ARTIGO
DOI: https://doi.org/10.21728/p2p.2025v12n1e-7624
Data de submissão: 27/05/2025 Data de aprovação: 15/07/2025 Data de publicação: 28/07/2025

REVOLUCIONANDO A INCLUSÃO DIGITAL
uma avaliação híbrida de interfaces potencializada pelo GPT-4

Christian Alex de Souza da Silva[1]
Universidade Federal do Amazonas
christian.silva@ufam.edu.br

Tayana Uchôa Conte[2]
Universidade Federal do Amazonas
tayana@icomp.ufam.edu.br

Wilson Silva Prata[3]
Universidade Federal do Amazonas
wilsonprata@gmail.com

Resumo

Este artigo apresenta uma abordagem híbrida para a avaliação de interfaces digitais inclusivas, combinando o método GenderMag ao modelo de linguagem GPT-4. A pesquisa explora técnicas de *Prompt Engineering* para identificar barreiras de inclusão, como rótulos confusos e falta de feedback visual, ao mesmo tempo em que mantém o olhar crítico de um inspetor humano. Aplicado à plataforma Kahoot, tomando a persona "Abby" como referência, o método demonstrou convergência moderada através de um comparativo estatístico (*Kappa* de Cohen) entre a análise automatizada e a tradicional, reforçando o potencial de o GPT-4 contribuir com *insights* úteis e de reduzir significativamente o tempo de avaliação. Ainda assim, limitações como a dificuldade de capturar todos os aspectos de uma navegação direta e o limite de imagens enviadas ao modelo destacam a importância de manter o papel do especialista humano na inspeção.

Palavras-chave: GPT-4; Gendermag; interfaces inclusivas;

REVOLUTIONIZING DIGITAL INCLUSION
a GPT-4 powered hybrid evaluation of interfaces

Abstract

This article presents a hybrid approach to evaluating inclusive digital interfaces by combining the GenderMag method with the GPT-4 language model. The research explores Prompt Engineering techniques to identify inclusion barriers, such as confusing labels and lack of visual feedback, while maintaining the critical perspective of a human inspector. Applied to the Kahoot platform, using the "Abby" persona as a reference, the method demonstrated moderate convergence—measured by Cohen's Kappa—between the automated and traditional analyses, reinforcing GPT-4's potential to offer useful insights and significantly reduce evaluation time. Even so, limitations such as the difficulty of capturing all aspects of direct navigation and the limit on images sent to the model highlight the importance of preserving the human specialist's role in the inspection process.

Keywords: GPT-4; Gendermag; inclusive interfaces;

[1] Especialização em Design de Produtos Digitais - UX/UI pela Universidade Anhanguera - Uniderp (2022) e bacharelado em Design com ênfase em Interfaces Digitais (2018).

[2] Graduação em Ciência da Computação pela Universidade Federal do Pará (1996), mestrado em Ciências da Computação pela Universidade Federal de Minas Gerais (2001) e doutorado em Engenharia de Sistemas e Computação pela Universidade Federal do Rio de Janeiro (2009). Pós-doutorado na University of California - Irvine (UCI) - (2019-2020), como bolsista do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). Professora Associada da Universidade Federal do Amazonas.

[3] Graduação em Desenho Industrial pela UFAM (Universidade Federal do Amazonas, 2003), especialização em Arte Multimídia pela mesma instituição (2003), MBA em Marketing pela FGV/ISAE AM (2011) e Mestrado (2013) e Doutorado (2017) em Design pela PUC-Rio. UX LEAD no Instituto de Pesquisas Eldorado

REVOLUCIONANDO LA INCLUSIÓN DIGITAL

una evaluación híbrida de interfaces potenciada por GPT-4

Resumen

Este artículo presenta un enfoque híbrido para la evaluación de interfaces digitales inclusivas, combinando el método GenderMag con el modelo de lenguaje GPT-4. La investigación explora técnicas de *Prompt Engineering* para identificar barreras de inclusión, como etiquetas confusas y la falta de retroalimentación visual, al tiempo que mantiene la perspectiva crítica de un inspector humano. Aplicado a la plataforma Kahoot, tomando la persona "Abby" como referencia, el método demostró una convergencia moderada a través de un comparativo estadístico (Kappa de Cohen) entre el análisis automatizado y el tradicional, reforzando el potencial de GPT-4 de aportar *insights* útiles y de reducir significativamente el tiempo de evaluación. Aun así, limitaciones como la dificultad de capturar todos los aspectos de una navegación directa y el límite de imágenes enviadas al modelo ponen de manifiesto la importancia de mantener el rol del especialista humano en la inspección.

Palabras clave: GPT-4; GenderMag; interfaces inclusivas;

1 INTRODUÇÃO

As tecnologias digitais não são neutras, pois refletem as ideias e os valores de seus criadores e do contexto social de sua produção. Dessa forma, a inclusão digital é um desafio central em um mundo cada vez mais dependente dessas tecnologias. Apesar dos avanços, estudos mostram que interfaces digitais frequentemente refletem normas sociais e estereótipos de gênero, perpetuando desigualdades estruturais em ambientes tecnológicos (Kovaleva; Happonen; Kindsiko, 2022). Essa exclusão está relacionada à falta de diversidade nas equipes de desenvolvimento, que tendem a projetar produtos baseados nas perspectivas do grupo dominante, negligenciando as necessidades de outros perfis de usuários (Nunes *et al.,* 2023). Nesse contexto, o crescente número de estudos, como o método GenderMag (Burnett *et al.,* 2016), tem se mostrado eficaz na identificação dessas barreiras de gênero presentes no desenvolvimento de softwares.

Nesse contexto, técnicas qualitativas específicas foram desenvolvidas para mitigar essas desigualdades e promover uma agenda mais equitativa, dentre elas o método GenderMag. Baseado nos conceitos do Percurso Cognitivo, o método utiliza cinco facetas cognitivas: motivação, estilos de processamento, autoeficácia, aversão ao risco e tinkering, que são incorporadas em personas para permitir uma análise direcionada e inclusiva (Burnett et al., 2016). Contudo, Burnett *et al.* (2016) ressaltam que a aplicação manual do método exige uma preparação detalhada, o que pode se tornar demorado e trabalhoso. Embora possa ser aplicado individualmente, o GenderMag Kit (GenderMag Manual, 2020) recomenda que uma inspeção de um único cenário dure de dois a três horas, com a participação de até três pessoas. Essa dificuldade na execução é comum em diversas variações do Percurso Cognitivo, que frequentemente requerem a adaptação dos treinamentos e carecem de ferramentas auxiliares para se obter resultados mais assertivos (Mahatody; Sagar, Kolski, 2010). Portanto, identifica-se uma oportunidade para explorar novas abordagens que permitam otimizar o processo de inspeção, reduzindo o tempo investido pela equipe sem comprometer a qualidade da análise.

Dentre as abordagens recentes para redução de esforço e tempo em tarefas relacionadas a linguagens, o uso de Modelos de Linguagem de Grande Escala (LLMs) tem se tornando cada vez mais comum e representa um caminho tecnológico viável para aprimorar o GenderMag. Estudos recentes demonstram que os LLMs são sistemas de inteligência artificial que identificam padrões contextuais e realizam análises com alta eficiência, especialmente em cenários multilinguísticos e multiculturais (Rathje *et al*., 2024), o que justifica a integração desses modelos ao método GenderMag. Neste estudo, foram empregadas técnicas do Prompt

Engineering, baseadas no modelo apresentado por Vogelsang (2024), para configurar o GPT-4 como um inspetor automatizado, estruturado em quatro componentes essenciais: Instrução, Contexto, Dados de Entrada e Indicador de Saída. Este modelo visa replicar a profundidade analítica do método tradicional de maneira mais ágil e com menor subjetividade, otimizando o processo de inspeção. Diante disto, este estudo visa investigar a seguinte questão de pesquisa: O ChatGPT (GPT-4) é capaz de substituir a inspeção humana no método GenderMag e reduzir o tempo necessário para a análise sem comprometer a qualidade dos resultados?

A principal contribuição deste estudo está na apresentação de um caminho inovador para a otimização do método GenderMag. Este estudo ressalta a necessidade de alternativas que mantenham a profundidade analítica do método tradicional, mas que ofereçam maior agilidade e consistência. Ao integrar o GPT-4 por meio de técnicas de Prompt Engineering, foi criado um inspetor automatizado que busca replicar os resultados do método tradicional, otimizando o tempo investido pela equipe. Assim, o estudo contribui para o aprimoramento das práticas de avaliação de inclusão em ambientes digitais, promovendo a evolução dos métodos de inspeção no desenvolvimento de softwares.

2 BACKGROUND E TRABALHOS RELACIONADOS

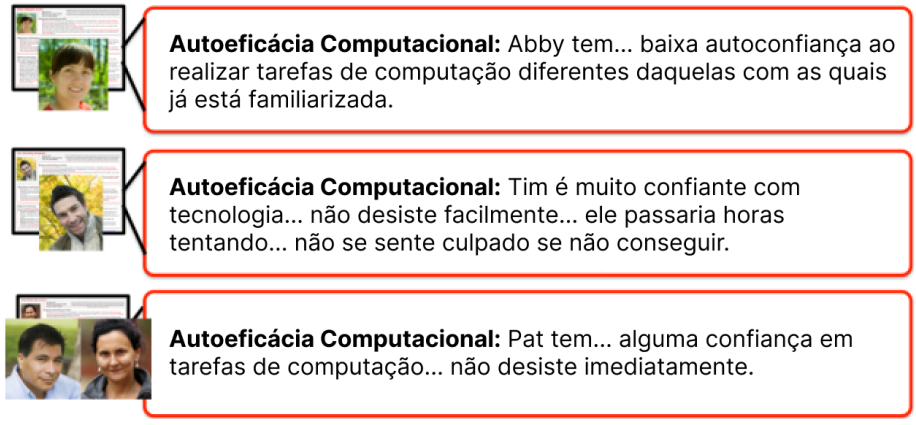
2.1 MÉTODO GENDERMAG

O GenderMag representa uma abordagem eficiente na detecção de vieses de gênero em software. Esse método baseia-se em pesquisas estatísticas que identificaram diferenças significativas nos modos de resolução de problemas entre usuários de diferentes gêneros, e propõe a avaliação da inclusividade de interfaces por meio de um processo sistemático de inspeção (Burnett *et al*., 2016). Estudos iniciais demonstram que, ao aplicar o GenderMag, é possível identificar problemas de inclusividade em até 25% dos cenários inspecionados, evidenciando a relevância do método para promover designs mais equitativos (Burnett *et al.,* 2016). Assim, o GenderMag se apresenta como uma ferramenta fundamental na inspeção de softwares, promovendo uma melhora significativa na experiência do usuário.

Nas inspeções com o GenderMag, o uso de personas é essencial para guiar a aplicação do método. Para validar as diferenças de comportamento observadas na prática, o método define personas como Abby, Tim e Pat(ricia)/Pat(rick), onde cada persona está associada a um conjunto de facetas. De acordo com Burnett *et al.* (2016), a faceta "Motivação" indica que, de forma geral, as mulheres valorizam a tecnologia pelo que ela permite realizar, enquanto os homens tendem a se motivar pela experiência e pelo entretenimento que a interface proporciona.

Em relação ao "Processamento de Informação", pesquisas sugerem que mulheres costumam adotar uma abordagem mais abrangente, buscando coletar detalhes completos antes de tomar decisões, enquanto os homens frequentemente se baseiam em informações mais resumidas para agir rapidamente. A "Autoeficácia Computacional" trata da confiança dos usuários para lidar com desafios tecnológicos, com estudos apontando que as mulheres podem apresentar níveis mais baixos de autoconfiança em contextos digitais. Já a faceta "Aversão ao Risco" evidencia que, estatisticamente, as usuárias são mais cautelosas ao experimentar novas funcionalidades, diferentemente dos homens, que tendem a assumir riscos de maneira mais natural. Por fim, "Tinkering" descreve a propensão à experimentação: embora mulheres sejam geralmente menos inclinadas a testar novas ferramentas, quando o fazem, costumam refletir mais sobre as consequências de suas ações em comparação aos homens. Por exemplo, a persona Abby exige informações detalhadas e feedback abrangente, enquanto Tim prefere interações rápidas e concisas. Se a interface for muito resumida, Abby pode ter dificuldades em compreender as funcionalidades, mas se for excessivamente detalhada, Tim pode se sentir sobrecarregado. Dessa forma, as personas possibilitam que os avaliadores entendam como diferentes perfis podem ser afetados pelos aspectos de design de um software, fundamentando a análise inclusiva.

Figura 1 - Faceta e sua diferença para cada persona



Fonte: Burnett *et al.* (2016).

O processo de inspeção com o GenderMag adapta o método Percurso Cognitivo (PC) para questões de gênero. Segundo Mahatody e Kolsky (2010), um PC viabiliza uma inspeção de usabilidade em interfaces baseando-se em um modelo cognitivo. O GenderMag é realizado a partir do planejamento dos cenários e definição de personas, que percorre um cenário de uso de acordo com um objetivo principal, respondendo a perguntas específicas para cada subobjetivo (passos necessários para alcançar o objetivo principal) e ações de cada fluxo de

interação presente em cada suobjetivo (Burnet *et al.* 2016). Essas perguntas, focadas em verificar se a interface suporta o estilo de interação característico da persona selecionada, permitirão identificar se um determinado elemento é visto como barreiras para determinados grupos. Ao combinar a fundamentação teórica das diferenças de gênero com um procedimento prático e estruturado, o método não só identifica problemas, mas também fornece subsídios para a reformulação do design das interfaces.

2.2 LARGE LANGUAGE MODELS E ENGENHARIA DE PROMPTS

Conforme pode ser concluído a partir da apresentação técnica, o GenderMag trabalha a linguagem para identificar e mitigar os vieses de gênero. Assim, para a otimização da aplicação do modelo, verificamos quais tecnologias possibilitam a redução de tarefas e análise que fazem uso da linguagem como objeto de pesquisa. Assim, no paradigma atual os Large Language Models (LLMs) representam a culminação da evolução histórica dos modelos de linguagem, combinando escalabilidade, diversidade de dados e arquiteturas de transformadores. Segundo Naveed *et al.* (2023), os LLMs são treinados de forma auto-supervisionada com enormes volumes de texto, aprendendo a identificar padrões contextuais complexos e a gerar textos coerentes, o que os torna superiores às abordagens estatísticas tradicionais e permite a realização de atividades que vão desde tradução até raciocínio e planejamento. Wang *et al.* (2023), complementam que a trajetória dos modelos de linguagem evoluiu dos modelos estatísticos (SLMs) para os neurais (NLMs) e, posteriormente, para os modelos pré-treinados em larga escala (PLMs), nos levando até os LLMs modernos, cujo desempenho é amplamente impulsionado pelo aumento do número de parâmetros e pela diversidade dos dados de treinamento. Dessa forma, os LLMs ampliam significativamente as fronteiras da inteligência artificial, estabelecendo uma base robusta para a aplicação de conhecimentos avançados, como a Engenharia de Prompts.

A Engenharia de Prompts tem se consolidado como um elemento essencial para otimizar a interação com LLMs em diversas aplicações. A engenharia de prompts consiste no processo de formular o texto em linguagem natural que será fornecido a um LLM, abrangendo tanto estratégias técnicas que melhoram o desempenho do modelo quanto estudos centrados em como os usuários interagem com esses comandos (Desmond, 2024). De acordo com Vogelsang (2024), a elaboração de prompts eficazes requer a integração de elementos específicos como: instruções, contexto, dados de entrada e indicador de saída, onde combinados, orientam o modelo na geração de respostas mais precisas e alinhadas com as necessidades do usuário. Assim, é possível facilitar as necessidades do usuário para uma linguagem compreensível pelos

LLMs além de servir como um elo vital, entre a complexidade dos modelos e sua aplicação prática com técnicas variadas de acordo com o contexto.

Neste estudo, a aplicação dessas técnicas complementou a estruturação geral do prompt final. A abordagem de *few-shot prompting*, segundo Brown *et al.* (2020), consiste em fornecer ao modelo alguns exemplos ilustrativos do tipo de entrada e saída esperadas, de modo que o LLM possa aprender indutivamente o formato e a lógica desejados. Essa prática é especialmente útil quando se pretende manter um nível de generalização adequado, mas ainda assim fornecer ao modelo informações suficientes para lidar com tarefas mais complexas. Já a técnica de *Chain-of-Thought (CoT) prompting,* possibilita que o modelo apresente etapas intermediárias de raciocínio antes de chegar à resposta definitiva, o que se assemelha à ideia de decompor a análise em passos lógicos (Wei *et al.,* 2023). Embora a estrutura de prompt adotada neste trabalho se beneficie de princípios semelhantes, ela não reproduz exatamente o método de CoT, pois não requer a exibição explícita de cada passo intermediário no processo de raciocínio. Em vez disso, o prompt integra instruções detalhadas, exemplos selecionados (*few-shot*) e ferramentas de análise (como o método GenderMag e o mapeamento de facetas) para guiar a compreensão do modelo. As técnicas de *few-shot* e a inspiração na decomposição lógica de CoT se somam às ferramentas na avaliação de interfaces, resultando em um prompt final que equilibra clareza, detalhamento e flexibilidade na interação com o LLM.

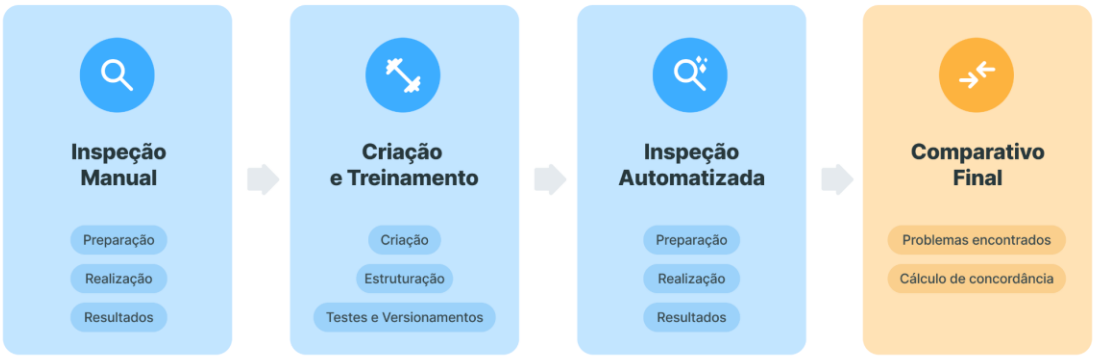
2.3 TRABALHOS RELACIONADOS

Pesquisas recentes demonstram o interesse em automatizar inspeções de interface, tanto por meio de ferramentas específicas para o método GenderMag quanto pela aplicação de LLMs em Percursos Cognitivos. Ferramentas como o AID e o GenderMag Recorder's Assistant foram desenvolvidas para automatizar partes do método GenderMag, reduzindo significativamente o tempo e o esforço das análises manuais e facilitando a identificação de problemas de inclusão em interfaces digitais (Chatterjee *et al*., 2021; Mendez *et al*., 2018). Embora essas ferramentas não sejam baseadas em LLMs, elas evidenciam a necessidade de métodos automatizados para a avaliação inclusiva. Paralelamente, a utilização de LLMs – exemplificada por iniciativas como o CWGPT, que utiliza o GPT-4 – permite simular interações humanas em avaliações do tipo Percurso Cognitivo, oferecendo feedback detalhado e identificando padrões contextuais complexos (Bisante *et al*., 2024).

3 METODOLOGIA

Esta seção descreve o processo metodológico adotado para comparar a aplicação manual do método GenderMag com a versão automatizada via GPT. O processo foi dividido em quatro grandes etapas, conforme ilustra a Figura 2: (1) Inspeção Manual, (2) Criação e Treinamento do Prompt, (3) Inspeção Automatizada e (4) Comparativo Final dos Resultados.

Figura 2 - Processo do estudo.



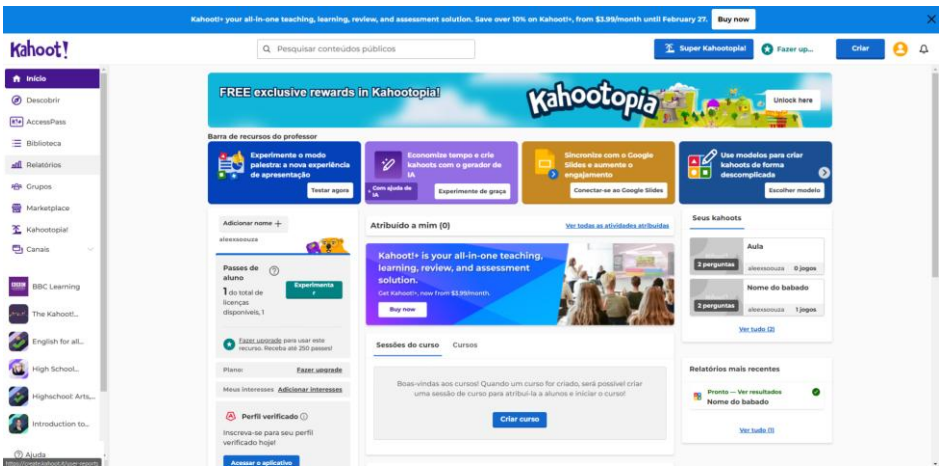
Fonte: Autor (2025).

3.1 INSPEÇÃO MANUAL

A inspeção manual foi dividida em preparação, realização e análise dos resultados. Na preparação, definiu-se o cenário e a persona a serem avaliados, além de registrar as interfaces em uma planilha de controle. O cenário escolhido foi a plataforma educacional *Kahoot* (Kahoot, 2024) (Figura 3), que já havia sido inspecionada por outra autora (Pinheiro, 2021) e apresentava possíveis problemas de gênero na interface. Como se trata de um fluxo complexo, com diversos passos e componentes, acredita-se que esse tipo de cenário seja capaz de evidenciar tanto falhas aparentes quanto problemas mais sutis de usabilidade (Ismagambetov, 2025). Por essa razão, o uso de um cenário extenso aumenta a probabilidade de se identificar barreiras de inclusão, uma vez que a variedade de interações expõe dificuldades que não surgiriam em um fluxo mais simples. A persona adotada foi a Abby, pois, conforme Burnett *et al.* (2016), suas facetas frequentemente ajudam a evidenciar problemas de gênero com maior facilidade em comparação às demais. Também foram utilizados os objetivos gerais, subobjetivos e ações relacionados ao fluxo já definidos por outra Pinheiro (2021), de modo que tanto a inspeção manual quanto a versão automatizada tivessem um direcionamento claro. Por fim, esta inspeção foi realizada somente por um inspetor, o autor principal.

Na realização, seguiu-se o protocolo do GenderMag Kit (2020), que consiste na estruturação do fluxo, objetivo da persona, subobjetivos e ações. Nesse processo, as justificativas das respostas foram embasadas nas facetas da Abby, verificando-se se a persona efetivamente executaria as ações previstas, considerando sua motivação, autoeficácia, aversão ao risco, estilo de processamento e *tinkering*. O objetivo geral estabelecido para a persona foi "criar um questionário em que os alunos concorram entre si ao vivo", refletindo uma tarefa típica de uso em plataformas como Kahoot.

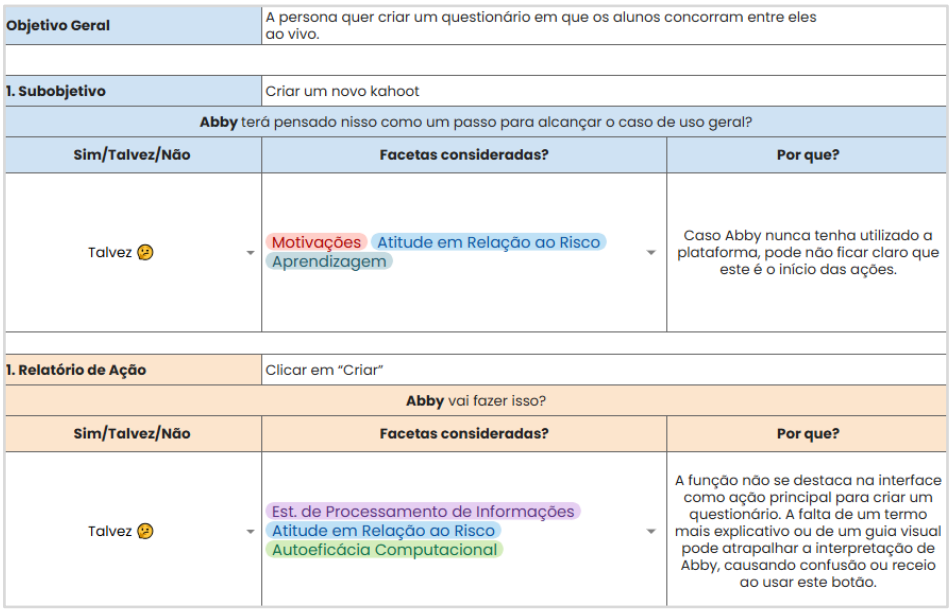
Figura 3 - Plataforma Kahoot



Fonte: Site Kahoot (2025).

Na análise dos resultados, as barreiras encontradas foram registradas em uma planilha (Figura 4), permitindo a visualização rápida das respostas, das facetas consideradas e do motivo das conclusões, organizadas segundo subobjetivos e ações. Esse registro tornou possível prosseguir para as próximas etapas de comparação entre a inspeção manual e a automatizada.

Figura 4 - Planilha de Resultado



Objetivo Geral	A persona quer criar um questionário em que os alunos concorram entre eles ao vivo.	
1. Subobjetivo	Criar um novo kahoot	
Abby terá pensado nisso como um passo para alcançar o caso de uso geral?		
Sim/Talvez/Não	**Facetas consideradas?**	**Por que?**
Talvez 😕	Motivações, Atitude em Relação ao Risco, Aprendizagem	Caso Abby nunca tenha utilizado a plataforma, pode não ficar claro que este é o início das ações.
1. Relatório de Ação	Clicar em "Criar"	
Abby vai fazer isso?		
Sim/Talvez/Não	**Facetas consideradas?**	**Por que?**
Talvez 😕	Est. de Processamento de Informações, Atitude em Relação ao Risco, Autoeficácia Computacional	A função não se destaca na interface como ação principal para criar um questionário. A falta de um termo mais explicativo ou de um guia visual pode atrapalhar a interpretação de Abby, causando confusão ou receio ao usar este botão.

Fonte: Autor (2025).

4.2 CRIAÇÃO E TREINAMENTO DE PROMPT

Este estudo utilizou o ChatGPT-4 (versão paga) como base para implementar um modelo personalizado de análise de interfaces digitais com foco no método GenderMag. A escolha pelo ChatGPT-4 deu-se por diversos benefícios voltado ao seu desempenho, segundo a OpenAI ([20--?]), sua robustez em tarefas de processamento de linguagem natural, flexibilidade para personalização por meio de técnicas de prompt engineering e alta capacidade de adaptação a diferentes contextos de análise. Essa escolha foi essencial para criar um sistema automatizado que identifica barreiras de inclusão de gênero em busca de precisão analítica e escalabilidade.

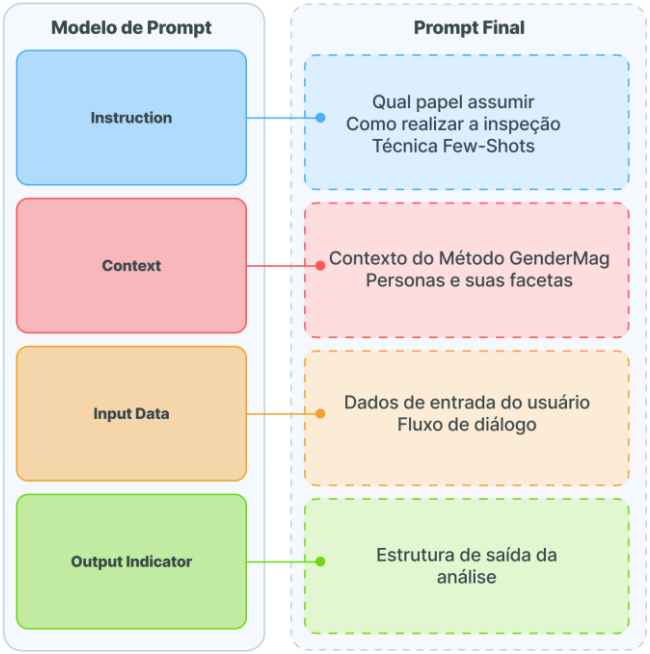
O ChatGPT-4 foi configurado com base em uma estrutura de prompts conforme descrito por Vogelsang (2024), composta por quatro elementos principais:

- **Instrução (Instruction):** Especificação da tarefa a ser executada, como a análise de sub passos e ações em interfaces digitais. Nesta etapa foram definidas instruções diretas sobre qual papel o GPT deve assumir como inspetor e direcionamentos a serem considerados durante sua inspeção.
- **Contexto (Context):** Informações adicionais que orientem o modelo para fornecer respostas mais precisas. Informações gerais sobre as facetas presentes em cada persona, além das delimitações do que deve ser integrado nas respostas finais foram delimitados neste elemento.

- **Dados de Entrada (Input Data):** Questões ou descrições fornecidas pelo usuário, como sub passos de interação ou imagens da interface. Este elemento permaneceu aberto para o usuário final interagir com o GPT, inserindo dados para gerar informações cruciais para a inspeção.
- **Indicador de Saída (Output Indicator):** O formato esperado para a resposta, garantindo estruturação e aplicabilidade nos resultados. Por fim, esta etapa permitiu delimitar o padrão de entrega da análise final, alinhando-se ao resultado esperado pela aplicação do GenderMag.

Essa organização de elementos segue uma estrutura modular (Figura 5), que permite adaptar o modelo e facilita no versionamento, além de direcionar às necessidades do método GenderMag, assegurando que ele responda de forma estruturada e coerente.

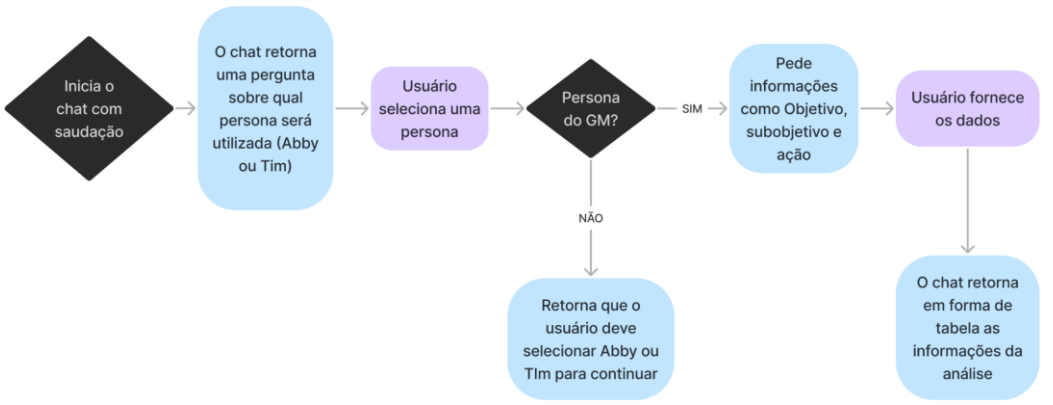
Figura 5 - Estrutura do Prompt



Fonte: Autor (2025).

Para auxiliar o usuário no uso do prompt, desenvolveu-se um fluxo de interação (Figura 6), inserido no módulo *Input Data*, que facilita a comunicação com o sistema, limita as respostas do GPT ao essencial e mantém o usuário focado na realização da inspeção.

Figura 6 - Fluxo de Interação com o usuário.



Fonte: Autor (2025).

Nos testes iniciais, adotou-se um cenário previamente inspecionado por Pinheiro (2021), cujos problemas serviram de base para a validação do prompt. Dessa forma, analisaram-se as respostas do GPT e sua eficácia em identificar as falhas. Ao longo dos testes, o prompt passou por iterações que aprimoraram tanto a estrutura das respostas quanto a capacidade de apontar eventuais problemas, como é evidente no Quadro 1 abaixo.

Quadro 1 - Versionamento da prompt.

Versão	Módulo Alterado	Descrição da alteração
1.1	*Instruction*	Foram adicionadas regras para o uso da interface do usuário como justificativa nas respostas
1.2	*Instruction; Context*	Foram adicionadas regras para consistência nas respostas, além de melhorias nas descrições das facetas
1.3	*Instruction*	Adição da técnica *Few-Shots*

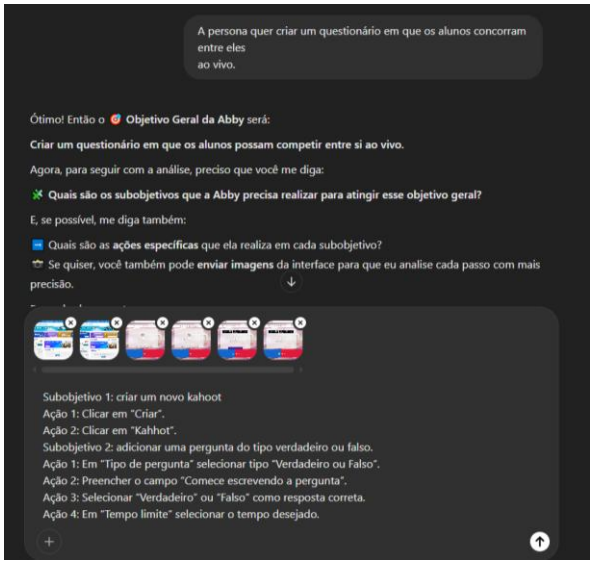
Fonte: Autor (2025).

4.3 INSPEÇÃO AUTOMATIZADA

Na preparação, forneceu-se ao GPT o mesmo cenário complexo utilizado na inspeção manual, a plataforma educacional *Kahoot*, juntamente com a mesma persona (Abby) e seus respectivos objetivos gerais, subobjetivos e ações. O autor interagiu com o modelo via chat, indicando qual persona deveria ser considerada, qual objetivo geral se almejava alcançar e quais

imagens representariam cada subobjetivo e ação da interface. Foram mantidas então, as mesmas condições iniciais da inspeção manual, assegurando equivalência no processo de comparação.

Figura 7 - Exemplo de interação GPT



Fonte: Autor (2025).

Na realização, o GPT foi solicitado a percorrer cada passo definido no fluxo de interação, analisando se a persona efetivamente executaria as ações previstas, com base nas suas características de cada faceta. Por fim, no resultado, assim como na inspeção manual, registraram-se as barreiras encontradas em uma planilha, classificadas de acordo com cada subobjetivo e ação. Em cada registro, figuravam as respostas do GPT, as facetas consideradas e os motivos que levaram à indicação de uma possível barreira. Esse registro estruturado permitiu, ao término do processo, comparar de forma direta os achados das duas modalidades (manual e automatizada), possibilitando verificar a eficácia do GPT na identificação de potenciais problemas de inclusão.

4.4 COMPARATIVO FINAL ENTRE OS RESULTADOS

Para comparar de forma sistemática os achados das inspeções manual e automatizada via GPT, elaborou-se uma planilha (Figura 7) que lista as respostas fornecidas pelo autor no método manual e pelo modelo no método automatizado. Cada linha representa um subobjetivo ou ação analisada, permitindo identificar facilmente onde há concordância ou discrepância.

Figura 7 - Planilha Comparativa.

Nº	Etapa / Pergunta	Autor	GPT
1	Subobjetivo: Criar um novo kahoot – A persona terá pensado nisso?	Talvez	Talvez
2	Relatório de Ação: Clicar em "Criar"	Talvez	Talvez
3	Relatório de Ação, pós-ação: A persona saberá que fez a coisa certa?	Sim	Não
4	Relatório de Ação: Clicar em "Kahoot"	Sim	Talvez
5	Relatório de Ação, pós-ação: A persona saberá que fez a coisa certa?	Sim	Sim
6	Subobjetivo: Adicionar uma pergunta do tipo verdadeiro ou falso – A persona terá pensado nisso?	Talvez	Sim
7	Relatório de Ação: Em "Tipo de questão" selecionar "Verdadeiro ou Falso"	Talvez	Sim
8	Relatório de Ação, pós-ação: A persona saberá que fez a coisa certa?	Sim	Sim
9	Relatório de Ação: Preencher o campo "Comece a digitar a pergunta"	Sim	Sim
10	Relatório de Ação, pós-ação: A persona saberá que fez a coisa certa?	Sim	Sim
11	Relatório de Ação: Selecionar "Verdadeiro" ou "Falso" como resposta correta	Talvez	Sim
12	Relatório de Ação, pós-ação: A persona saberá que fez a coisa certa?	Sim	Sim
13	Relatório de Ação: Em "Tempo limite" selecionar o tempo desejado	Sim	Talvez
14	Relatório de Ação, pós-ação: A persona saberá que fez a coisa certa?	Talvez	Não

Fonte: Autor (2025).

A fim de quantificar essa concordância, recorreu-se ao *Kappa de Cohen* (Cohen, 1960), coeficiente estatístico que indica até que ponto duas avaliações convergem além do que seria esperado por puro acaso. Seu valor varia de -1 a 1:

- Valores próximos de 1 sinalizam forte concordância,
- Valores próximos de 0 sugerem concordância meramente aleatória,
- Valores negativos indicam menos concordância do que o esperado ao acaso.

Assim, calculou-se o Kappa considerando os itens em que o GPT concordou ou não com o autor, permitindo avaliar objetivamente quanto a inspeção automatizada se aproxima do método manual. Esse procedimento assegura uma análise numérica e transparente da eficácia do GPT em reproduzir os resultados do GenderMag tradicional.

5 RESULTADOS

5.1 INSPEÇÃO MANUAL

A inspeção manual contou com o cenário determinado, a partir da persona Abby na plataforma *Kahoot*. Cada ação/subobjetivo foi classificado como "Sim", quando a usuária potencial (Abby) demonstraria segurança em seguir adiante, ou "Talvez", indicando incertezas ou riscos de barreira. No final da inspeção nenhum passo recebeu o rótulo de "Não".

Durante um período de 2 horas, foram analisados 46 subpassos, entre subobjetivos e ações. Desses, 37 receberam classificação "Sim", sugerindo que a persona não encontraria obstáculos relevantes para prosseguir, enquanto 9 foram marcados como "Talvez", indicando pontos de ambiguidade ou risco de abandono da tarefa. Esses 9 casos configuram potenciais problemas de inclusão, relacionados principalmente a rótulos pouco claros em botões ou ícones,

que geram dúvidas sobre o propósito da ação, à falta de feedback explícito após determinadas etapas, o que pode inibir a confiança de uma usuária cautelosa, e ao excesso de etapas de configuração sem auxílio ou instruções diretas, exigindo maior esforço de exploração, algo que Abby tende a evitar.

A maior parte das incertezas (“Talvez”) envolve as facetas Estilo de Processamento de Informações e Atitude em Relação ao Risco, sinalizando que a usuária pode hesitar ao tomar decisões sem garantias visuais e sem instruções claras na tela. Em contrapartida, as ações marcadas como “Sim” mostram que a plataforma oferece sinais suficientes (por exemplo, feedback imediatos ou rótulos autoexplicativos), permitindo que a persona se mantenha engajada.

5.2 INSPEÇÃO AUTOMATIZADA

Para verificar em que medida o GPT-4 reproduz as observações feitas na inspeção manual, foi simulada a mesma análise na plataforma *Kahoot*, orientando o modelo a seguir os mesmos subobjetivos e mesmas ações. A persona adotada continuou sendo a Abby, com foco nas suas facetas. A seguir, um resumo do que o GPT indicou como respostas (Sim/Talvez/Não) para cada subobjetivo/ação.

Por um período de aproximadamente 5 minutos, foram analisados 46 subpassos. Desses, 38 receberam a classificação “Sim”, indicando que Abby tenderia a prosseguir sem grandes problemas; 5 foram avaliados como “Talvez”, sugerindo um risco potencial de barreira; e 3 foram marcados como “Não”, apontando pontos em que a usuária ficaria insegura quanto ao progresso. Assim, 8 subpassos (os 5 “Talvez” e os 3 “Não”) representam possíveis problemas, principalmente relacionados a botões ou opções pouco descritivos, como o botão “Criar”, que não deixa claro se serve para criar um quiz ou outro tipo de conteúdo , à falta de feedback ou recomendações na escolha de configurações como o “Tempo limite”, que gera dúvidas sobre a adequação do valor selecionado, e à ausência de rótulos claros em determinados menus, o que pode fazer com que abby, com sua baixa autoeficácia e alta cautela, hesite antes de avançar.

Assim como na abordagem manual, o modelo identificou ações bem sustentadas por feedback imediato (ícones de confirmação, mudança de layout), que diminuem as incertezas de Abby. No entanto, sempre que a interface se mostrava ambígua ou pouco instrutiva, o GPT classificou as ações como “Talvez” ou “Não”, reforçando a importância de fornecer pistas visuais e textuais mais consistentes.

5.3 COMPARATIVO FINAL E CONCORDÂNCIA

Ao final das inspeções manual e automatizada, criou-se uma planilha comparativa para verificar quantos problemas foram indicados em comum e quantos eram divergentes. Com o intuito de padronizar para a análise estatística, toda ocorrência de “Talvez” foi tratada como “Não” (ou seja, um problema), de modo a produzir uma matriz de concordância binária: Sim (sem problema) ou Não (problema).

- **Manual:** Foram identificados 9 subpassos como potencialmente problemáticos entre 46 subpassos analisados, envolvendo sobretudo as facetas de Autoeficácia (Baixa) e Atitude em Relação ao Risco (Alta).
- **Automatizado (GPT):** Ao aplicar o mesmo critério, 8 subpassos foram classificados como potencialmente problemáticos. As mesmas duas facetas se destacaram na maioria dos casos.

Figura 8 - Comparativo de Problemas.

Fonte: Autor (2025).

Para aferir a concordância entre as classificações (Sim/Não) de ambos os inspetores (manual vs. GPT), foi usado o coeficiente *Kappa de Cohen*. Com base na matriz de confusão derivada dos dados, chegou-se ao valor de 0,35. Conforme escalas usuais de interpretação, esse resultado indica um nível de concordância razoável (“*fair agreement*”), mas não muito alto.

Em síntese, as duas inspeções apontaram problemas semelhantes (sobretudo em passos que exigiam maior segurança ou feedback), mas houve algumas discrepâncias pontuais, refletindo diferenças na forma de julgar um subpasso como “Talvez” ou não. Ainda assim, o Kappa de 0,35 sugere que o método automatizado se aproximou de forma moderada às conclusões do método manual, validando em parte a capacidade do GPT de replicar o raciocínio do avaliador humano.

6 DISCUSSÃO

Os resultados evidenciam que embora o GPT-4 reproduza a maioria das observações realizadas na inspeção manual, há pontos de discordância que merecem atenção. Por exemplo, o modelo indicou que a descrição do menu "Kahoot" pode causar confusão para a persona Abby, enquanto a inspeção manual não apontou problemas neste subpasso. Essa divergência sugere que o GPT capta nuances específicas, como a falta de clareza em rótulos e a ausência de feedback imediato, que ainda fazem sentido do ponto de vista do autor e podem passar despercebidas em uma análise exclusivamente humana. Dessa maneira, a utilização do GPT como ferramenta complementar evidencia potencial para identificar problemas sutis, embora suas interpretações nem sempre estejam facilmente detectáveis pela percepção humana.

Além das diferenças qualitativas na identificação de barreiras, o tempo de execução das análises mostrou-se um fator crucial. Enquanto a inspeção manual, seguindo o protocolo completo do GenderMag, demandou aproximadamente duas horas para avaliar os 46 subpassos, a abordagem automatizada utilizando o GPT-4 foi capaz de realizar a mesma análise em apenas 5 minutos. Esse ganho expressivo de tempo confirma que a automatização pode otimizar o processo de inspeção, entretanto, seu resultado não expressou a qualidade de uma avaliação humana na captação de nuances mais profundas. Assim, o GPT-4 se apresenta como um parceiro eficiente para reduzir o esforço temporal, complementando a análise manual e ampliando a capacidade de detecção de problemas de inclusão.

Diante disso, respondemos a pergunta de pesquisa, pois os achados apontam que o GPT-4 não deve ser considerado um substituto do inspetor humano, mas sim um parceiro auxiliar na avaliação de inclusão. A análise dos resultados, inclusive através do cálculo do *Kappa de Cohen*, demonstrou uma concordância moderada entre os métodos, o que indica que o modelo automatizado contribui com insights que podem complementar a inspeção manual. No entanto, limitações como o número restrito de imagens que podem ser enviadas e a dificuldade de capturar todas as nuances de uma navegação direta restringem sua aplicação. Assim, para

alcançar uma análise robusta, é recomendável que o GPT-4 seja utilizado em conjunto com a avaliação humana, ampliando a perspectiva sobre possíveis problemas de inclusão sem comprometer a profundidade da análise.

7 CONCLUSÃO

Os resultados evidenciam que o GPT-4 pode fortalecer o método GenderMag ao ampliar o escopo de possíveis problemas detectados, mas não substitui o inspetor humano. Apesar de a ferramenta auxiliar na identificação de barreiras de inclusão, há nuances contextuais que exigem o julgamento especializado, principalmente no que se refere à navegação em tempo real e à interpretação de elementos mais subjetivos da interface. Dessa forma, recomenda-se que a avaliação automatizada seja utilizada em conjunto com a análise manual, resultando em um processo mais robusto e abrangente.

Em síntese, a implementação do modelo em um processo híbrido garante agilidade e amplitude sem sacrificar a profundidade da inspeção, impulsionando a construção de interfaces cada vez mais equitativas e acessíveis. Pesquisas futuras poderão explorar a aplicação deste método em diferentes domínios, bem como aperfeiçoar o uso do GPT-4 para tarefas mais complexas, consolidando o potencial de modelos de linguagem no desenvolvimento inclusivo de software.

REFERÊNCIAS

BISANTE, Alba; DATLA, Venkata Srikanth Varma; PANIZZI, Emanuele; TRASCIATTI, Gabriella; ZEPPIERI, Stefano. Enhancing Interface Design with AI: An Exploratory Study on a ChatGPT-4-Based Tool for Cognitive Walkthrough Inspired Evaluations. *In:* International Conference on Advanced Visual Interfaces (AVI' 2024), 2024, Arenzano, Genoa, Itália. **Proceedings [...]** Arenzano, Genoa, Itália, 2024. p. 1-5. DOI: 10.1145/3656650.3656676. Disponível em: https://doi.org/10.1145/3656650.3656676. Acesso em: 02 fev. 2025.

BROWN, Tom B.; MANN, Benjamin; RYDER, Nick; SUBBIAH, Melanie; *et al.* Language Models are Few-Shot Learners. **Portal Cornell University – ArXiv**, 2020. Disponível em: https://arxiv.org/abs/2005.14165. Acesso em: 25 mar. 2025.

BURNETT, Margaret; STUMPF, Simone; MACBETH, Jacob; MAKRI, Stephann; BECKWITH, Laura; KWAN, Irene; PETERS, Anneliese; JERNIGAN, Wendy. GenderMag: A Method for Evaluating Software's Gender Inclusiveness**. Interacting with Computers**, v. 28, n. 6, p. 760-787, 2016. DOI: 10.1093/iwc/iwv046. Disponível em: https://doi.org/10.1093/iwc/iwv046. Acesso em: 01 mar. 2025.

CHATTERJEE, Amreeta; GUIZANI, Mariam; STEVENS, Catherine; EMARD, Jillian; MAY, Mary Evelyn; BURNETT, Margaret; AHMED, Iftekhar; SARMA, Anita. AID: An automated detector for gender-inclusivity bugs in OSS project pages. *In:* International Conference on Software Engineering (ICSE), 43, 2021. **Proceedings [...]** [S.l.]: [s.n.], 2021. p. 1423-1435. Disponível em: https://stairs.ics.uci.edu/papers/2021/AID_An_Automated_Detector_for_Gender-Inclusivity_Bugs_in_OSS_Project_Pages.pdf. Acesso em: 02 fev. 2025.

COHEN, Jacob. A Coefficient of Agreement for Nominal Scales**. Educational and Psychological Measurement**, v. 20, n. 1, 960. Disponível em: https://journals.sagepub.com/doi/10.1177/001316446002000104. Acesso em: 25 mar. 2025.

DESMOND, Michael; BRACHMAN, Michelle. Exploring Prompt Engineering Practices in the Enterprise. **Portal Cornell University – ArXiv**, [2024]. Disponível em: https://arxiv.org/abs/2403.08950. Acesso em: 20 fev. 2025.

ISMAGAMBETOV, Zhenis. User Interface Testing Methods in Complex Software Systems. **International Research Journal of Modernization in Engineering Technology and Science**, v. 07, n. 02, fev. 2025. DOI: https://www.doi.org/10.56726/IRJMETS68058. Disponível em: https://www.doi.org/10.56726/IRJMETS68058. Acesso em: 20 mar. 2025.

KAHOOT. Disponível em: https://kahoot.com/. Acesso em: 20 fev. 2025.

KOVALEVA, Yekaterina; HAPPONEN, Ari; KINDSIKO, Eneli. Designing gender-neutral software engineering program: stereotypes, social pressure, and current attitudes based on recent studies. *In:* International Workshop on Gender Equality, Diversity and Inclusion in Software Engineering (GE@ICSE'22), 3, 2022, Pittsburgh, PA, EUA. **Proceedings [...].** Pittsburgh, PA, EUA: ACM, 2022. p. 43-50. DOI: 10.1145/3524501.3527600. Disponível em: https://doi.org/10.1145/3524501.3527600. Acesso em: 13 fev. 2025.

MAHATODY, Thomas; SAGAR, Mouldi; KOLSKI, Christophe. State of the Art on the Cognitive Walkthrough Method, Its Variants and Evolutions. **International Journal of**

Human-Computer Interaction, v. 26, n. 8, p. 741-785, 2010. DOI: 10.1080/10447311003781409. Disponível em: https://www.researchgate.net/publication/220302514. Acesso em: 13 mar. 2025.

MENDEZ, Christopher; ANDERSON, Andrew; BHUVA, Brijesh; BURNETT, Margaret. The GenderMag Recorder's Assistant**.** *In:* Symposium on Visual Languages and Human-Centric Computing (VL/HCC), 2018, Lisboa, Portugal. **Proceedings […]** Lisboa, Portugal: IEEE, 2018. Disponível em: https://ieeexplore.ieee.org/document/8506505. Acesso em: 20 mar. 2025.

NAVEED, Humza; KHAN, Asad Ullah; QIU, Shi; SAQIB, Muhammad; ANWAR, Saeed; USMAN, Muhammad; AKHTAR, Naveed; BARNES, Nick; MIAN, Ajmal. A Comprehensive Overview of Large Language Models. **Portal Cornell University – ArXiv**, 2023. Disponível em: https://arxiv.org/abs/2307.06435. Acesso em: 21 mar. 2025.

NUNES, Inês; MOREIRA, Ana; ARAUJO, João. GIRE: Gender-Inclusive Requirements Engineering. **Data & Knowledge Engineering**, v. 143, p. 102108, 2023. DOI: 10.1016/j.datak.2022.102108. Disponível em: https://doi.org/10.1016/j.datak.2022.102108. Acesso em: 13 dez. 2024.

OPENAI. Creating a GPT. **Portal OpenAi,** [20--?]. Disponível em: https://help.openai.com/en/articles/8554397-creating-a-gpt. Acesso em: 16 dez. 2024.

PINHEIRO, Rayane Marques. **Usando GenderMag como técnica para avaliação de requisitos de inclusão em ferramentas de ensino online.** 2021. 54 f. Trabalho de Conclusão de Curso (Bacharelado em Engenharia de Software) – Instituto de Computação, Universidade Federal do Amazonas, Manaus, 2021.

RATHJE, Steve; MIREA, Dan-Mircea; SUCHOLUTSKY, Ilia; MARJIEH, Raja; ROBERTSON, Claire E.; VAN BAVEL, Jay J. GPT is an effective tool for multilingual psychological text analysis. **PNAS**, v. 121, n. 34, p.1-11, 2024. DOI: 10.1073/pnas.2308950121. Disponível em: https://doi.org/10.1073/pnas.2308950121. Acesso em: 01 mar. 2025.

VOGELSANG, Andreas. From Specifications to Prompts: On the Future of Generative LLMs in Requirements Engineering. **IEEE Software**, v. 41, n. 5, p. 9-13, Sept.-Oct. 2024. Disponível em: https://ieeexplore.ieee.org/document/10629163. Acesso em: 13 dez. 2024.

WANG, Xinyuan; LI, Chenxi; WANG, Zhen; BAI, Fan; LUO, Haotian; ZHANG, Jiayou; JOJIC, Nebojsa; XING, Eric; HU, Zhiting. PromptAgent: Strategic Planning with Language Models Enables Expert-Level Prompt Optimization. **Portal Cornell University – ArXiv**, 2023. Disponível em: https://github.com/XinyuanWangCS/PromptAgent. Acesso em: 15 mar. 2025.