar uma maior variedade de problemas; ao detalhar e organizar os resultados, justificando-os à luz da literatura pertinente, considerando aspectos visuais, estruturais e cognitivos da interface na avaliação. O ChatGPT também deu indícios de considerar a experiência do usuário descrita no *prompt*.

Considerações Finais

Os achados indicam que as GenAIs são capazes de identificar problemas de usabilidade, embora haja diferenças no detalhamento descritivo da avaliação e em sua precisão, na consideração ou não do perfil do usuário descrito no *prompt* e na atribuição do grau de severidade dos problemas encontrados. Entre as ferramentas, o ChatGPT apresentou desempenho mais robusto, com uma avaliação mais completa e contextualizada, aproximando-se de uma avaliação humana por especialista. Já o Copilot e o Gemini

apresentaram limitações, resultando em avaliações com alguns pontos incompletos e imprecisos.

Verificou-se que a qualidade e completude das análises estavam diretamente relacionada ao nível de estruturação do *prompt*, por isso foi necessário realizar vários testes com variações no *prompt*, incluindo, modificando ou retirando elementos, até se chegar a versão final aqui apresentada. Por isso, considera-se que o *prompt* apresentado também é uma contribuição deste trabalho. Assim como o teste de qual ferramenta o executa com melhor precisão e completude nos resultados.

Ressalta-se que este estudo é um teste inicial do uso das GenAI para realização da avaliação heurística, que focou na construção do *prompt* avaliativo e na submissão dele à GenAI para refinamentos, além da identificação da GenAI mais adequada para este tipo de avaliação. A limitação da análise à página *web* inicial do repositório fez com que a avaliação de algumas heurísticas ficasse limitada, pois dependiam de fluxos e ações dinâmicas, relacionadas à navegação pelo ambiente digital.

Como trabalho futuro, pretende-se ampliar a análise para considerar as interfaces e o fluxo de navegação das principais atividades realizadas no repositório, tais como a tela de autodepósito de uma produção e a navegação e uso da busca dentro do repositório. Além disso, sugere-se como trabalhos futuros utilizar diferentes *frameworks* de construção de *prompts* para a avaliação heurística, a fim de comparar os resultados obtidos; assim como pode-se verificar a execução do *prompt* deste estudo em GenAI diferentes das utilizadas.

Conclui-se que os resultados dão indícios de que o uso de GenAI é uma alternativa promissora para apoiar a avaliação heurística de usabilidade, sobretudo como ferramenta de pré-diagnóstico de usabilidade, o que pode contribuir para o desenvolvimento de interfaces mais intuitivas, que possam promover uma boa experiência para seus usuários.

Referências bibliográficas

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS

2011 *NBR ISO 9.241-11: Requisitos ergonômicos para trabalho de escritório com computadores – Parte 11: Orientações sobre usabilidade*. Rio de Janeiro: ABNT, 2011.

BARBOSA, Simone Diniz Junqueira; SILVA, Bruno Santana da

2010 *Interação humano-computador*. Rio de Janeiro: Elsevier, 2010.

BARBOSA FILHO, Luiz Henrique

2025 Como usar engenharia de prompt para negócios. *Analisemacro*. [Em linha]. (20 jan. 2025). Disponível em: <https://analisemacro.com.br/inteligenciaartificial/como-usar-engenharia-de-prompt-para-negocios/>.

BASTIEN, Jean-Marie Christian; SCAPIN, Dominique

1993 *Ergonomic criteria for the evaluation of human-computer interfaces*. [Em linha]. Rocquencourt: INRIA, 1993. Disponível em: <https://inria.hal.science/inria-00070012/file/RT-0156.pdf>.

CHOO, Chun Wei

2006 *A Organização do conhecimento: como as organizações usam a informação para criar significado, construir conhecimento e tomar decisões*. São Paulo: Editora Senac, 2006.

COUTINHO, João Gabriel

2022 *Um Estudo sobre Sistemas de Recomendação baseados em transformers*. [Em linha]. São Carlos, 2022. 33 p. Disponível em:

https://bdta.abcd.usp.br/directbitstream/aaf7111e-9c33-4540-931c-12d1cfe78aab/Joao+Gabriel+Coutinho_TCC_MBA_JOAO_COUTINHO_206611.pdf.

Monografia - MBA in Artificial Intelligence and Big Data - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo.

CYBIS, Walter; BETIOL, Adriana Holtz; FAUST, Richard

2017 *Ergonomia e usabilidade: conhecimentos, métodos e aplicações*. 3ª ed. São Paulo: Novatec, 2017.

DORTA HERNÁNDEZ, Joel Aday

2025 *Inteligencia Artificial Generativa guiada por ontologias: un estudio de ingeniería del prompt em modelos de evaluación económica en salud*. [Em linha]. San Cristóbal de La Laguna 2025. 83 f. Disponível em:

<https://riull.ull.es/xmlui/bitstream/handle/915/42526/Inteligencia%20artificial%20generativa%20guiada%20por%20ontologias%20Un%20estudio%20de%20ingenieria%20del%20prompt%20en%20modelos%20de%20evaluacion%20economica%20en%20salud.pdf?sequence=1&isAllowed=y>.

Trabalho de Conclusão de Curso em Ingeniería Informática – Escuela Superior de Ingeniería y Tecnología, Universidad de La Laguna.

DUAN, Peitong [et al.]

2024 Exploring the potential of large language models for heuristic evaluation of user interfaces. In CHI CONFERENCE ON HUMANFACTORSIN COMPUTING SYSTEMS -CHI '24 (2024) – *Proceedings...* [Em linha]. Honolulu: ACM, 2024. Disponível em: <https://people.eecs.berkeley.edu/~bjoern/papers/duan-heuristic-chi2024.pdf>.

DUARTE, Roberto Dias

2025 O Framework de prompts de Greg Brockman : um guia completo para interações eficazes com IA. *RDD10+*, [Em linha]. (24 mar. 2025). Disponível em: <https://www.robertodiasduarte.com.br/o-framework-de-prompts-de-greg-brockman-um-guia-completo-para-interacoes-eficazes-com-ia/>.

GAO, Jie; JIA, Wenjing; YIN, Jun

2024 Exploring smartphone user interface experience-sharing behavior: design perception and motivation-driven mechanisms through the SOR model. *Sustainability*. [Em linha]. 16:15, (2024) 6.670. Disponível em: <https://www.mdpi.com/2071-1050/16/15/6670>. DOI: 10.3390/su16156670.

GEORGIA TECH

2024 *Prompt Engineering: The art of getting what you need from Generative AI*. [Em linha]. Atlanta: Georgia Tech Ivan Allen College, 2024. Disponível em: <https://iac.gatech.edu/featured-news/2024/02/AI-prompt-engineering-ChatGPT>.

GIL, Antônio Carlos

2002 *Como elaborar projetos de pesquisa*. 4ª ed. São Paulo: Atlas, 2002.

GOODFELLOW, Ian; BENGIO, Yoshua; COURVILLE, Aaron
2016 *Deep Learning*. Cambridge: MIT Press, 2016.

HESSEL, Ana Maria Di Grado; LEMES, David
2023 Criatividade da inteligência artificial generativa. *TECCOGS: Revista Digital de Tecnologias Cognitivas*. [Em linha]. 28 (2023) 119-130. Disponível em: <https://revistas.pucsp.br/index.php/teccogs/article/view/67075/45082>.

HUANG, Lei [et al.]
2024 A Survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*. [Em linha]. 1:1 (jan. 2024) artigo 1. Disponível em: <https://arxiv.org/abs/2311.05232>.

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION
2018 *ISO 9.241-11:2018: Ergonomics of human-system interaction. Part 11: Usability: Definitions and concepts*. [Em linha]. London: ISO, 2018. Disponível em: <https://www.iso.org/standard/63500.html>.

KRAWLEC, Lukasz; DUDYCZ, Helena
2020 A Comparison of heuristics applied for studying the usability of websites. *Procedia Computer Science*. [Em linha]. 176:1 (2020) 3.571-3.580. Disponível em: <https://doi.org/10.1016/j.procs.2020.09.029>.

KRUG, Steve
1993 *Não me faça pensar!: uma abordagem de bom senso à usabilidade na web*. 2ª ed. Rio de Janeiro: Alta Books, 2025, cap. 1, p. 16-20.
1ª ed.: 2008.

MAYHEW, Deborah J.
1999 *The Usability engineering lifecycle: a practitioner's handbook for user interface design*. San Francisco: Morgan Kaufmann, 1999.

MCCARTHY, John
2007 *What is Artificial Intelligence?* [Em linha]. Califórnia: Stanford University, Computer Science Department, 2007. Disponível em: <https://www-formal.stanford.edu/jmc/whatisai.pdf>.

MENEZES, Leonardo Mendes Lacerda de [et al.]
2024 Empreendedor melhorado pela IA: criando propostas de valor com IA generativa. In SIMPÓSIO DE INOVAÇÃO, TECNOLOGIA E EMPREENDEDORISMO – SITE, 32º, São Paulo, 2024 – *Anais...* [Em linha]. São Paulo: ANPAD, 2024. Disponível em: <https://anpad.com.br/uploads/articles/130/approved/4cce0644e37052e4c750f194dd5cccfdf.pdf>.

NIELSEN, Jakob
2024 *10 Usability heuristics for user interface design*. [Em linha]. [S. l.]: Nielsen Norman Group, April, 2024. Disponível em: <https://www.nngroup.com/articles/ten-usability-heuristics/>.

NIELSEN, Jakob
2005 *Ten usability heuristics for user interface design*. [Em linha]. [S. l.]: Nielsen Norman Group, 2005. Disponível em: <https://www.nngroup.com/articles/ten-usability-heuristics/>.

NIELSEN, Jakob

1994 *How to rate the severity of usability problems*. [Em linha]. [S. l.]: Nielsen Norman Group, 1 nov. 1994. Disponível em: <https://www.nngroup.com/articles/how-to-rate-the-severity-of-usability-problems/>.

NIELSEN, Jakob

1993 *Usability Engineering*. San Francisco: Morgan Kaufmann, 1993.

NIELSEN NORMAN GROUP

2025 *Heuristic evaluation workbook*. [Em linha]. 2025. Disponível em: https://media.nngroup.com/media/articles/attachments/Heuristic_Evaluation_Workbook_1_Fillable.pdf.

NORVIG, Peter; RUSSELL, Stuart

2020 *Inteligência Artificial: uma abordagem moderna*. 4ª ed. Porto Alegre: Ed. LTC, 2020.

OLIVEIRA, Ranna Redlley Silva

2025 *Contribuições da Inteligência Artificial Generativa para a avaliação de usabilidade: a criação de prompts avaliativos*. [Em linha]. Recife, 2025. Disponível em: https://repositorio.ufpe.br/bitstream/123456789/67498/1/TCC_RannaRedlleySilvaOliveira.pdf.

Trabalho de Conclusão de Curso em Gestão da Informação - Universidade Federal de Pernambuco.

PAIVA, Vítor Heitor de

2022 *Avaliação de usabilidade da interface do repositório digital ATTENA da UFPE utilizando heurísticas de Nielsen*. [Em linha]. Recife, 2022. Disponível em: <https://repositorio.ufpe.br/bitstream/123456789/47453/1/TCC%20Vitor%20Heitor%20de%20Paiva.pdf>.

Trabalho de Conclusão de Curso em Ciência da Informação – Universidade Federal de Pernambuco.

SANTAELLA, Lucia; KAUFMAN, Dora

2024 *A Inteligência Artificial Generativa como quarta ferida narcísica do humano. MATRIZES*. [Em linha]. 18:1 (2024) 37-53. Disponível em: <https://revistas.usp.br/matrizes/article/view/210834/204171>.

SHORT, Cole; SHORT, Jeremy Clark

2023 *The Artificially intelligent entrepreneur: ChatGPT, prompt engineering, and entrepreneurial rhetoric creation. Journal of Business Venturing Insights*. [Em linha]. 19 (Jun. 2023) e00388. Disponível em: <https://doi.org/10.1016/j.jbvi.2023.e00388>.

TECHTARGET

2025 *What is a Prompt?: Definition from TechTarget*. [Em linha]. 2025. Disponível em: <https://www.techtarget.com/searchenterpriseai/definition/AI-prompt>.

TRINDADE, Alessandra Stefane Cândido Elias da; OLIVEIRA, Henry Poncio Cruz de

2024 *Inteligência Artificial (IA) generativa e competência em informação: Habilidades informacionais necessárias ao uso de ferramentas de ia generativa em demandas informacionais de natureza acadêmica-científica. Perspectivas em Ciência da Informação*. [Em linha]. 29 (2024) e-47485. Disponível em: <https://www.scielo.br/j/pci/a/GVCW7KbcRjGVhLSrmy3PCng/?format=pdf&lang=pt>.

ZHANG, Yuxin; WANG, Lin

2024 Survey of user interface design and interaction techniques in generative AI applications. [Em linha]. *ArXiv preprint*. (2024). Disponível em: <https://arxiv.org/pdf/2410.22370>.

Ranna Redlley Silva Oliveira | rannaredlley@gmail.com

Universidade Federal de Pernambuco (UFPE), Brasil

Sandra de Albuquerque Siebra | sandra.siebra@ufpe.br

Universidade Federal de Pernambuco (UFPE), Brasil