Weynner Kenneth Bezerra Santos, Felipe de Castro Souza, Bárbara Coutinho da Silva, Paula Peres Gimenez, Ana Karoline Valadão Mendes, Giovanna de Souza Leal Paixão, Walter Franklin Marques Correia, Alexandre Gomes de Siqueira*

Autodesign? Avaliações de Modelo de Linguagem de Grande Escala para Assistentes de UX: Revisão e Diretrizes

Resumo Modelos de linguagem de grande escala (LLMs) estão sendo rapidamente integrados aos fluxos de trabalho de Design Systems, pesquisa em UX e UX Writing, mas ainda faltam métodos sistemáticos para avaliar sua segurança e utilidade. Com base em uma revisão de escopo, este artigo mapeia os principais desafios de avaliação e propõe três frameworks específicos para essas atividades. Cada framework relaciona tarefas de design a critérios mensuráveis de sucesso (por exemplo, ≤3% de alucinações e latência <2s) e sugere conjuntos de dados para testes reprodutíveis. Uma demonstração utilizando GPT-4o e Gemini 2.5 ilustra como essas métricas permitem identificar os pontos fortes e as limitações dos modelos. As diretrizes propostas oferecem às equipes de design um roteiro prático para incorporar testes contínuos baseados em evidências, promovendo maior rigor metodológico e responsabilização no uso de IA como apoio à prática de design.

Palavras Chave LLM, Avaliação, Design, Pesquisa em ux, Escrita em ux

* **Weynner Kenneth Bezerra Santos** PhD candidate in FiGitAL Design (UFPE, 2028); MBA in Project Management (USP, 2022); Master's in Digital Artifact Design (UFPE, 2019); and a Bachelor's degree in Design (UFPE, 2016). Works as Senior Research Analyst at PicPay's Design Center of Excellence, helping the company to apply AI for user research and other design contexts, leading the ResearchOps squad.
wey.uxr@gmail.com
ORCID 0000-0001-7420-8718

Felipe de Castro Souza Bachelor's degree in Business Administration from the Federal University of Uberlândia (2020). In his recent role at PicPay, he led the initiative to develop and adopt the company's new Design System, achieving a cost reduction of almost 30% in mobile development. In addition, he implemented AI assistants to scale design processes in the company, with monthly use by more than 100 employees in the area.

DESIGN, ARTE E TECNOLOGIA

Autodesign? Large Language Model Evaluations for UX Assistants: Review and Guidelines

Abstract *Large language models (LLMs) are being rapidly integrated into design workflows for Design Systems, UX research, and UX writing, but systematic methods to evaluate their safety and utility are lacking. Based on a scoping review, this paper maps key evaluation challenges and proposes three frameworks specific to these tasks. Each framework connects design activities to measurable success criteria (e.g., ≤3% hallucination, <2s latency) and suggests datasets for reproducible testing. A demonstration with GPT-4o and Gemini 2.5 illustrates how the metrics identify the models' strengths and weaknesses. The guidelines offer design teams a practical roadmap to incorporate continuous, evidence-based testing, promoting greater rigor and accountability in AI-assisted design practice.*

Keywords *LLM, Evaluation, Design, Ux research, Ux writing*

¿Autodiseño? Evaluaciones de modelos de lenguaje de gran escala para asistentes de UX: revisión y directrices

Resumen *Los modelos de lenguaje de gran escala (LLMs) se están integrando rápidamente en los flujos de trabajo de Design Systems, investigación en UX y UX Writing; sin embargo, aún faltan métodos sistemáticos para evaluar su seguridad y utilidad. A partir de una revisión de alcance, este artículo identifica los principales desafíos de evaluación y propone tres frameworks específicos para estas actividades. Cada framework relaciona tareas de diseño con criterios de éxito medibles (por ejemplo, ≤ 3 % de alucinaciones y latencia < 2 s) y sugiere conjuntos de datos para realizar pruebas reproducibles. Una demostración con GPT-4o y Gemini 2.5 muestra cómo estas métricas permiten identificar las fortalezas y limitaciones de los modelos. Las directrices propuestas ofrecen a los equipos de diseño una hoja de ruta práctica para incorporar pruebas continuas basadas en evidencia, promoviendo un mayor rigor metodológico y una mayor responsabilidad en el uso de la IA como apoyo a la práctica del diseño.*

Palabras clave *LLM, evaluación, diseño, investigación en UX, redacción UX.*

felipecastrosousa@gmail.com
ORCID 0009-0009-1095-6093

Bárbara Coutinho da Silva Bachelor's degree in Social Communication with a major in Journalism from UFES (2018). She previously worked as a Content Designer at PicPay's Design Center of Excellence, focusing on creating the Content System and implementing an AI assistant to optimize processes, scale content creation, and ensure the best user experience.
barbara_coutinho@hotmail.com
ORCID 0009-0007-6441-0905

Paula Peres Gimenez MBA in Strategic Marketing Management (2015); and a Bachelor's degree in Advertising (2014). She currently works as a Content Designer at PicPay's Design Center of Excellence, where she is involved in creating the company's new Content System, Voice Tone manual, and Writing Style Guide. More recently, she has also been participating in the initiative to apply AI assistants to content creation.
paulaperesgimenez@gmail.com
ORCID 0009-0003-7099-3319

Ana Karoline Valadão Mendes Technologist's Degree in Social Communication (2021) and a member of Mensa Brasil. Digital Product Designer at PicPay, responsible for operations involving tools, people, and processes. Played a key role in enabling the adoption of artificial intelligence solutions to increase operational efficiency and eliminate silos within the company.
valadaokaroline@gmail.com
ORCID 0009-0008-5869-9420

Giovanna de Souza Leal Paixão Bachelor's degree in Social Commu-

nication with a major in Radio, TV and Internet (2023). Product Manager at PicPay, as part of the AI Products team, which is responsible for expanding and democratizing the use of Generative AI (GenAI) within the company. Her work is dedicated to fostering innovation, optimizing processes, and making GenAI solutions accessible to different teams and audiences.
giopxo@gmail.com
ORCID 0009-0007-1952-6204

Walter Franklin Marques Correia Post-Doctorate in the Department of Digital Media Concepts, Academy for AI, Games Media at BUas (Breda University of Applied Sciences, 2024/25); PhD in Production Engineering (UFPE with a Sandwich at the Technical University of Lisbon, 2007); Master in Production Engineering from UFPE (2002); Specialist (Lato Sensu) in Ergonomics from UFPE (2001); and Bachelor in Industrial Design from UFPE (1999).
walter.franklin@ufpe.br
ORCID 0000-0002-6491-9783

Alexandre Gomes de Siqueira Assistant Professor in the Department of Computer & Information Science & Engineering, Post-Doctorate - Virtual Environments Research Lab (University of Florida, 2021), Ph.D. in Human-centered Computing from Clemson University (2019), MBA from Fundação Getulio Vargas - FGV (2011), Bachelor's degree in Computer Science from the University of São Paulo - USP (2005).
agomesdesiqueira@ufl.edu
ORCID 0000-0002-9213-9602

Introduction

Design's consolidation within the Product-Design-Technology triad (CRUTH, 2022) and process automation have intensified the pursuit of efficiency, perhaps leading to its sixth paradigm: Autodesign (adapted from NIJS, 2019), grounded in Artificial Intelligence and deep simulations. Consolidated in Design's third paradigm, user research was the last practice to achieve significant industrial adoption. Subsequent paradigms like Participatory Design (SCHULER & NAMIOKA, 1993) and Emergent Design (NIJS, 2019) saw limited application, but Autodesign may represent the next global shift.

Unlike Metadesign (WOOD, 2020), which focuses on self-reflection, Autodesign is anchored in Large Language Models (LLMs) that execute creative tasks. However, testing and monitoring are often neglected (GOTHELF & SEIDEN, 2022), reinforcing a culture of speed that suppresses essential phases even in frameworks like Design Thinking (BROWN, 2020). This addresses a core issue, as designers often fail to measure their impact, leaving the return on investment (ROI) in Design absent despite its relevance (MORAN, 2020). Despite unresolved ethical challenges in AI (COSKUN, 2023), Autodesign enables the precise and systematic measurement of AI tools.

Applying these systems to design teams requires expertise in Product Design, UI, UX Writing, and User Research, alongside technical skills in prompt engineering, SQL, and Python. This article addresses how to continuously evaluate LLMs through structured systems, filling a critical gap in AI and Conversational Design. The lack of evaluation systems implemented from inception is a primary reason LLM-based products often fail (HUSAIN, 2024).

Teams often prioritize model tuning over continuous testing, a practice that hinders sustainable improvements (RIBEIRO et al., 2020; NG, 2025). The "Lucy" case study (HUSAIN, 2024) illustrates this: after initial gains from prompt tuning, the assistant stagnated. The turning point was implementing systematic, multi-level tests (unit, human, A/B). As Husain (2024) notes, a well-planned evaluation architecture eliminates friction and sustains virtuous cycles of improvement.

Well-structured evaluation systems are crucial for driving model fine-tuning and debugging, where data quality is central. Logs and automated criteria facilitate failure identification. The recommended practice is a testing progression, from unit tests to A/B tests. Metrics must be context-specific, avoiding generic benchmarks, and updated as the model evolves (ZHANG et al., 2023). LLMs can be used to generate test cases and evaluate model outputs.

This approach mirrors engineering principles like test-driven development and is endorsed by the HELM methodology (LIANG et al., 2022). However, its implementation requires expertise to define valid criteria. It also demands attention to the risk of over-optimizing for flawed or static metrics, a critical concern highlighted in recent studies (LIANG et al., 2022; ZHANG et al., 2023; HUSAIN, 2024).

Methodology

This article develops a quality evaluation system for LLM-powered assistants in two phases. The first phase reviews current practices for the continuous evaluation of generative AI. Building on this, the second phase proposes specific testing and remediation guides for three core Design areas: Product (Design System), UX Research, and UX Writing. This paper aims to answer the following research questions, formulated according to FINER criteria and evaluated on a one-to-five scale (HULLEY et al., 2014):

Table 1 Research questions
Fonte The authors., 2025

Class	Questions
Primary	What are the primary challenges in the creation, application, and improvement of evaluation systems for large language models (LLMs)?
Secondary	What are the specific challenges of applying these systems to the evaluation of Design and UX assistants?
	Which existing proposals for evaluation systems are most suitable for application in assistants for Product Design, UX Research, and UX Writing?
	Can the application of evaluation frameworks ('evals') to Design and UX assistants improve the work of designers during the testing and monitoring stages of conversational products?

All questions were approved with high scores, following a logical progression from a broad problem (Main Question) to a specific domain (Q1), practical solutions (Q2), and a discussion of their impact (Q3). For a critical analysis of the questions qualification, a GPT-4.1 model (OPENAI, 2025) was utilized, previously contextualized with the FINER method and trained to adopt a critical stance. While not a full systematic review, PRISMA method (TRICCO et al., 2018) was employed to ensure a minimum standard scope for the study to address source traceability, reproducibility, and replicability leading to a review protocol to guide the literature selection process.

Table 2 Literature review protocol
Fonte Adapted from TRICCO, 2018

Item	Description
Search period	January 2019 — June 2025.
Databases	Google Scholar, ACM Digital Library, arXvi.org, Nielsen Norman Group (nn/g) archives, and pre-print repositories.
Tools	Two pre-trained AI assistants (GPT-4.1 and Gemini 2.5 Pro) in PRISMA method, which were prompted with to suggest references based on pre-injected criteria.
Initial results	A total of 62 documents were initially retrieved: — 48 academic sources (journals, conferences, arXiv); — 14 non-academic sources (expert blogs, NNG articles, product case studies).

Item	Description
Deduplication	— 5 duplicate entries.
Inclusion criteria	— Open access or publicly available full text; — Direct relevance to LLMs, UX assistants, design tools, or AI evaluation; — Published between 2014 and 2025; — At least 50% of the final sources had to be peer-reviewed or academic; — Inclusion of high-authority blogs (e.g., NNG, authors from Google, OpenAI, or Microsoft Research); — Texts cited in Husain (2024) were used as seeds for snowballing
Exclusion criteria	— Paid articles; — Purely theoretical or speculative discussions not grounded in LLM applications; — Articles without clear evaluation focus; — Unverified or anonymous blog posts.
Screening	— Title and abstract screening excluded 19 documents (e.g., duplicates, unrelated to LLMs or UX); — Full-text review excluded another 18 articles (e.g., conceptual only, paywalled, insufficient relevance); — Final selection consisted of 26 documents, but only 20 were used directly in the article's evaluation synthesis, due to some overlapping and repetitive content. — 10 complementary papers were introduced in the final version of the paper, despite not entering in the Literature Review dedicated topic.
Final body	20 total sources formed the review base: — (55%) 11 academic publications (journals/conferences/arXiv); — (45%) 9 professional and industry sources (e.g., NNg, AI blogs).
Language	Articles in all languages were considered. However, all included texts were ultimately in English, as other-language texts failed relevance or accessibility filters. No exclusion was made based on language a priori, but none were retained after screening.
Citation impact	— 5 highly cited sources (>100 citations, e.g., Bender et al., 2021; Ribeiro et al., 2020); — 6 preprints with emerging relevance (arXiv 2022–2025); — 5 expert-authored blogs or industry reports; — 4 recent low-citation peer-reviewed articles, selected for their relevance to LLMs and UX.

Item	Description
Sampling	After initial retrieval, a snowball sampling strategy was applied from Husain (2024) and Liang et al. (2022), whose references led to more recent evaluation frameworks (e.g., Schwartz et al., 2025; Dietz et al., 2025). Only one layer of snowballing was applied.
Replicability guidance	A reader attempting replication should: — Start with Husain (2024) and Liang et al. (2022); — Apply keyword queries on Google Scholar and arXiv from 2019–2025; — Apply the same filters: open access, LLM + UX focus, with exclusion of pure theory and paywalled texts; — Use GPT or Gemini to assist in unclear cases.
Relevance domain	Final sources covered: — LLM evaluation theory (6 papers); — UX alignment and assistant design (5 papers); — Human-computer interaction frameworks (4 papers); — Product/UX industry practices (5 sources).

To review the protocol, two AI assistants—ChatGPT-4.1 and Gemini Pro 2.5—were contextualized with the PRISMA method to provide observations on the inclusion and exclusion criteria. Following this, specific keywords and search strings were adopted to conduct the literature review.

Table 3 Keywords and search strings
Fonte The authors, 2025

Item	Description
Keywords	Searches used combinations of isolated and compound keywords, such as "LLM", "GPT", "large language evaluation", "UX assistant", "AI in design", "LLM evaluation UX", "design system assistant", and "LLM hallucination".
Some search strings	LLM in UX context: — ("LLM" OR "large language models") AND ("UX" OR "user experience") AND (evaluation OR metrics OR testing).
	Design assistants with AI: — ("design assistant" OR "AI assistant") AND ("UX" OR "UI") AND ("language model" OR "GPT") AND (evaluation OR assessment).
	Automatic vs. human evaluation: — ("LLM evaluation" OR "language model metrics") AND ("human evaluation" OR "automated scoring") AND ("UX research" OR "UX design").
	Performance and hallucination in LLM for UX: — ("LLM performance" OR "LLM hallucination") AND ("UX writing" OR "UX assistant") AND (truthfulness OR grounding OR consistency).

Item	Description
Some search strings	Language models comparison applied to Design: — ("GPT-4" OR "ChatGPT" OR "Gemini") AND ("design system" OR "UX writing") AND ("evaluation criteria" OR "performance analysis").
	Test simulation with AI assistants: — ("LLM-based assistant" OR "conversational AI") AND ("UX" OR "user interface") AND ("usability testing" OR "task evaluation").
	Academic sources applied to AI and Design: — ("language model" OR "LLM") AND ("UX research" OR "design tools") AND ("evaluation framework" OR "evaluation methods") AND (conference OR journal OR "industry blog"). overlapping and repetitive content.

As detailed in Table 3, search operators were used to account for orthographic variations like singular and plural forms, while the Boolean connectors AND and OR were utilized to combine and intersect content, ensuring a more comprehensive and precise search strategy.

Theoretical reference: Challenges in LLM Evals for UX

Evaluating large language models (LLMs) remains a multidimensional challenge. Research in NLP and HCI highlights limitations in current approaches concerning data representativeness, response fidelity, consistency, and the reliability of automated metrics (LIN et al., 2022; WANG et al., 2023; MIZRAHI et al., 2024). Aligning evaluation criteria with UX and design objectives is also difficult, which compromises the practical applicability of the results (RICHARDS & WESSEL, 2025).

Representativeness of Evaluation Data and Scenario Coverage

Current benchmarks often fail to represent the true diversity of LLM inputs, with many using single prompt styles that mask performance variations. Mizrahi et al. (2024) demonstrated that reformulating instructions drastically alters outcomes, proposing aggregated metrics with multiple paraphrases for stability. It is also essential to test across different populations, as Bender et al. (2021) advocate for an "evaluation culture" that considers worst-case scenarios to reveal critical flaws and amplified biases.

In HCI, Richards & Wessel (2025) advocate for user-centered evaluations beyond traditional benchmarks. They recommend simulating multiple personas to detect "inclusivity bugs," where systems fail certain user groups. Evaluating LLMs thus requires representing diverse demographic subgroups, digital literacy levels, and contexts of use. Without this variety, a model may perform well in standard tests but fail real or marginalized users.

Response fidelity and grounding

Measuring the factual fidelity of LLM responses to avoid hallucinations is pivotal (GAO et al., 2024). Tasks like summarizing UX data, yet assessing factuality are complex as traditional metrics fail to capture veracity (GOYAL et al., 2022). The TruthfulQA benchmark demonstrated that advanced models like GPT-3 can confidently answer factually incorrect questions, highlighting the gap between apparent authority and actual truthfulness (LIN et al., 2022).

Recent solutions include specialized metrics and human evaluation protocols. More effective evaluations require the model to ground its responses in verifiable sources, especially in Retrieval-Augmented Generation (RAG) systems (GAO et al., 2024). While metrics like "faithfulness" have been proposed, specialized human evaluations—such as sentence-by-sentence verification—remain essential (GOYAL et al., 2022). The most promising approach combines automated tools to flag potential hallucinations with human evaluators to validate critical cases, a necessary step as models become more fluent, making errors harder to detect (GAO et al., 2024).

Consistency and Context of Responses

An LLM's consistency is its ability to maintain coherence and avoid self-contradiction during continuous interactions, a crucial aspect in conversational assistants (WANG et al., 2023). Inconsistencies arise when a model gives divergent answers to equivalent questions or ignores prior information (MIZRAHI et al., 2024). Common detection methods include tests with paraphrasing, repetition, and contradictory premises (RIBEIRO et al., 2020). While not an evaluation metric, the concept of "self-consistency" suggests convergent answers are more reliable (WANG et al., 2023).

Evaluation can involve checking response stability through prompt ensembles; significant divergences from minor changes indicate model fragility. Context memory in long interactions is assessed by verifying if the model correctly references information from previous turns (RICHARDS & WESSEL, 2025). This is a critical challenge for design: modern evaluations, therefore, should include multi-turn use cases and metrics like coherence scores or contradiction rates, leveraging NLP tools for detection (RICHARDS & WESSEL, 2025).

Automated metrics vs. Human Evaluation

A central challenge in evaluating LLMs is balancing automated metrics with human judgments. Traditional NLP metrics like BLEU or ROUGE have limitations for generative models (GAO et al., 2024). Goyal et al. (2022) reported that GPT-3, despite lower ROUGE scores, achieved superior human ratings, showing that such metrics may not reflect real-world utility.

Schwartz et al. (2025) warn that outdated metrics can favor models that maximize scores without delivering quality.

Given these limitations, using LLMs as automated evaluators is a growing trend. This approach is scalable and shows good correlation with human evaluations, but its reliability is questioned. Dietz et al. (2025) warn of circularity and self-reinforcing biases when systems are optimized to please an LLM evaluator rather than the user. This can lead to "evaluator overfitting," where a model is rated highly because it was trained for that outcome.

Alignment with Design and Interaction Objectives

Authors highlight risks in using LLMs to evaluate other LLMs, such as reinforcing biases and a lack of reproducibility (DIETZ et al., 2025). They do not replace the need for human evaluators for subtle criteria like ethics. Consequently, a hybrid approach is ideal: automated metrics handle initial screening, while human evaluation remains the standard for complex quality (RICHARDS & WESSEL, 2025). This model should combine quantitative metrics with qualitative indicators like perceived utility, cognitive load, and usability.

HCI researchers propose expanding AI evaluations with human-centered metrics through "Human Interaction Evaluations" (HIEs), which assess models within simulated or real interactions (RICHARDS & WESSEL, 2025). This method observes not only accuracy but also impacts on human behavior and risks of overreliance, based on the rationale that evaluating static outputs is insufficient (SCHWARTZ et al., 2025). It needs a new ecosystem that integrates traditional benchmarks with usage studies to understand real-world effects (SCHWARTZ et al., 2025).

In Product Design, UX Writing, and UX Research, evaluation metrics must reflect specific design and business objectives. For a Design System assistant, this means adherence to visual guidelines (FESSENDEN, 2021); for UX Writing, tone and style guide consistency are crucial (GIBBONS, 2023); and for UX Research, utility can be measured by time saved or insights generated (SPONHEIM, 2024).

Scoring and Primary Definitions

To implement the evaluation metrics, a scoring system will be developed and tested through empirical iterations (ZHANG et al., 2023). Initial evaluation cycles will be conducted via heuristic analyses where specialists assign and justify scores. In later phases, human assessments will be reserved for semantic evaluations where automatic scoring is unreliable.

A key example is usability testing (MORAN, 2019), which will be the dominant method for qualitative user evaluation. The scoring will use Likert-type scales (SAURO & LEWIS, 2016) to quantify quality and performance, while user-based tests will focus on the five foundational usability components: learnability, efficiency, memorability, errors, and satisfaction (Nielsen, 2012).

Evaluations for a Digital Product Design System Assistant

A Product Design assistant for Design Systems (DS) aims to automate and scale the use of an interface component library, helping teams create consistent interfaces faster. The evaluation must measure the assistant's ability to generate artifacts faithful to the DS, accelerate workflows, and reduce manual verification. This includes ensuring generated designs meet accessibility standards (e.g., WCAG):

- **Component & Screen Generation**
  Creating layouts from prompts, using a prompt benchmark set with "gold-standard" designer versions for comparison.
- **Compliance Verification**
  Analyzing designs for inconsistencies against the Design System library, which serves as the "ground truth."
- **Inconsistency Detection**
  Using existing, approved product screens to test the assistant's ability to identify DS violations.
- **Component Suggestion & Consultation**
  Recommending appropriate components and answering questions about DS rules and usage.
- **Continuous Feedback**
  Collecting metrics and reports from designers on the assistant's daily utility and precision.

Based on these tasks, the following specific evaluation criteria are defined:

Table 4 Evaluation criteria and metrics (Design System Assistant)
Fonte The authors, 2025.

Item	Description	Evaluation System
Compliance with DS	Assess whether generated or revised designs adhere to the Design System rules (e.g., colors, fonts, spacing).	Automated Metric: Percentage of screens and features, whether generated or revised, that comply with the training models. Human Metric: Verification by specialists from the Design System Ops team. Goal: >95% compliance with the DS tokens and rules.
Correct component usage	Verify if the suggested or applied components are correct for the intended function and context.	Automated Metric: Percentage of suggested or applied components that are compliant with those available in the DS. Human Metric: Semantic evaluation by senior designers on the appropriateness of the component for the given context. Goal: >90% of suggestions considered semantically correct and appropriate.

Item	Description	Evaluation System
Workflow acceleration	Measure whether the assistant streamlines the designer's work by reducing manual effort.	Automated Metric: Measurement of time taken to complete design tasks (with and without AI). Human Metric: Satisfaction and perceived utility surveys conducted with designers. Goal: A 20% reduction in the time required to create standard screens and components.
Artifact quality	Evaluate the technical quality of the generated file (e.g., layer organization and use of auto-layout in Figma).	Automated Metric: Analysis by design linters, if applicable. Human Metric: Review of the file structure by experienced designers. Goal: Files are generated with a clear structure and require no manual reorganization.
Layout coherence	Verify if layouts generated from a prompt are logical, functional, and visually coherent.	Automated Metric: Percentage of layout compliance with usability heuristics (e.g., Nielsen's), Gestalt principles, and Semiotics. Human Metric: Evaluation by UI/UX specialists. Goal: >80% of layouts rated as "ready for use."

This allows for a mixed-method evaluation dashboard, combining automated validations (e.g., token analysis, linters) with expert reviews. This approach ensures the generated artifacts uphold the quality, consistency, and usability standards defined by the Design System, adding tangible value to the product development process.

Evaluations for a UX Research Assistant

A UX Research assistant is designed to help researchers analyze data, offering insights and summaries while also training users on institutional methodologies. Its objectives are to accelerate analysis, ensure findings are faithful to the data, support design decisions, and drive research compliance. The evaluation system must measure performance across these dimensions, verifying if the assistant saves time, preserves data integrity, and contributes reliably to decision-making:

- **Data Summarization & Insight Extraction**
  Generating summaries and extracting themes from qualitative data, using annotated real-world interviews or previous studies as a baseline for fidelity.
- **Data-Driven Query Answering**
  Responding to questions based on collected evidence, using open datasets (e.g., AMI Meeting Corpus, CoQA) to test consistency and summarization.

- **Report Generation & Data Organization**
  Compiling findings into reports and assisting with tagging or categorizing user feedback, using LLM-generated synthetic data to ensure realism and persona diversity.
- **Teaching & Compliance**
  Providing information on research processes according to institutional rules, ensuring ethical standards are met, and verifying for biases in analysis.
- **Contextual Human-in-the-Loop Evaluation**
  Conducting studies with researchers to collect qualitative feedback on the utility and precision of the generated insights.

Based on these tasks, the following specific evaluation criteria are defined:

Table 5 Evaluation criteria and metrics (UX Research Assistant)
Fonte The authors, 2025.

Item	Description	Evaluation System
Factual fidelity to data	Ensures the assistant does not invent or distort information from the source research.	Automated Metric: Comparative simulation against a core knowledge base, where the response is compared with multiple ground-truth versions to measure deviation. Human Metric: Assessment of the conformity of outputs for critical responses against established model answers. Goal: 0% to 3% hallucination rate.
Coverage and representation	Ensures that the main themes raised by users are represented in the output.	Automated Metric: Recall/Precision of key themes identified by the LLM. Human Metric: Recall/Precision of critical key themes identified by the ResearchOps team. Goal: >90% coverage of themes present in the data.
Insight relevance and unity	Measures whether the generated insights are useful, applicable, and effectively answer the core research questions.	Automated Metric: Measurement and simulation of insight impact using established risk analysis and ROI methods. Human Metric: Evaluation by designers (on a scale of 1–5). Goal: Average score of ≥ 4/5 and utility rated as equal to or greater than human-level work.

Item	Description	Evaluation System
Clarity and coherence of presentation	Verifies that the response is easy to understand, free of ambiguities or contradictions.	Automated Metric: Readability, intelligibility, and conciseness scores of the responses. Human Metric: Scores assigned for the same indices by human evaluators. Goal: Communication across three complexity levels, where the depth of the response correlates with its technical complexity.
Consistency and context memory	Ensures coherence across responses in multi-turn or sequential query sessions.	Automated Metric: Dialogue scans to verify contradictions or loss of context. Human Metric: Comparison of the most problematic dialogues to establish correction patterns. Goal: Zero contradictions in test sessions; context memory maintained for 'n' turns.

An evaluation dashboard can be assembled for the UX Research assistant, combining automated metrics (like readability and contradiction detection) with human assessments (utility and fidelity). By applying the assistant to real or simulated cases, the system would generate comprehensive performance reports. This approach ensures the model produces insights that are useful, reliable, and aligned with user data.

Evaluations for a UX Writing Assistant

A UX Writing assistant helps create clear, concise interface text, such as button labels and error messages, that aligns with a brand's tone of voice. The evaluation must verify if the assistant generates suggestions consistent with UX Writing guidelines, maintains textual quality, and accelerates the writing workflow. The evaluation should also include bias verification to ensure the assistant does not produce insensitive or inappropriate language:

- **Microcopy & Variation Generation**
  Proposing text for UI components based on context and creating alternative options, evaluated against custom benchmarks with reference texts from senior UX writers.
- **Text Refinement & Tone Adaptation**
  Suggesting improvements for clarity and converting text to align with the brand's style guide, using

"situation/recommended text" pairs from the guide as the ground truth.

- **Terminological Consistency**
  Ensuring uniform use of key terms across the product, validated using datasets of high-quality microcopy from established applications.
- **Continuous Usage Feedback**
  Collecting data on how designers interact with suggestions (accepting, editing, or discarding them) during the actual use of the tool.

Considering these functions, the proposed evaluation criteria are:

Table 6 Evaluation criteria and metrics (UX Writing Assistant)
Fonte The authors, 2025.

Item	Description	Evaluation System
Adherence to Tone of Voice	Evaluates whether the text adheres to the brand's style guide (e.g., voice, persona).	**Automated Metric:** Automated appropriateness classifier. **Human Metric:** Analysis by specialists based on the style guide. **Goal:** >80% approval rate by specialists; zero use of forbidden terms.
Clarity and Conciseness	Verifies if the text is clear, objective, and easy to understand.	**Automated Metric:** Scores on readability indices; word and character counts. **Human Metric:** Comprehension tests with users. **Goal:** Texts within defined character limits; >90% comprehension rate in user tests.
Grammatical Correctness	Assesses the grammatical and orthographic correctness of the text.	Automated Metric: Analysis with automated checkers (e.g., error count). Human Metric: Human input for cases not yet mapped by automated tools. Goal: Zero grammatical or orthographic errors in the evaluated texts.
Consistency	Measures whether terms and tone of voice are used consistently throughout the product.	Automated Metric: Automated check for the use of standard terms. Human Metric: Analysis of tonal consistency by specialists. Goal: >97% usage of defined terms; consistent tone of voice throughout a task flow.
Utility of the Suggestion	Assesses whether the suggestions are useful and if they streamline the designers' workflow.	Automated Metric: Percentage of suggestions accepted/edited; measurement of time taken to complete tasks with and without AI. Human Metric: Satisfaction surveys with users (designers). Goal: >70% of suggestions used in practice; a 20% reduction in work time.

This allows for an evaluation dashboard that combines automated checks (e.g., orthography, consistency) with expert evaluations of tone, clarity, and utility. By submitting the assistant to curated design scenarios, the system can generate performance reports, ensuring the model produces text that is correct, useful, and aligned with the brand identity.

Conclusions

This study examined the challenges of evaluating Large Language Models (LLMs) for UX assistants, proposing tailored evaluation systems for Design Systems, UX Research, and UX Writing that combine automated and human metrics. The work aims to provide practical tools for design teams, addressing a notable gap in the academic literature, which currently lacks applied, peer-reviewed frameworks for UX-centered LLM evaluation. Methodologically, while this paper used elements of the PRISMA framework and AI assistants (GPT-4.1, Gemini 2.5) to guide its review, it is not a full systematic review due to scope and transparency limitations.

Future work involves refining a dedicated, empirically-developed scoring system. This framework will be advanced through iterative testing, initially using heuristic analyses by specialists before incorporating Likert-type scales (Sauro & Lewis, 2016) to quantify quality, coherence, and performance. This will be complemented by usability tests on the five classic dimensions: learnability, efficiency, memorability, errors, and satisfaction (Nielsen, 2012). The ultimate goal is to create a scalable and reliable framework that reserves human input for complex semantic tasks while automating quantitative checks, guiding the responsible integration of LLMs into UX.

Referências

BENDER, Emily M. et al. ***On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?*** In: ACM CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY (FAccT), 2021. Virtual Event, Canada. Proceedings [...]. New York: ACM, 2021. p. 610–623. DOI: 10.1145/3442188.3445922.

BROWN, Tim. **Design Thinking: uma metodologia poderosa para decretar o fim das velhas ideias.** Rio de Janeiro: Alta Books, 2020.

COSKUN, M. N. K. ***Spotify's latest layoff is an opportunity to reconsider how we look at UX research.*** Fast Company, 2023. Available at: https://www.fastcompany.com/90999717/spotifys-latest-layoff-ux-research. Accessed on: Jun. 25, 2025.

CRUTH, M. ***Discover the Spotify model.*** Atlassian, 2022. Available at: https://www.atlassian.com/agile/agile-at-scale/spotify. Accessed on: Jun. 25, 2025.

DIETZ, Laura et al. **LLM-*Evaluation Tropes: Perspectives on the Validity of LLM-Evaluations.*** 2025. Preprint. DOI: https://doi.org/10.48550/arXiv.2504.19076.

FESSENDEN, T. ***Design Systems 101.*** Nielsen Norman Group, 2021. Available at: https://www.nngroup.com/articles/design-systems-101/. Accessed on: Jun. 26, 2025.

GAO, Mingqi et al. ***LLM-based NLG Evaluation: Current Status and Challenges.*** 2024. Preprint. DOI: https://doi.org/10.48550/arXiv.2402.01383.

GOTHELF, Jeff; SEIDEN, Josh. **Lean UX: projetando ótimos produtos com times ágeis.** 3. ed. Rio de Janeiro: Novatec Editora, 2022.

GOYAL, Tanya; LI, Junyi Jessy; DURRETT, Greg. ***News Summarization and Evaluation.*** 2022. Preprint. DOI: https://doi.org/10.48550/arXiv.2209.12356.

HULLEY, Stephen B. et al. **Delineando a pesquisa clínica.** 4. ed. Porto Alegre: Artmed, 2015.

HUSAIN, Hamel. ***Your AI Product Needs Evals.*** Hamel.dev, 2024. Available at: https://hamel.dev/blog/posts/evals/#motivation. Accessed on: Jun. 11, 2025.

LIN, Stephanie; HILTON, Jacob; EVANS, Owain. TruthfulQA: ***Measuring How Models Mimic Human Falsehoods.*** In: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS (ACL), 60., 2022, Dublin. Proceedings [...]. Stroudsburg, PA: Association for Computational Linguistics, 2022. p. 3214–3252. DOI: https://doi.org/10.18653/v1/2022.acl-long.229.

LIANG, Percy et al. ***Holistic Evaluation of Language Models.*** 2022. Preprint. Available at: https://arxiv.org/abs/2211.09110. Accessed on: Jun. 25, 2025.

MIZRAHI, Moran et al. ***State of What Art? A Call for Multi-Prompt LLM Evaluation. Transactions of the Association for Computational Linguistics***, Cambridge, MA, v. 12, p. 933–949, 2024. DOI: https://doi.org/10.1162/tacl_a_00681.

MORAN, Kate. ***Usability (User) Testing.*** Nielsen Norman Group, 2019. Available at: https://www.nngroup.com/articles/usability-testing-101/. Accessed on: Jul. 18, 2025.

_____. ***Calculating ROI for Design Projects in 4 Steps.*** Nielsen Norman Group, 2020. Available at: https://www.nngroup.com/articles/calculating-roi-design-projects/. Accessed on: Aug. 26, 2025.

_____. ***The Four Dimensions of Tone of Voice.*** Nielsen Norman Group, 2023. Available at: https://www.nngroup.com/articles/tone-of-voice-dimensions/. Accessed on: Jun. 26, 2025.

NG, Andrew. ***Machine Learning Engineering for Production (MLOps) Specialization: Course 1.*** Coursera; DeepLearning.AI, 2025. Online course. Available at: https://www.coursera.org/learn/introduction-to-machine-learning-in-production. Accessed on: Jun. 25, 2025.

NIELSEN, Jakob. ***Usability 101.*** Nielsen Norman Group, 2012. Available at: https://www.nngroup.com/articles/usability-101-introduction-to-usability/. Accessed on: Jul. 18, 2025.

NIJS, Diane. ***Lens: a big shift in science – seeing change and innovation as a matter of emergence.*** In: _____ (Ed.). Advanced Imagineering: Designing Innovation as Collective Creation. Cheltenham, UK: Edward Elgar Publishing, 2019. chap. 2, p. 27–55.

RIBEIRO, Marco Tulio et al. Beyond Accuracy: Behavioral ***Testing of NLP Models with CheckList.*** In: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS (ACL), 58., 2020. Proceedings [...]. Stroudsburg, PA: Association for Computational Linguistics, 2020. p. 4902–4912. DOI: https://doi.org/10.18653/v1/2020.acl-main.442.

RICHARDS, Julian; WESSEL, Merten. ***Bridging HCI and AI Research for the Evaluation of Conversational SE Assistants.*** In: INTERNATIONAL WORKSHOP ON BOTS IN SOFTWARE ENGINEERING (BOTSE), 6., 2025, Ottawa. Proceedings [...]. IEEE/ACM, 2025. (In press). Available at: https://arxiv.org/abs/2502.07956. Accessed on: Aug. 28, 2025.

SAURO, Jeff; LEWIS, James R. ***Quantifying the User Experience: Practical Statistics for User Research.*** 2. ed. Cambridge, MA: Morgan Kaufmann, 2016.

SCHULER, Douglas; NAMIOKA, Aki (Org.). ***Participatory Design: Principles and Practices.*** Hillsdale, NJ: Lawrence Erlbaum Associates, 1993.

SCHWARTZ, Reva et al. ***Reality Check: A New Evaluation Ecosystem is Necessary to Understand AI's Real World Effects.*** 2025. Preprint. DOI: https://doi.org/10.48550/arXiv.2505.18893.

SPONHEIM, Christina. ***What Is User Research?*** Nielsen Norman Group, 2024. 1 video (2 min 37 s). Available at: https://www.nngroup.com/videos/what-is-user-research/. Accessed on: Jun. 26, 2025.

TRICCO, Andrea C. et al. ***PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation.*** Annals of Internal Medicine, Philadelphia, PA, v. 169, n. 7, p. 467–473, Oct. 2018. DOI: https://doi.org/10.7326/M18-0850.

WANG, Xuezhi et al. Self-Consistency Improves Chain of Thought Reasoning in Language Models. 2022. Preprint. Available at: https://doi.org/10.48550/arXiv.2203.11171. Accessed on: Aug. 28, 2025.

WOOD, Ian. M***etadesign: Designing in the Anthropocene.*** London: Routledge, 2020.

ZHANG, Ziyin et al. ***Rethinking the A/B Test for Large Language Model-based Features: A Case Study from GitHub Copilot.*** In: ACM JOINT EUROPEAN SOFTWARE ENGINEERING CONFERENCE AND SYMPOSIUM ON THE FOUNDATIONS OF SOFTWARE ENGINEERING (ESEC/FSE), 31., 2023, San Francisco. Proceedings [...]. New York: ACM, 2023. p. 1000–1012. DOI: https://doi.org/10.1145/3611643.3616319.

Recebido: 28 de agosto de 2025
Aprovado: 27 de junho de 2026