Uma Taxonomia de Abordagens Baseadas em LLMs para Automação da Inspeção de Usabilidade

Ana Beatriz de Araújo Farias[1], **Emanuel Dantas Filho**[2], **Danyllo Wagner Albuquerque**[3], **Bruno Neiva Moreno**[4]

[1] Mestrado Profissional em Tecnologia da Informação
Instituto Federal de Educação, Ciência e Tecnologia da Paraíba (IFPB)
Paraíba – PB – Brasil

ana.farias.1@academico.ifpb.edu.br, emanuel.filho@jaboatao.ifpe.edu.br

danyllo.albuquerque@ifpb.edu.br, bruno.moreno@ifpb.edu.br

Abstract. *This paper proposes a taxonomy of LLM-based approaches for automating usability inspection. To this end, a systematic mapping study was conducted across established scientific databases, considering publications from 2022 to 2025. The search yielded 401 studies, of which 14 were selected based on exclusion and quality criteria. The taxonomy was derived from a qualitative synthesis guided by the research questions. The results indicate a predominance of automated approaches, a recurrent adoption of GPT-4 family models, and the central role of prompt engineering in operationalizing inspections. Based on these patterns, the taxonomy organizes the literature into five dimensions: LLM usage strategies, level of automation, input artifacts, inspection references, and validation procedures. The analysis also revealed recurring challenges, such as response inconsistency, false positives, difficulties in contextual interpretation, reliance on human validation, and concentration on proprietary models. We conclude that the proposed taxonomy not only structures the state of the art but also highlights research gaps and outlines a research agenda toward more robust, comparable, and reliable approaches.*

Resumo. *Este artigo propõe uma taxonomia de abordagens baseadas em Modelos de Linguagem de Grande Escala (LLMs) para automação da inspeção de usabilidade. Para isso, foi realizado um mapeamento sistemático da literatura em bases de dados científicas consolidadas, considerando publicações entre 2022 e 2025. A busca resultou em 401 estudos, dos quais 14 foram selecionados com base em critérios de exclusão e de qualidade. A taxonomia foi derivada de uma síntese qualitativa orientada pelas questões de pesquisa. Os resultados indicam a predominância de abordagens automatizadas, a adoção recorrente de modelos da família GPT-4 e o papel central do prompt engineering na condução das inspeções. Com base nesses padrões, a taxonomia organiza a literatura em cinco dimensões: estratégias de uso de LLMs, nível de automação, artefatos de entrada, referenciais de inspeção e formas de validação. A análise também evidenciou desafios recorrentes, como inconsistência nas respostas, falsos positivos, dificuldades de interpretação contextual, dependência de validação humana e concentração em modelos proprietários. Conclui-se que a taxonomia proposta não apenas organiza o estado da arte, mas também explicita lacunas e orienta uma agenda de pesquisa voltada ao desenvolvimento de abordagens mais robustas, comparáveis e confiáveis.*

1. Introdução

A usabilidade constitui um atributo central da qualidade de sistemas interativos, estando diretamente relacionada à eficácia, à eficiência e à satisfação dos usuários durante a

interação com interfaces digitais [International Organization for Standardization 2018]. Problemas de usabilidade podem comprometer não apenas a experiência de uso, mas também a execução adequada de tarefas, a adoção de sistemas e o alcance de objetivos organizacionais e de negócio [Norman 2006]. Nesse contexto, a identificação precoce de problemas de interação torna-se um aspecto relevante no processo de desenvolvimento de software.

Entre as estratégias empregadas para esse fim, a inspeção de usabilidade ocupa posição de destaque por permitir a identificação sistemática de problemas em interfaces, com base em princípios, heurísticas e diretrizes previamente estabelecidos [Doğan et al. 2014, Fernandez et al. 2011, Nielsen and Mack 1994]. Em geral, esse processo é conduzido manualmente por especialistas, o que exige tempo, conhecimento do domínio e interpretação criteriosa dos artefatos avaliados, o que limita sua escalabilidade e a frequência de aplicação em cenários reais.

Nos últimos anos, o avanço da Inteligência Artificial, em especial dos Modelos de Linguagem de Grande Escala (LLMs), tem ampliado as possibilidades de apoio e automação de tarefas analíticas tradicionalmente realizadas por humanos [Shin et al. 2025]. No contexto da inspeção de usabilidade, esses modelos vêm sendo explorados em atividades como interpretação de heurísticas, análise de capturas de tela e protótipos, processamento de representações estruturadas de interfaces e geração de relatórios ou recomendações a partir de critérios de avaliação [Guerino et al. 2025, Lubos et al. 2025, Lu et al. 2025]. Paralelamente, a literatura recente já registra iniciativas secundárias voltadas ao uso de Inteligência Artificial e LLMs em atividades relacionadas à usabilidade, acessibilidade e UX, evidenciando o crescimento do interesse por abordagens baseadas em IA para avaliação de interfaces e a diversidade de estratégias, critérios e contextos de aplicação atualmente investigados [Santos et al. 2025, Kuric et al. 2025, Liu 2025].

Nesse cenário, torna-se importante compreender não apenas quais modelos são utilizados, mas como as abordagens se diferenciam quanto aos artefatos de entrada, ao nível de automação, aos referenciais de inspeção e às formas de validação. Organizar esses elementos em uma estrutura analítica comum pode consolidar o estado da arte e favorecer comparações mais sistemáticas entre as propostas existentes.

Diante disso, este artigo tem como objetivo propor uma taxonomia de abordagens baseadas em LLMs para automação da inspeção de usabilidade, construída a partir de um mapeamento sistemático da literatura. Para isso, foi conduzido um estudo em bases científicas consolidadas, abrangendo publicações entre 2022 e 2025, de modo a contemplar tanto trabalhos recentes baseados em LLMs quanto antecedentes próximos relacionados à automação da inspeção de usabilidade apoiada por IA. A busca resultou em 401 estudos inicialmente identificados, dos quais 14 foram selecionados com base em critérios de exclusão e de qualidade. A partir da análise desses estudos, foi definida uma taxonomia que organiza o domínio em cinco dimensões: estratégias de uso de LLMs, nível de automação, artefatos de entrada, referenciais de inspeção e formas de validação.

Ao propor essa taxonomia, o estudo busca oferecer uma estrutura analítica para organizar a literatura recente, apoiar comparações entre abordagens e evidenciar lacunas metodológicas no uso de LLMs na inspeção de usabilidade. O restante do artigo está organizado da seguinte forma: a Seção 2 apresenta o método de pesquisa adotado no mapeamento sistemático; a Seção 3 descreve o processo de análise qualitativa conduzido para a construção da taxonomia proposta; a Seção 4 apresenta e discute os resultados do estudo à luz das questões de pesquisa; a Seção 5 discute as ameaças à validade; e, por fim, a Seção 6 apresenta as conclusões do estudo.

2. Método de Pesquisa

Este estudo adotou o Mapeamento Sistemático da Literatura como estratégia metodológica, por permitir identificar, classificar e sintetizar evidências científicas, bem como mapear tendências e lacunas em um domínio específico [Petersen et al. 2008]. A condução da pesquisa compreendeu as etapas de definição do escopo, planejamento do protocolo, execução das buscas, seleção dos estudos, avaliação da qualidade e extração dos dados, com apoio da ferramenta Parsifal[1].

2.1. Objetivo e Questões de Pesquisa

O objetivo do estudo foi compreender como os LLMs têm sido utilizados na automação da inspeção de usabilidade, considerando estratégias de uso, níveis de automação, artefatos de entrada, referenciais, formas de validação e limitações reportadas. Para isso, foram definidas cinco questões de pesquisa (QPs), conforme a Tabela 1.

Tabela 1. Questões de pesquisa

QP	Questão de Pesquisa
QP1	Quais LLMs e estratégias de uso, incluindo técnicas de *prompting*, têm sido empregados na automação da inspeção de usabilidade?
QP2	Como as abordagens se caracterizam quanto ao nível de automação, aos tipos de interface e aos domínios de aplicação?
QP3	Quais artefatos de entrada e referenciais de inspeção são mobilizados para operacionalizar a análise?
QP4	Como os estudos validam os resultados obtidos e quais métricas são utilizadas?
QP5	Quais desafios e limitações são reportados na literatura?

As QPs definidas orientaram diretamente a elaboração da estratégia de busca, assegurando o alinhamento entre os objetivos do estudo e o processo de identificação dos trabalhos relevantes.

2.2. Estratégia de Busca

A estratégia de busca foi definida com base nos principais *constructos* do estudo: inspeção de usabilidade e uso de LLMs ou de IA generativa. Com base nesses elementos, identificaram-se sinônimos e variações terminológicas para compor as *strings* de busca. O recorte temporal de 2022 a 2025 foi definido para capturar a emergência e a consolidação recentes do uso de LLMs no contexto da inspeção de usabilidade. Esse período contempla tanto os avanços mais recentes impulsionados por LLMs quanto estudos imediatamente anteriores que exploram a automação da inspeção por meio de técnicas de IA, permitindo uma visão contínua da evolução do campo. A Tabela 2 apresenta as expressões utilizadas.

As buscas foram realizadas em bases científicas consolidadas, considerando trabalhos publicados entre 2022 e 2025. As fontes selecionadas foram Scopus, ACM Digital Library, IEEE Digital Library e SBC OpenLib.

2.3. Seleção dos Estudos

A seleção dos estudos seguiu diretrizes para mapeamentos sistemáticos [Petersen et al. 2015]. A busca automatizada resultou em 401 publicações: Scopus

[1] `https://parsif.al/`

Tabela 2. Strings de busca utilizadas no estudo

String 1: (”Usability inspection”OR ”Heuristic evaluation”OR ”Heuristic analysis”OR ”UX inspection”OR ”user experience inspection”) AND (”LLM”OR ”LLMs”OR ”large language model”OR ”Generative AI”)
String 2 – Adaptada para o SBC OpenLib: (”Inspeção de usabilidade”OR ”avaliação heurística”OR ”análise heurística”OR ”inspeção de UX”OR ”inspeção da experiência do usuário”)

(203), ACM Digital Library (188), IEEE Digital Library (7) e SBC OpenLib (3). Após a remoção de duplicados, realizou-se triagem em dois níveis: análise de títulos, palavras-chave e resumos; e leitura do texto completo dos estudos potencialmente relevantes para confirmar sua aderência ao escopo.

Foram incluídos estudos que abordavam explicitamente o uso de LLMs ou de IA generativa na automação da inspeção de usabilidade, descrevendo, avaliando ou aplicando abordagens automatizadas ou semiautomatizadas, com informações suficientes para análise. Foram excluídos estudos fora do escopo, duplicados, fora do recorte temporal, sem acesso ao texto completo ou redigidos em idiomas diferentes do inglês ou português.

A aplicação dos critérios foi conduzida por dois pesquisadores de forma independente, sendo as divergências resolvidas por um terceiro avaliador. Após esse processo, os estudos remanescentes foram submetidos à avaliação de qualidade metodológica, conforme descrito na subseção seguinte. Ao final do processo de seleção, 14 estudos foram incluídos no *corpus* final de análise. A Tabela 3 apresenta a distribuição dos estudos recuperados e incluídos por fonte de busca. A lista completa dos trabalhos selecionados encontra-se disponível em um repositório público: `https://doi.org/10.6084/m9.figshare.31895143`.

Tabela 3. Distribuição dos estudos por fonte de busca

Fonte	Recuperados	Incluídos	Excluídos
ACM Digital Library	188	1	187
IEEE Digital Library	7	0	7
SBC OpenLib	3	0	3
Scopus	203	13	190
Total	401	14	387

2.4. Avaliação da Qualidade

Os estudos selecionados foram submetidos a uma avaliação de qualidade metodológica, com o objetivo de analisar a clareza, o rigor e a validade das evidências apresentadas. Foram considerados cinco critérios: (i) clareza do objetivo ou da questão de pesquisa; (ii) uso explícito de LLMs na automação da inspeção de usabilidade; (iii) detalhamento metodológico suficiente para reprodutibilidade; (iv) validação dos resultados por comparação com especialistas ou outros métodos de referência; e (v) discussão de desafios, limitações ou restrições reportadas pelos estudos. Cada critério recebeu pontuação 1,0 para “sim”, 0,5 para “parcialmente” e 0,0 para “não”, totalizando até 5,0 pontos por estudo. Foi adotado o ponto de corte de 3,0 pontos para inclusão na análise final. As pontuações detalhadas estão disponíveis em um repositório público[2].

2.5. Extração e Síntese dos Dados

Após a seleção final, os estudos incluídos foram lidos integralmente e submetidos à extração sistemática de dados (vide repositório). A extração foi orientada pelas questões

[2]`https://doi.org/10.6084/m9.figshare.31895143`

de pesquisa e contemplou informações sobre os modelos utilizados, as estratégias de uso de LLMs, o nível de automação, os artefatos de entrada, os referenciais de inspeção, as formas de validação e as limitações reportadas.

Nesta etapa, os dados extraídos serviram a dois propósitos complementares: responder às QP do mapeamento e subsidiar a análise qualitativa empregada na construção da taxonomia proposta. O processo de síntese qualitativa que deu origem às dimensões da taxonomia é descrito na Seção 3.

3. Construção da Taxonomia

Esta seção descreve o processo analítico empregado para derivar a taxonomia proposta a partir do *corpus* final. Embora o mapeamento sistemático tenha identificado e estruturado os estudos relevantes, a taxonomia demandou uma síntese qualitativa adicional para transformar os achados em uma estrutura classificatória do estado da arte, conforme comum em mapeamentos de domínios emergentes em Engenharia de Software [Petersen et al. 2008, Petersen et al. 2015].

Para isso, adotou-se uma estratégia de codificação temática inspirada em Braun e Clarke [Braun and Clarke 2006], adequada à identificação de padrões recorrentes em dados qualitativos e compatível com estudos de síntese em Engenharia de Software [Huang et al. 2018]. O processo também foi conduzido em consonância com a noção de desenvolvimento de taxonomias como atividade de classificação sistemática de objetos de um domínio [Nickerson et al. 2013], de modo que a taxonomia resultante não se limita à mera enumeração de achados, mas constitui uma estrutura analítica derivada iterativamente da literatura selecionada.

3.1. Procedimento de Análise Qualitativa

O ponto de partida da análise foi o conjunto final de 14 estudos incluídos no mapeamento sistemático. Inicialmente, os dados extraídos na Seção 2 foram revisados integralmente, com foco nos elementos relacionados às QPs: modelos empregados, estratégias de uso de LLMs, nível de automação, artefatos de entrada, referenciais de inspeção, formas de validação e limitações reportadas. Essa etapa correspondeu à familiarização com o corpus, com o objetivo de identificar regularidades, diferenças e recorrências relevantes para a classificação das abordagens [Braun and Clarke 2006].

Em seguida, foi realizada uma codificação inicial dos estudos, na qual informações extraídas foram convertidas em rótulos descritivos de características observáveis. Por exemplo, estudos que utilizaram capturas de tela foram codificados como *entrada visual*; abordagens que instruíam o modelo a assumir o papel de avaliador foram codificadas como *role prompting*; e estudos que comparavam resultados do LLM com especialistas foram codificados como *validação humana*. A codificação teve caráter predominantemente indutivo, embora orientada pelas questões de pesquisa.

Na sequência, códigos semanticamente próximos foram agrupados em categorias mais amplas. Assim, *zero-shot prompting*, *chain-of-thought*, *role prompting* e *few-shot prompting* compuseram a dimensão *estratégias de uso de LLMs*, enquanto capturas de tela, protótipos, DOM/JSON, transcrições e logs foram consolidados em *artefatos de entrada*. O agrupamento foi iterativo, conduzido por dois pesquisadores, com divergências resolvidas por consenso [Braun and Clarke 2006, Nickerson et al. 2013].

Por fim, as categorias foram revisadas e consolidadas em dimensões taxonômicas de nível superior. O critério adotado para essa consolidação foi sua capacidade de representar aspectos centrais e recorrentes das abordagens identificadas no corpus, mantendo aderência

ao objetivo do estudo. Como resultado, foram definidas cinco dimensões principais: (i) estratégias de uso de LLMs, (ii) nível de automação, (iii) artefatos de entrada, (iv) referenciais de inspeção e (v) formas de validação.

3.2. Dimensões da Taxonomia

A taxonomia organiza as abordagens em cinco dimensões. A primeira, *estratégias de uso de LLMs*, inclui formas de operacionalização como *prompting* direto ou estruturado, *role prompting*, *chain-of-thought*, *few-shot prompting*, agentes e abordagens multimodais. A segunda, *nível de automação*, distingue entre abordagens assistivas, semiautomatizadas e totalmente automatizadas. A terceira, *artefatos de entrada*, contempla capturas de tela, protótipos, DOM/JSON, transcrições, logs e combinações multimodais. A quarta, *referenciais de inspeção*, inclui heurísticas de Nielsen, diretrizes de acessibilidade, princípios de design, *guidelines* genéricas e critérios de domínio. Por fim, *formas de validação* abrangem comparações com especialistas, métodos de referência, métricas quantitativas e estudos empíricos com usuários ou designers.

A Figura 1 sintetiza a estrutura central da taxonomia definida neste estudo, destacando as cinco dimensões utilizadas para organizar o estado da arte sobre o uso de LLMs na automação da inspeção de usabilidade.

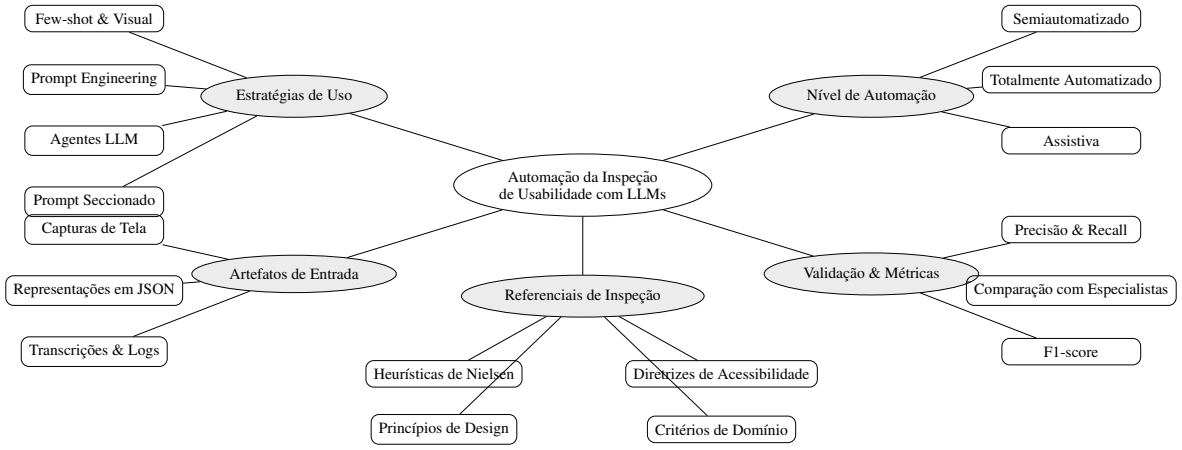

Figura 1. Estrutura da taxonomia proposta

A taxonomia resultante fornece uma estrutura analítica para organizar as abordagens identificadas no mapeamento e orientar a interpretação dos resultados apresentados na seção seguinte. Em vez de tratar os estudos apenas como exemplos isolados, a taxonomia permite compará-los de forma sistemática, destacando padrões recorrentes, diferenças metodológicas e lacunas de pesquisa no campo. Finalmente, a taxonomia atua como estrutura analítica para organizar a interpretação dos resultados e apoiar a resposta sistemática às questões de pesquisa.

A validação da taxonomia foi conduzida analiticamente, por meio da reaplicação das cinco dimensões aos 14 estudos do corpus, verificando se cada estudo poderia ser classificado de forma consistente e se as categorias cobriam as variações observadas. Esse procedimento avaliou a cobertura interna da taxonomia; contudo, ela ainda não foi validada por especialistas independentes nem aplicada a um novo conjunto de estudos, o que constitui uma limitação e uma oportunidade para trabalhos futuros.

4. Resultados e Discussão

Esta seção apresenta os achados do estudo, articulando as QPs às dimensões da taxonomia proposta. A análise aborda os LLMs e estratégias de uso, o nível de automação, os

artefatos de entrada e referenciais de inspeção, as formas de validação e os principais desafios identificados na literatura.

4.1. LLMs e estratégias de uso

No que se refere à QP1, os estudos analisados indicam predominância de modelos da família GPT (OpenAI) na automação da inspeção de usabilidade, especialmente GPT-4, GPT-4o e GPT-4 Turbo with Vision. Também foram identificadas aplicações envolvendo Gemini Pro Vision, Gemini 2.5 Flash, Claude 3 Opus, PaLM 2 e LLaMA, embora com menor frequência. Esse resultado sugere uma concentração em poucos modelos dominantes, especialmente proprietários. Tal "monocultura" tecnológica tem implicações para reprodutibilidade e comparação entre estudos, pois versões, políticas de acesso, parâmetros internos e atualizações dos modelos podem variar ao longo do tempo. A baixa presença de modelos abertos, como LLaMA, também limita auditoria, adaptação local e replicação controlada.

Tabela 4. Frequência de LLMs utilizadas nos estudos

LLM	Quantidade
GPT-4	6
GPT-4o	3
Gemini 2.5 Flash[3]	1
GPT-4 Turbo with Vision	1
Gemini Pro Vision	1
Anthropic Claude-3-Opus	1
Gemini-1.0-Pro	1
LLaMA	1

Ainda em relação à QP1, a análise evidenciou que os estudos diferem não apenas pelos modelos empregados, mas principalmente pelas estratégias de operacionalização adotadas. Foram identificadas abordagens baseadas em *prompting* direto, *role prompting*, *chain-of-thought*, *few-shot prompting*, *visual prompting*, *self-refine*, *least-to-most*, agentes e configurações multimodais. [Shin et al. 2025], por exemplo, comparam diferentes técnicas de *prompt engineering* e apontam melhor desempenho da combinação entre *role prompting* e CoT. [Duan et al. 2024b] exploram representações estruturadas em JSON associadas a *few-shot* e *visual prompting*, enquanto [Lu et al. 2025] utilizam agentes com múltiplos *loops* e fluxo de memória. Esses resultados reforçam que, na literatura atual, a eficácia das abordagens depende não apenas do LLM utilizado, mas da forma como ele é configurado e instruído.

4.2. Nível de automação

No que diz respeito à QP2, observou-se predominância de abordagens totalmente automatizadas, presentes em 13 dos 14 estudos analisados. Apenas um trabalho foi classificado como semiautomatizado, por depender de intervenção humana na coleta ou no fornecimento dos insumos para análise [Bisante et al. 2024]. Esse resultado mostra que o campo tem privilegiado propostas voltadas à automação ampla da inspeção, em detrimento de ferramentas de apoio incremental ao especialista. Quanto aos contextos de aplicação, predominam as interfaces *web* (5 estudos; 35,7%), seguidas por aplicações *mobile* (4; 28,6%), artefatos de design e prototipação (3; 21,4%) e abordagens multiplataforma (2; 14,3%). Não foram identificados estudos com foco exclusivo em aplicações *desktop*, enquanto os domínios mais recorrentes concentram-se em e-commerce, sistemas de recomendação, aplicativos produtivos e aplicações esportivas.

[3]A nomenclatura dos modelos foi revisada conforme reportado nos estudos primários. Quando aplicável, manteve-se o nome informado pelos autores do estudo analisado, evitando inferir versões não explicitadas.

Embora a taxonomia contemple abordagens assistivas, o *corpus* concentrou-se em sistemas semiautomatizados ou totalmente automatizados. Esse achado revela uma tensão: os estudos propõem automação ampla, mas a validação ainda depende do julgamento humano. Assim, a automação total aparece mais como objetivo de projeto do que como evidência de substituição do especialista; no estágio atual, os LLMs parecem mais adequados ao apoio, triagem ou priorização de problemas.

4.3. Artefatos de entrada e referenciais de inspeção

As evidências relacionadas à QP3 mostram que a operacionalização da inspeção automatizada varia significativamente em função dos artefatos de entrada e dos referenciais de inspeção adotados. Entre os artefatos identificados destacam-se capturas de tela, protótipos, representações estruturadas de interface, como DOM e JSON, transcrições de leitores de tela, metadados de hierarquia visual e logs de interação. [Duan et al. 2023], por exemplo, utilizam representações JSON derivadas de mockups do Figma; [Zhong et al. 2025] baseiam-se em transcrições do TalkBack e metadados de visão; e [Hsueh et al. 2024] integram logs de interação capturados via ferramenta Rapi. Em conjunto, esses resultados mostram que os estudos não diferem apenas pelos modelos empregados, mas também pelo tipo de insumo utilizado para viabilizar a inspeção.

Quanto aos referenciais de inspeção, as 10 Heurísticas de Nielsen aparecem como base mais recorrente [Guerino et al. 2025, Duan et al. 2024a, Hsueh et al. 2024], embora também tenham sido identificadas diretrizes de acessibilidade, princípios de design visual, *guidelines* genéricas e estratégias híbridas. [Wu et al. 2024], por exemplo, utilizam os princípios CRAP, enquanto outros estudos combinam mais de um conjunto de critérios para orientar a análise. Essa diversidade indica que a automação da inspeção de usabilidade não depende apenas da capacidade inferencial dos LLMs, mas também do enquadramento conceitual que orienta a avaliação. De forma complementar, alguns estudos também recorrem a bases e conjuntos de dados específicos, como o UICrit Dataset, com 3.059 críticas de design, e o JitterWeb, com 2,3 milhões de exemplos sintéticos, sugerindo diferentes graus de formalização e escalabilidade nas abordagens propostas.

4.4. Formas de validação

No que se refere à QP4, a validação das abordagens é realizada predominantemente por meio de comparação com especialistas humanos ou com métodos de referência. [Guerino et al. 2025] comparam o GPT-4o com três especialistas utilizando testes estatísticos como Qui-quadrado de Pearson e Mann-Whitney. [Zhong et al. 2025] contrastam a cobertura do ScreenAudit (69,2%) com a de ferramentas tradicionais, como o Accessibility Scanner (31,3%). Já [Xiang et al. 2024] conduzem estudo empírico com 48 usuários e 21 designers, reportando similaridade de 80% na cobertura de problemas. Entre as métricas mais frequentes estão precisão, *recall*, *F1-score*, cobertura e porcentagem de concordância; em [Pourasad and Maalej 2025], por exemplo, a precisão variou entre 0,61 e 0,66, enquanto [Bisante et al. 2024] reportam concordância de 93,55% com especialistas humanos.

De modo geral, os resultados mostram que, mesmo em abordagens altamente automatizadas, a validação permanece fortemente ancorada em referência humana. Isso sugere que o campo ainda se encontra em estágio de consolidação metodológica: embora existam evidências promissoras de desempenho, os protocolos de validação são heterogêneos, o que dificulta comparações diretas entre estudos.

4.5. Desafios, limitações e lacunas

Em relação à QP5, os estudos relatam limitações recorrentes que caracterizam o estágio atual da área. Entre os problemas mais frequentes estão alucinações, falsos positivos, di-

ficuldade de interpretação contextual, dependência de *prompt engineering* e limitações na análise de interfaces complexas ou dinâmicas. [Guerino et al. 2025], por exemplo, relatam que 24,3% dos problemas identificados pelo GPT-4o foram classificados como falsos positivos. [Shin et al. 2025] apontam a chamada "distorção de hiper-acurácia", na qual o modelo produz avaliações mecanicamente precisas, mas desconsidera aspectos do julgamento intuitivo humano. [Lu et al. 2025] observam que dados artificiais não reproduzem integralmente o comportamento humano e que o processamento pode ficar restrito a informações textuais, sem capturar adequadamente aspectos do layout visual. [Duan et al. 2023] mencionam a aplicação literal de diretrizes sem consideração suficiente do contexto, além de limitações impostas pela janela de contexto em interfaces mais extensas. Por fim, [Zhong et al. 2025] destacam a dificuldade de interpretar diretrizes abstratas, como as WCAG.

De forma geral, o estado da arte sugere que abordagens baseadas em LLMs apresentam melhor desempenho quando combinam instruções explícitas de inspeção, referenciais claros, como as heurísticas de Nielsen ou as diretrizes de acessibilidade, e artefatos de entrada suficientemente estruturados. Estratégias como *role prompting*, *chain-of-thought*, *few-shot prompting* e uso de representações estruturadas parecem favorecer análises mais direcionadas e comparáveis. Por outro lado, abordagens baseadas apenas em prompts genéricos ou em entradas visuais pouco contextualizadas tendem a ser mais suscetíveis a falsos positivos, a interpretações superficiais e à aplicação literal de diretrizes. Assim, a literatura aponta potencial relevante para apoio à inspeção, mas ainda não demonstra robustez suficiente para substituir integralmente avaliadores humanos em contextos reais.

Em conjunto, esses achados indicam que a literatura já apresenta avanços consistentes na automação da inspeção de usabilidade com LLMs, mas ainda depende fortemente de escolhas metodológicas específicas e de validação humana para garantir confiabilidade. A partir da taxonomia proposta, observa-se que as principais diferenças entre as abordagens não se limitam ao modelo de linguagem adotado, mas também envolvem a combinação entre estratégia de uso, nível de automação, artefatos de entrada, referencial de inspeção e forma de validação. Essa heterogeneidade, embora revele riqueza de possibilidades, também evidencia a ausência de padrões consolidados para comparação e replicação, o que configura uma agenda importante para pesquisas futuras.

5. Ameaças à Validade

Como em todo estudo secundário, os resultados deste trabalho estão sujeitos a limitações metodológicas que devem ser consideradas na interpretação dos achados. Uma primeira ameaça diz respeito à *validade de busca*, uma vez que, embora tenham sido utilizadas bases científicas consolidadas e *strings* elaboradas a partir dos principais *constructos* do estudo, é possível que trabalhos relevantes não tenham sido recuperados devido à diversidade terminológica na área. Esse risco é particularmente relevante em um domínio emergente, no qual diferentes estudos podem utilizar denominações distintas para se referir ao uso de LLMs, de IA generativa ou de automação da inspeção de usabilidade. Uma segunda ameaça refere-se à *validade de seleção*. Apesar da adoção de critérios explícitos de exclusão e de avaliação por múltiplos pesquisadores, o processo de triagem envolve julgamento interpretativo, especialmente na leitura de títulos, resumos e textos completos. Para mitigar esse risco, a inclusão dos estudos foi conduzida com base na concordância entre dois pesquisadores, com as divergências resolvidas por um terceiro avaliador.

Também há ameaças relacionadas à *validade de extração e de classificação*. A extração de dados e a construção da taxonomia dependeram da interpretação das informações relatadas nos estudos primários. Como diferentes artigos apresentam níveis distintos de detalhamento metodológico, algumas classificações exigiram padronização terminológica e agrupamento

analítico. Para reduzir esse risco, a taxonomia foi construída por meio de um processo iterativo de síntese qualitativa, orientado pelas questões de pesquisa e apoiado na codificação temática. Por fim, há ameaças à *validade externa* dos resultados. O corpus final é composto por 14 estudos, o que limita generalizações amplas sobre o campo. Além disso, por se tratar de um domínio em rápida evolução, novas abordagens, modelos e protocolos de avaliação podem surgir em curto prazo, alterando o panorama aqui apresentado. Assim, os resultados deste artigo devem ser interpretados como um retrato estruturado do estado da arte no recorte analisado, e não como uma visão definitiva ou exaustiva do domínio.

6. Conclusão

Este artigo teve como objetivo propor uma taxonomia de abordagens baseadas em LLMs para automação da inspeção de usabilidade, a partir de um mapeamento sistemático da literatura. Para isso, foram analisados 14 estudos selecionados com base em critérios de exclusão e qualidade, abrangendo diferentes contextos e estratégias de aplicação. A partir dos dados extraídos, foi conduzida uma síntese qualitativa orientada pelas questões de pesquisa, permitindo a derivação de uma estrutura classificatória organizada em cinco dimensões centrais: estratégias de uso de LLMs, nível de automação, artefatos de entrada, referenciais de inspeção e formas de validação.

A análise dos estudos evidenciou padrões relevantes em relação às QPs. Observou-se predominância de abordagens automatizadas, com forte concentração no uso de modelos da família GPT-4 e na adoção recorrente de estratégias de *prompt engineering* como principal mecanismo de operacionalização. Além disso, identificou-se diversidade significativa quanto aos artefatos de entrada, incluindo representações estruturadas, capturas de tela e dados de interação, bem como variedade nos referenciais de inspeção adotados. No que se refere à validação, os estudos apresentam abordagens heterogêneas, frequentemente baseadas na comparação com especialistas humanos e no uso de métricas como precisão, *recall* e *F1-score*.

Apesar desses avanços, os resultados também evidenciaram limitações. Entre os principais desafios, destacam-se a ocorrência de falsos positivos, dificuldades na interpretação do contexto, sensibilidade às variações de *prompting* e dependência de validação humana para assegurar a confiabilidade dos resultados. Tais limitações indicam que, embora promissoras, as abordagens baseadas em LLMs ainda carecem de maior maturidade metodológica, especialmente quanto à robustez das análises e à comparabilidade entre estudos.

Nesse contexto, a taxonomia proposta apresenta implicações relevantes. Do ponto de vista teórico, contribui para a organização e sistematização de um domínio ainda em consolidação, oferecendo uma lente analítica para interpretar e comparar abordagens de forma estruturada. Do ponto de vista prático, fornece um referencial que pode apoiar pesquisadores e profissionais na concepção, seleção e avaliação de soluções, evidenciando como diferentes decisões de projeto impactam os resultados obtidos.

Por fim, como direções para trabalhos futuros, destacam-se: (i) desenvolver protocolos de avaliação padronizados, com conjuntos de interfaces, heurísticas e métricas compartilhadas; (ii) comparar sistematicamente técnicas de *prompt engineering*, como *role prompting*, *few-shot*, *chain-of-thought*, *self-refine* e agentes; (iii) avaliar abordagens híbridas, nas quais LLMs atuem como apoio à triagem, priorização ou explicação de problemas, mantendo especialistas no ciclo de decisão; (iv) investigar modelos abertos e proprietários sob os mesmos protocolos, ampliando a reprodutibilidade; e (v) explorar estratégias mais robustas para lidar com contexto, multimodalidade, interfaces dinâmicas e julgamento subjetivo. Adicionalmente, a rápida evolução dos modelos de linguagem sugere a necessidade de revisões contínuas da taxonomia.

Referências

Bisante, A. et al. (2024). Enhancing interface design with ai: An exploratory study on a chatgpt-4-based tool for cognitive walkthrough inspired evaluations. In *Proceedings of the 2024 International Conference on Advanced Visual Interfaces*, AVI '24, New York, NY, USA. Association for Computing Machinery.

Braun, V. and Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101.

Doğan, S., Betin-Can, A., and Garousi, V. (2014). Web application testing: A systematic literature review. *Journal of Systems and Software*, 91:174–201.

Duan, P., Cheng, C.-Y., Li, G., Hartmann, B., and Li, Y. (2024a). Uicrit: Enhancing automated design evaluation with a ui critique dataset. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, UIST '24, New York, NY, USA. Association for Computing Machinery.

Duan, P., Warner, J., and Hartmann, B. (2023). Towards generating ui design feedback with llms. In *Adjunct Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23 Adjunct, New York, NY, USA. Association for Computing Machinery.

Duan, P., Warner, J., Li, Y., and Hartmann, B. (2024b). Generating automatic feedback on ui mockups with large language models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Fernandez, A., Insfran, E., and Abrahão, S. (2011). Usability evaluation methods for the web: A systematic mapping study. *Information and Software Technology*, 53(8):789–817.

Guerino, G., Rodrigues, L., Capeleti, B., Mello, R. F., Freire, A., and Zaina, L. (2025). Can gpt-4o evaluate usability like human experts? a comparative study on issue identification in heuristic evaluation. Accepted at INTERACT 2025.

Hsueh, N.-L., Lin, H.-J., and Lai, L.-C. (2024). Applying large language model to user experience testing. *Electronics*, 13(23).

Huang, X., Zhang, H., and Ali Babar, M. (2018). Synthesizing qualitative research in software engineering: A critical review. In *Proceedings of the 40th International Conference on Software Engineering: New Ideas and Emerging Results*, pages 41–44.

International Organization for Standardization (2018). Iso 9241-11:2018 — ergonomics of human-system interaction – part 11: Usability: Definitions and concepts. Disponível em: `https://www.iso.org/standard/63500.html`. Acesso em: 10 maio 2025.

Kuric, E., Demcak, P., Krajcovic, M., and Lang, J. (2025). Systematic literature review of automation and artificial intelligence in usability issue detection.

Liu, J. (2025). Ai-powered automated and remote ux evaluation methods: A systematic literature review. In *Proceedings of the 43rd ACM International Conference on Design of Communication*, SIGDOC '25, page 10–16, New York, NY, USA. Association for Computing Machinery.

Lu, Y., Yao, B., Gu, H., Huang, J., Wang, J., Li, Y., Gesi, J., He, Q., Li, T. J.-J., and Wang, D. (2025). Uxagent: An llm agent-based usability testing framework for web design. Also available as arXiv:2502.12561 and listed by Amazon Science as CHI 2025.

Lubos, S., Felfernig, A., Garber, D., Le, V. M., and Tran, T. N. T. (2025). Towards llm-based usability analysis for recommender user interfaces. *CEUR Workshop Proceedings*,

4027. Presented at the 12th Joint Workshop on Interfaces and Human Decision Making for Recommender Systems (IntRS 2025).

Nickerson, R. C., Varshney, U., and Muntermann, J. (2013). A method for taxonomy development and its application in information systems. *European Journal of Information Systems*, 22(3):336–359.

Nielsen, J. and Mack, R. L. (1994). *Usability Inspection Methods*. John Wiley & Sons, New York, NY, USA.

Norman, D. A. (2006). *O design do dia-a-dia*. Rocco, Rio de Janeiro.

Petersen, K., Feldt, R., Mujtaba, S., and Mattsson, M. (2008). Systematic mapping studies in software engineering. In *Proceedings of the 12th International Conference on Evaluation and Assessment in Software Engineering (EASE)*, pages 68–77.

Petersen, K., Vakkalanka, S., and Kuzniarz, L. (2015). Guidelines for conducting systematic mapping studies in software engineering: An update. *Information and Software Technology*, 64:1–18.

Pourasad, A. E. and Maalej, W. (2025). Does genai make usability testing obsolete?

Santos, S., Alves, R., Oliveira, F., and Araujo, F. (2025). Investigating llm-based tools to support usability, accessibility, user experience in hci activities: A systematic literature mapping. In *Proceedings of the 31st Brazilian Symposium on Multimedia and the Web*, pages 631–645, Porto Alegre, RS, Brasil. SBC.

Shin, S., Oh, J., and Lee, S. (2025). Can llms see what i see? a study on five prompt engineering techniques for evaluating ux on a shopping site. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '25, New York, NY, USA. Association for Computing Machinery.

Wu, J., Peng, Y.-H., Li, A., Swearngin, A., Bigham, J. P., and Nichols, J. (2024). Uiclip: A data-driven model for assessing user interface design.

Xiang, W., Zhu, H., Lou, S., Chen, X., Pan, Z., Jin, Y., Chen, S., and Sun, L. (2024). Simuser: Generating usability feedback by simulating various users interacting with mobile applications. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Zhong, M., Chen, R., Chen, X., Fogarty, J., and Wobbrock, J. O. (2025). Screenaudit: Detecting screen reader accessibility errors in mobile apps using large language models. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. Association for Computing Machinery.