

Uso de grandes modelos de linguagem na análise da experiência do usuário

Yvens Almeida Girão
Universidade Federal do Ceará (UFC)
Russas, CE, Brasil
yvensalmeida21@gmail.com

Denise da Silva Santos
Universidade Federal do Ceará (UFC)
Russas, CE, Brasil
denisevania5@gmail.com

Pedro Oliveira
Universidade Estadual de Campinas
(Unicamp)
Campinas, SP, Brasil
P300636@dac.unicamp.br

Marília Mendes
Universidade Federal do Ceará (UFC)
Fortaleza, CE, Brasil
marilia.mendes@virtual.ufc.br

Alexandre Arruda
Universidade Federal do Ceará (UFC)
Quixadá, CE, Brasil
alexandre.arruda@ufc.br

ABSTRACT

This work investigates the application of Large Language Models (LLMs) to the classification stage of the MALTU method for analyzing Postings Related to Use (PRUs). Using 299 reviews extracted from the Gov.br app page on Google Play, annotated by human evaluators as a gold standard, we evaluated five LLMs (GPT-4o, GPT-4o-mini, GPT-OSS-120B, Kimi K2, and Sabiazinho-4) under two prompting strategies (zero-shot and few-shot). Models were instructed to return structured JSON outputs for the six MALTU dimensions: type, intention, sentiment, functionality, usability, and UX. In total, models produced 2,990 classifications that were compared against the human gold standard using accuracy, precision, recall, and macro F1. Average accuracy ranged approximately from 74% to 78% (Kimi K2 achieved the best performance in the few-shot setting: 77.8%; GPT-OSS-120B led the zero-shot setting: 76.8%). Vocabulary dimensions (intention, type, sentiment, and UX) exhibited high performance (~85–95%), whereas functionality (33–49%) and usability (57–65%) were more challenging due to terminological variation and semantic ambiguity.

KEYWORDS

Large Language Models; MALTU; automated classification; user experience; prompt engineering.

1 INTRODUÇÃO

Grandes modelos de linguagem (*Large Language Models*, LLMs) são redes neurais treinadas em enormes volumes de texto para modelar e gerar linguagem natural, mostrando grande capacidade de compreensão contextual, generalização e síntese de informações [28]. Eles vêm sendo aplicados em diversas áreas, que variam desde a Saúde e a Medicina até a Computação e a Educação. Na Computação, muitos trabalhos têm investigado o uso de LLMs na geração de código, testes de software, otimização de redes neurais, sistemas de recomendação, jogos, planejamento, entre outros [30]. Na área de Interação Humano-Computador (IHC), os estudos se concentram desde o apoio à escrita e produtividade até aplicações voltadas para

aprendizagem, desenvolvimento de software, interfaces acessíveis e suporte à saúde e bem-estar [23]. Os autores [23] complementam que os LLMs estão sendo aplicados em diferentes contextos da IHC, abrangendo tanto tarefas técnicas quanto sociais e criativas.

Ao mesmo tempo, o ecossistema digital produz um volume crescente de conteúdo gerado por usuários (*User-Generated Content*, UGC), como avaliações em lojas de aplicativos, comentários em *e-commerce* e postagens em fóruns de suporte, que constituem uma fonte rica porém custosa e demorada de evidências sobre a experiência de uso (*User eXperience*, UX) de sistemas [15]. O MALTU (Modelo para Avaliação da Interação em Sistemas Sociais a partir da Linguagem Textual do Usuário) propõe um método estruturado para analisar Postagens Relacionadas ao Uso (PRUs), cobrindo extração, classificação, interpretação e relato de resultados [15]. Entretanto, a etapa de classificação e a consequente interpretação de PRUs dependem de especialistas em IHC, o que torna o processo escalável apenas com grande esforço humano. Essa limitação cria uma oportunidade para investigar se LLMs podem automatizar e enriquecer o pipeline do MALTU, reduzindo tempo e custo sem sacrificar a qualidade analítica.

Nesse contexto, este artigo investiga a adaptação de um instrumento de avaliação de IHC, o MALTU, para utilização com modelos LLMs. Dessa forma, o trabalho contribui para a integração de tecnologias baseadas em IA aos processos de avaliação da Usabilidade e UX (UUX), ampliando as possibilidades de análise automatizada e apoio à pesquisa em IHC. Para tanto, este estudo aborda as seguintes Questões de Pesquisa (QP):

- **QP1:** Quais modelos apresentam melhor desempenho na classificação das PRUs de acordo com as categorias do MALTU (tipo, intenção, sentimento, funcionalidade e metas de UUX)?
- **QP2:** Qual o desempenho dos modelos na classificações de PRUs, e como esses resultados se comparam à interpretação humana?
- **QP3:** Quais são os principais desafios e oportunidades da integração de LLMs no método MALTU?

Este artigo está organizado da seguinte forma: a próxima seção apresenta o background, cobrindo o método MALTU, fundamentos sobre LLMs e Engenharia de prompt. A terceira seção apresenta trabalhos relacionados a este. Na quarta seção, descreve-se a metodologia utilizada, seguida dos resultados, discussão, conclusão e trabalhos futuros.

2 BACKGROUND

2.1 Método MALTU

O método MALTU [15] é uma abordagem de avaliação no campo da IHC. Ele foi proposto na premissa de que as expressões textuais dos usuários, publicadas em plataformas digitais, constituem uma fonte de dados para mensurar a qualidade de um determinado sistema. Para aplicar o método, o avaliador precisa cumprir cinco etapas essenciais: 1) Definição do contexto de avaliação; 2) Extração das PRUs; 3) Classificação das PRUs; 4) Interpretação; e 5) Relato dos resultados. Como o objetivo deste trabalho consiste em apoiar a etapa de classificação, detalha-se neste artigo apenas esta etapa.

As PRUs são classificadas por, pelo menos, dois avaliadores, tendo um terceiro para validar [15] nas seguintes categorias: a) tipo: classifica a natureza da postagem (crítica, elogio, dúvida, ajuda, comparação ou sugestão); b) intenção: faz uma classificação de intenção do usuário, adaptando a teoria de Norman [17] na classificação de PRUs em visceral, comportamental e reflexiva; c) Análise de Sentimentos, avalia a polaridade do texto, classificando a PRU em positiva, negativa ou neutra; d) Funcionalidade, que indica qual funcionalidade específica do sistema o usuário está mencionando, como, por exemplo: "login", "busca", "pagamento" etc.; e) Critérios de Qualidade de Uso, que classifica a PRU em facetas de usabilidade: eficácia, eficiência, satisfação ou segurança e UX: motivação, suporte, frustração, e f) Artefato, que identifica o dispositivos utilizado (celular, computador, TV etc.) [15].

2.2 Grandes Modelos de Linguagem

Segundo o ranking interativo da LMSYS Chatbot Arena [14], que é baseado na preferência humana e em votos ocultos, estes são os principais modelos de LLM do ano de 2026:

- **Família Claude (Anthropic):** As variantes mais recentes do *Claude Opus* (Claude Opus 4.6 e 4.7).
- **Família Gemini (Google DeepMind):** Consolidou-se com o *Gemini 3.1 Pro* e o *Gemini 3.5 Flash*.
- **Família GPT (OpenAI):** Mantém-se com os modelos da série *GPT-5* (incluindo as versões GPT-5.4 e GPT-5.2).

Como destacado em [25], a base estrutural desses modelos é a arquitetura Transformer, introduzida por [28]. Nela, utilizam-se tokens compostos por unidades de palavras ou subpalavras como dados de treinamento. Zhao e seus colegas [31] definem os dados de treinamento como centenas de bilhões (ou trilhões) de palavras e subpalavras (tokens). É através da leitura exaustiva dessas palavras que o modelo aprende quais expressões costumam aparecer juntas [31]. As grandes empresas de IA utilizam *datasets* (conjuntos de dados) extraídos publicamente da internet para utilização no treinamento, as principais fontes são: Páginas de Web, Enciclopédias, Livros, Artigos Científicos etc. [31]. Essas palavras ficam alocadas em um vetor numérico, de modo que uma identificação numérica possa ser relacionada a um determinado token, mapeando uma informação linguística para uma representação vetorial, agrupando palavras similares ou frases de situações parecidas. Contudo, segundo [28], a depender da ordem em que uma palavra é inserida em um LLM, poderia haver diferentes significados.

A fim de resolver essa problemática, a codificação posicional é utilizada, aplicando funções trigonométricas (senos e cossenos) para

inserir informações de posição em cada token, fazendo com que o modelo compreenda a ordem correta de um texto. Para além disso, o processamento interno do *Transformer* conta com um codificador, que tem como função principal gerar uma série de estados ocultos que representam o texto de entrada de maneira significativa [28].

Há também a presença do decodificador, responsável por processar tanto as incorporações de entrada com codificação posicional quanto as de saída. O mesmo possui um mecanismo de atenção que melhora a qualidade das saídas, utilizando o contexto recebido pelo codificador. Ele cria sequências de saída com base nas sequências de entrada codificadas, prevendo a próxima palavra de acordo com as geradas anteriormente (mecanismo conhecido como processamento autorregressivo) [28]. Cabe ressaltar que diversos LLMs modernos, como o GPT-3 e o GPT-4, utilizam uma arquitetura *decoder-only*, processando a entrada e gerando a saída na mesma estrutura [1].

Antes de estar apto para interação, o modelo passa por uma fase de pré-treinamento, consumindo volumes massivos de dados textuais, a fim de aprender a estrutura, a sintaxe e a semântica da linguagem. Em seguida, para que um LLM atinja um alto nível de desempenho, é necessário passar por um processo final chamado de alinhamento [22]. O alinhamento serve para ajustar o modelo fazendo com que ele entenda e responda aos prompts de acordo com os desejos do usuário. Em modelos de Código Aberto como o LLaMA¹ (mantido pela Meta), o ajuste é realizado por técnicas de Aprendizado por Reforço, por meio do Feedback Humano (RLHF) ou até mesmo um Feedback realizado pela própria IA (RLAIF) [22].

2.3 Engenharia de Prompt

A interação do usuário com LLMs ocorre por meio de "prompts", ou instruções em linguagem natural. Estes "prompts" servem como o início da conversa para que a máquina gere diversos formatos de saída, tais como: textos, imagens, documentos ou códigos [27]. A Engenharia de Prompt é focada em técnicas validadas que têm como objetivo aprimorar e aumentar a relevância das respostas retornadas pela LLM. Segundo Gao et al. [5], a aplicação dessas técnicas práticas permite obter respostas mais precisas e alinhadas com as expectativas humanas. Alguns exemplos são:

- (1) **Think Step by Step:** uma das técnicas mais utilizadas e fáceis de implementar [5]. Sua ideia se resume em adicionar apenas a frase "Pense passo a passo" ao final do prompt. Segundo [11] após a utilização dessa frase, houve um aumento na precisão do GPT-3 em diversas tarefas. Questões aritméticas obtiveram um salto relevante de precisão de 17,7% para 78,7% [11]. No entanto, em modelos mais recentes focados em raciocínio nativo, como o GPT-4, essa técnica se torna redundante, uma vez que o comportamento de decomposição do problema já é inserido no processo de inferência do modelo [20].
- (2) **Few-shot Learning:** esta técnica consiste no aprendizado com poucos exemplos. Desta forma, são fornecidos para a LLM apenas exemplos do que é desejado [1]. Segundo [1], ao utilizar esse método, apenas com saídas esperadas, a LLM se torna capaz de retornar respostas condizentes com a expectativa do usuário. Tal fato pode estar relacionado às diversas opções de saída para um determinado prompt,

¹<https://llama.meta.com/>

quando o usuário ensina a LLM como responder por meio de exemplos, isso reduz e restringe o espaço de saída da resposta ao formato demonstrado [1]. Segundo os autores, para garantir o entendimento da LLM, os exemplos precisam ser diversos e equilibrados, além de estarem devidamente inseridos no cenário desejado pelo utilizador [1].

3 TRABALHOS RELACIONADOS

As LLMs vêm inaugurando uma nova era de pesquisas no âmbito da experiência humana com computadores [18, 23], impactando diretamente a forma como a UX é analisada e compreendida. Seus principais usos estão relacionados ao funcionamento como motores centrais de sistemas interativos, ferramentas de apoio à pesquisa, participantes simulados em estudos, objetos de investigação científica e foco de análises sobre a percepção dos usuários [23].

Nesse contexto, diversas pesquisas têm explorado o uso de LLMs para otimizar processos de avaliação heurística [2], projetar e avaliar interfaces de usuário geradas por modelos de linguagem LLMs [4, 9, 10, 12, 24] e compreender diferentes aspectos relacionados à UX. Esses trabalhos fornecem fundamentos importantes para a construção do nosso modelo de análise baseado em LLMs.

O estudo [18], por exemplo, investiga o uso do ChatGPT no design de visualização de dados, evidenciando o potencial dessa ferramenta para auxiliar profissionais iniciantes na área a desenvolver reflexões metodológicas e encontrarem soluções adequadas para seus estudos. Além disso, o trabalho apresenta o conceito de *design-oriented LLM*, no qual o uso combinado entre humanos e LLMs cria uma nova forma de orientação voltada ao design. Entre os principais benefícios identificados estão a rápida geração de ideias e a adaptabilidade de prompts. Em relação ao nosso trabalho, embora a dimensão explorada seja distinta centrada na análise de comentários textuais produzidos por usuários, o estudo contribui significativamente ao demonstrar o potencial avaliativo das LLMs em contextos de IHC.

De forma semelhante, o trabalho [13] apresenta um sistema de recuperação semântica de interfaces baseado em modelos multimodais de linguagem. O sistema utiliza LLMs para analisar interfaces e reorganizá-las segundo seus aspectos semânticos, permitindo a identificação de padrões linguísticos presentes na construção de interfaces de usuário. Essa capacidade de reconhecer padrões e conceitos semânticos reforça o potencial das LLMs na criação de análises padronizadas e consistentes. Tal perspectiva dialoga diretamente com nossa proposta, uma vez que buscamos identificar padrões semânticos em comentários de usuários relacionados a diferentes dimensões da avaliação MALTU, como tipo de problema, UX, usabilidade, entre outros.

O trabalho de [9] teve como objetivo avaliar se prompts voltados à acessibilidade seriam capazes de melhorar a conformidade de interfaces com as diretrizes WCAG 2.1, além de investigar se os próprios LLMs conseguem identificar e corrigir problemas de acessibilidade em códigos gerados automaticamente. Em contexto brasileiro, [29] apresentou uma análise sobre o uso de código gerado por LLMs, buscando compreender os vieses presentes nas interfaces produzidas por meio de técnicas de engenharia de prompts. Embora a acessibilidade não constitua uma dimensão central do nosso trabalho, as abordagens metodológicas apresentadas nesses

estudos contribuem para ampliar as possibilidades de aplicação de LLMs em diferentes perspectivas de avaliação.

O trabalho [19] investiga a identificação automática de PRUs no contexto do método MALTU. Os autores construíram uma base anotada de 3.865 tweets (2.020 PRUs, 1.845 não-PRUs) com validação manual e compararam os algoritmos C-BLSTM a SVM, Naive Bayes, árvores, MLP, CNN, LSTM e BLSTM. Os autores concluíram que a combinação CNN+BLSTM melhora significativamente a detecção de PRUs em português frente a abordagens tradicionais e modelos individuais, indicando viabilidade para automatizar etapas do método MALTU.

No contexto deste trabalho, a LLM será utilizada como ferramenta de apoio à pesquisa, no que se refere à classificação de comentários de usuários. O objetivo é investigar a possibilidade de obter classificações compatíveis com aquelas realizadas por avaliadores humanos, contribuindo para tornar a aplicação do MALTU mais ágil e viável. Diferentemente de estudos como [6], que utilizam LLMs e técnicas de Processamento de Linguagem Natural (PLN) para análise de sentimentos positivos e negativos, nossa proposta busca avaliar múltiplas dimensões da UX a partir de comentários humanos, estabelecendo um novo parâmetro de análise no campo científico da IHC.

4 METODOLOGIA

Este trabalho apresenta um estudo empírico, quantitativo e comparativo que investiga o uso de LLMs na etapa de classificação do método MALTU. O experimento foi organizado em oito etapas: (1) seleção do sistema avaliado; (2) extração das postagens; (3) classificação humana das PRUs; (4) seleção das LLMs; (5) criação dos prompts; (6) classificação de PRUs com LLM; (7) interpretação de resultados com LLM; e (8) comparação dos resultados. O fluxo geral do procedimento é apresentado na Figura 1.

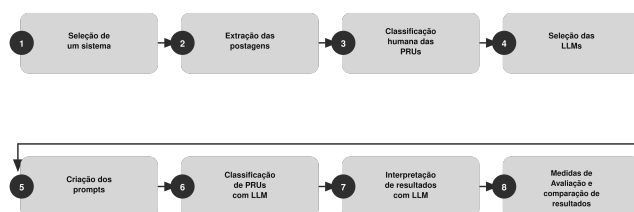


Figura 1: Metodologia

Na primeira etapa, o sistema escolhido foi o Gov.br, principal plataforma de serviços digitais do governo federal brasileiro, com mais de 100 milhões de downloads e 13,5 milhões de avaliações na Google Play Store [26]. Esses dados foram considerados indicadores de sua ampla disseminação e da disponibilidade de um volume expressivo de avaliações públicas para análise. Além disso, a escolha se justifica pela relevância do aplicativo como objeto de investigação em IHC, uma vez que ele reúne serviços digitais destinados ao atendimento de cidadãos em diferentes contextos de uso.

A segunda etapa envolveu extrair postagens do sistema escolhido. Para filtrar os reviews, pode-se escolher os seguintes filtros: dispositivo (telefone, chromebook ou tablet), postagens mais recentes, por avaliação ou mais relevantes. As mais recentes retornam,

como o próprio nome diz, as avaliações com a data mais próxima ao momento da extração, independente do seu nível de detalhe. O filtro por avaliação permite a escolha por nota, que varia de 1 estrela a 5 estrelas. O filtro mais relevante na Google Play Store utiliza algoritmos de aprendizado de máquina e IA para destacar avaliações e aplicativos que melhor representam a experiência geral dos usuários, focando na qualidade, atualização e relevância do conteúdo, em vez de apenas ordem cronológica [7]. Segundo [7], este filtro destaca opiniões detalhadas e recentes em detrimento de avaliações curtas ou antigas. Deste modo, na etapa de extração das postagens foi utilizado este filtro e no dia 13 de maio de 2026 foram extraídas as 400 primeiras postagens. A extração foi realizada usando a linguagem de programação Python e a plataforma online Google Colab [8].

As postagens extraídas continham respostas da própria plataforma (Serviços e Informações do Brasil). Por este motivo, tais postagens foram eliminadas, resultando em 299 PRUs. As PRUs extraídas tinham as seguintes colunas: autor, data, PRUs e curtidas. A data mais antiga foi 16/05/2023 e a data mais recente foi 09/05/2026. A maioria das PRUs se concentraram na data 16/03/2026. A maior quantidade de curtidas em uma PRU

("O aplicativo tem apresentando instabilidade, lentidão e não consigo concluir o login. Por causa disso, recebi uma notificação de acesso indevido e precisei RECUPE-RAR a senha. Só todo procedimento de SEGURANÇA EXIGE o aplicativo, e você não consegue concluir as etapas, e acabou que meu nível foi rebaixado para bronze, devido as falhas de recuperar a senha. Reconhecimento FACIAL não funciona independente do aparelho que você usa, sempre é BAIXA QUALIDADE.")

foi: 4751 e a menor quantidade de curtidas em uma PRU foi 0. A média de curtidas por PRU foi 81,28. Após esta análise as seguintes colunas foram eliminadas: autor, data e curtidas, sobrando então apenas a coluna das PRUs.

A terceira etapa envolveu a classificação das PRUs por pessoas, autores deste artigo. Eles classificaram as postagens² nas seguintes categorias do método MALTU: a) Tipo; b) Intenção; c) Análise de Sentimentos; d) Funcionalidade; e) Usabilidade; e f) UX. Artefato não foi colocado, pois as postagens já foram extraídas com o filtro Telefone.

A quarta etapa envolveu a seleção dos LLMs a serem avaliados. Foram escolhidos cinco modelos com critérios distintos, conforme apresentado na Tabela 1. O **GPT-4o** foi incluído por ser o modelo de maior capacidade da família GPT no período do estudo; o **GPT-4o-mini** foi selecionado por seu baixo custo, permitindo avaliar a relação entre desempenho e custo-benefício; o **GPT-OSS-120B** foi escolhido por ser um modelo *open-weight* da OpenAI [21], acessado via OpenRouter, oferecendo alto potencial com baixo custo de inferência; o **Kimi K2**, da Moonshot AI, foi incluído por ser um modelo de raciocínio recente, utilizado em sua variante com *thinking* habilitado; e o **Sabiazinho-4**, desenvolvido pela Maritaca AI, foi selecionado por ser o único modelo especializado em português brasileiro, permitindo avaliar se a especialização linguística impacta a acurácia na classificação de PRUs em língua portuguesa.

Tabela 1: Modelos selecionados e respectivos custos de inferência por milhão de tokens.

Modelo	Provedor	Input (USD)	Output (USD)	Input (BRL)	Output (BRL)
GPT-4o	OpenAI	\$2,50	\$10,00	R\$12,65	R\$50,60
GPT-4o-mini	OpenAI	\$0,15	\$0,60	R\$0,76	R\$3,04
GPT-OSS-120B	OpenRouter	\$0,039	\$0,18	R\$0,20	R\$0,91
Kimi K2 Thinking	OpenRouter	\$0,60	\$2,50	R\$3,04	R\$12,65
Sabiazinho-4	Maritaca AI	R\$0,20*	R\$0,79*	R\$1,00	R\$4,00

* Conversão aproximada. Sabiazinho-4 é precificado nativamente em BRL.
Cotação de referência: US\$1,00 = R\$5,06 (29/05/2026, Banco Central do Brasil).

A quinta etapa consistiu na elaboração do prompt³. Foram desenvolvidas duas estratégias de prompt: zero-shot, sem exemplos, e few-shot, com exemplos anotados de PRUs classificadas segundo o MALTU. Ambas as estratégias seguiram o mesmo esquema de saída estruturada em JSON, com as seis dimensões do método. A execução foi automatizada por meio da plataforma n8n [16], que orquestrou o envio de cada PRU para os cinco modelos em paralelo e consolidou os resultados em um arquivo .xlsx. Todos os modelos tiveram acesso à ferramenta Think do n8n, com exceção do GPT-OSS-120B, cujo modo de raciocínio nativo não estava disponível via OpenRouter na configuração utilizada.

A sexta etapa consistiu na classificação das 299 PRUs pelos cinco modelos selecionados. O processo foi executado em duas rodadas distintas, uma para cada estratégia de prompt: zero-shot e few-shot. Em cada rodada, todos os comentários foram enviados individualmente a todos os modelos em paralelo, resultando em 1.495 classificações por rodada (299 PRUs × 5 modelos) e 2.990 classificações no total. Os modelos foram instruídos a retornar as classificações no formato JSON estruturado, contendo as seis dimensões do método MALTU. Os resultados foram coletados e consolidados automaticamente pela plataforma n8n [16] em um arquivo .xlsx, com uma linha por PRU e colunas separadas para as classificações de cada modelo e do gabarito humano, permitindo a comparação direta na etapa seguinte.

A sétima etapa consistiu na interpretação dos resultados das classificações geradas pelos modelos. A partir das PRUs classificadas, buscou-se identificar padrões recorrentes nas dimensões do MALTU como os tipos de postagem predominantes, as funcionalidades mais mencionadas, os sentimentos expressos e as metas de usabilidade e UX mais frequentes reportadas pelos usuários do Gov.br. Essa interpretação seguiu as diretrizes do método MALTU [15] e foi aplicada tanto às classificações humanas quanto às geradas pelas LLMs, permitindo verificar se os modelos produzem interpretações coerentes com as do avaliador humano.

A oitava etapa consistiu na avaliação quantitativa e comparação dos resultados. As classificações das LLMs foram confrontadas com o gabarito humano por meio das métricas de acurácia, precision, recall e F1-score macro, calculadas por modelo e por dimensão do MALTU. A comparação foi realizada separadamente para as estratégias zero-shot e few-shot, permitindo avaliar o impacto da engenharia de prompt no desempenho dos modelos. Para as dimensões com vocabulário fechado (tipo, intenção e sentimento), adotou-se correspondência exata após normalização textual; para as dimensões com saída aberta ou multi-valor (funcionalidade, usabilidade e UX), utilizou-se correspondência parcial, considerando acerto quando ao menos um dos valores gerado pela LLM coincidia com

²<https://zenodo.org/records/20450962>

³<https://doi.org/10.5281/zenodo.20311698>

o gabarito humano. O processamento e a geração dos resultados foram realizados por meio de um notebook Python disponibilizado publicamente ⁴.

5 RESULTADOS

Os resultados das classificações realizadas pelos cinco modelos estão disponíveis publicamente no repositório Zenodo: estratégia zero-shot⁵ e estratégia few-shot⁶.

5.1 Zero-shot

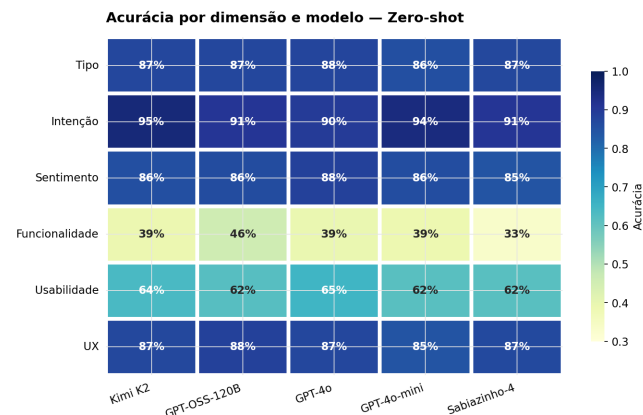


Figura 2: Acurácia por dimensão e modelo estratégia zero-shot.

A Figura 2 apresenta a acurácia por dimensão e modelo na estratégia zero-shot. O GPT-OSS-120B obteve a maior acurácia média (76,8%), seguido pelo Kimi K2 (76,4%) e GPT-4o (76,3%). O GPT-4o-mini (75,1%) e o Sabiazinho-4 (74,0%) completaram o ranking. As dimensões Intenção (90–95%), Tipo (86–88%), Sentimento (85–88%) e UX (85–88%) apresentaram os melhores resultados. Funcionalidade obteve os menores valores (33–46%), seguida por Usabilidade (62–65%).

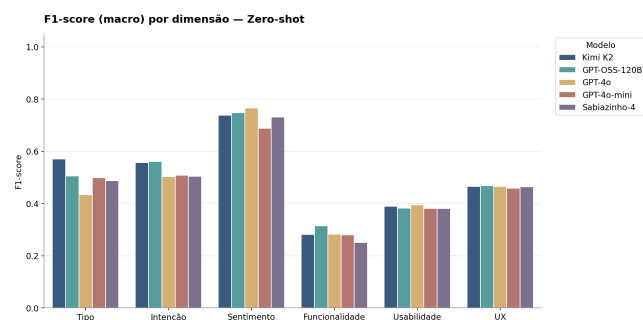


Figura 3: F1-score (macro) por dimensão e modelo estratégia zero-shot.

⁴<https://doi.org/10.5281/zenodo.20451981>

⁵<https://doi.org/10.5281/zenodo.20420785>

⁶<https://doi.org/10.5281/zenodo.20451936>

A Figura 3 apresenta o F1-score macro por dimensão na estratégia zero-shot. Para as dimensões de vocabulário fechado, Sentimento alcançou os maiores valores de F1 (0,69–0,77), seguido por Intenção (0,50–0,56) e Tipo (0,43–0,57). Para as dimensões multi-valor, o F1 manteve-se consistentemente abaixo de 0,50, com Funcionalidade apresentando os menores valores (~0,25–0,33), seguida por Usabilidade (~0,36–0,39) e UX, que se aproximou mais desse teto (~0,46). Esse padrão decorre da natureza da correspondência parcial adotada no método de avaliação, que penaliza saídas multi-valor mesmo quando parcialmente corretas, não refletindo necessariamente o desempenho real dos modelos nessas dimensões.

5.2 Few-shot

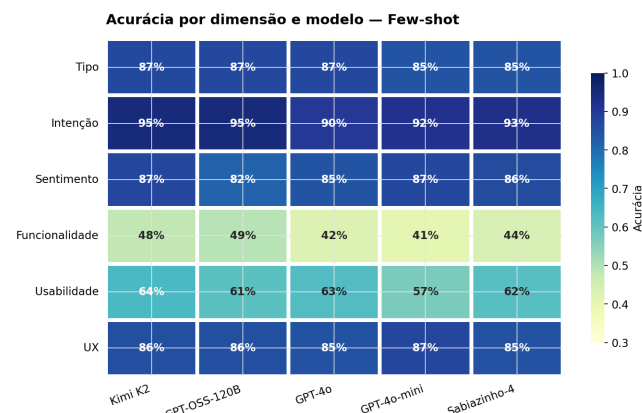


Figura 4: Acurácia por dimensão e modelo estratégia few-shot.

A Figura 4 apresenta a acurácia por dimensão e modelo na estratégia few-shot. O Kimi K2 obteve a maior acurácia média (77,8%), seguido pelo GPT-OSS-120B (76,7%) e Sabiazinho-4 (75,7%). O GPT-4o (75,3%) e o GPT-4o-mini (74,8%) completaram o ranking. As dimensões Intenção (90–95%), UX (85–87%), Sentimento (82–87%) e Tipo (85–87%) apresentaram os melhores resultados. Funcionalidade obteve os menores valores (41–49%), seguida por Usabilidade (57–64%).

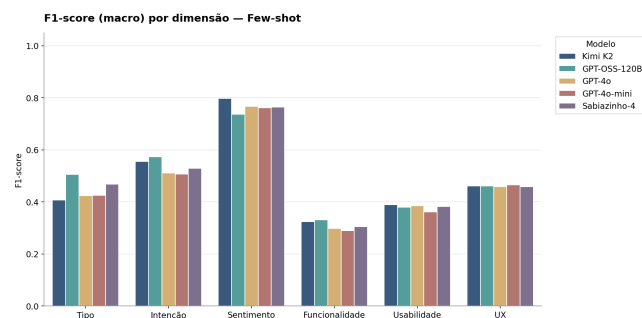


Figura 5: F1-score (macro) por dimensão e modelo estratégia few-shot.

A Figura 5 apresenta o F1-score macro por dimensão na estratégia *few-shot*. Para as dimensões de vocabulário fechado, Sentimento alcançou os maiores valores de F1 (0,74–0,80), seguido por Intenção (0,51–0,57) e Tipo (0,41–0,51). Assim como na estratégia *zero-shot*, os valores de F1 para Funcionalidade, Usabilidade e UX devem ser interpretados com cautela pela razão descrita anteriormente.

5.3 Comparação entre estratégias

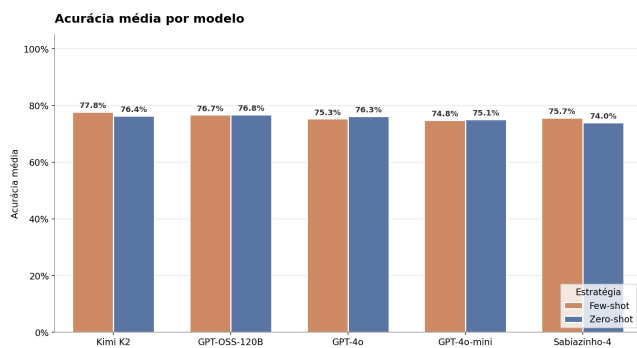


Figura 6: Acurácia média por modelo nas estratégias *zero-shot* e *few-shot*.

A Figura 6 compara a acurácia média dos cinco modelos nas duas estratégias. De modo geral, as diferenças foram pequenas, indicando que o uso de exemplos no prompt não impactou de forma expressiva o desempenho médio. O Kimi K2 e o Sabiazinho-4 beneficiaram-se mais do *few-shot*, com ganhos de 1,4 e 1,7 pontos percentuais, respectivamente. Os demais modelos obtiveram desempenho ligeiramente superior no *zero-shot*. A dimensão Funcionalidade foi a que mais se beneficiou do *few-shot*, com ganho médio de 7 pontos percentuais em relação ao *zero-shot*, sugerindo que exemplos concretos auxiliam os modelos na identificação de funcionalidades específicas. Vale destacar que a base de PRUs é desbalanceada em algumas dimensões, com predominância de uma classe majoritária em Tipo e Intenção. Isso pode inflar a acurácia, já que um modelo tende a acertar a classe mais frequente. Por isso, *precision* e *recall* são métricas mais representativas nesses casos, e melhorá-las permanece um desafio para a classificação automática de PRUs.

6 DISCUSSÃO

Em relação à QP1, os resultados indicam que o **Kimi K2** apresentou o melhor desempenho geral na estratégia *few-shot*, com acurácia média de 77,8%, enquanto o **GPT-OSS-120B** liderou na estratégia *zero-shot*, com 76,8%. No entanto, a diferença entre todos os modelos avaliados foi pequena, variando entre 74% e 78%, independentemente da estratégia utilizada. Esse resultado sugere que as LLMs modernas, mesmo com arquiteturas e tamanhos distintos, apresentam capacidade similar de compreender e aplicar as categorias do método MALTU a partir de instruções textuais.

Em relação à QP2, os resultados revelam que a estratégia *zero-shot* apresentou desempenho igual ou superior ao *few-shot* na maioria dos modelos avaliados. O GPT-OSS-120B, o GPT-4o e o GPT-4o-mini obtiveram acurácia ligeiramente maior no *zero-shot*,

enquanto apenas o Kimi K2 e o Sabiazinho-4 se beneficiaram do uso de exemplos no prompt, com ganhos de 1,4 e 1,7 pontos percentuais, respectivamente. Esse achado vai contra a expectativa inicial de que exemplos anotados melhorariam consistentemente o desempenho dos modelos, como sugere a literatura sobre *few-shot learning* [1]. Uma possível explicação é que o prompt *zero-shot* utilizado era suficientemente detalhado com definições claras, regras de desempate e vocabulário controlado reduzindo a necessidade de exemplos adicionais. Outra hipótese é que os exemplos do *few-shot*, por terem sido extraídos de um conjunto específico de PRUs do Gov.br, podem ter introduzido viés em alguns modelos, limitando sua capacidade de generalização para comentários com características distintas.

Em relação à QP3, a integração de LLMs ao método MALTU apresenta oportunidades relevantes para a avaliação de UX em larga escala. O principal ganho está na escalabilidade: enquanto a classificação manual de 299 PRUs por avaliadores humanos demanda horas de trabalho especializado, os modelos avaliados realizaram a mesma tarefa em questão de minutos, com custo financeiro reduzido. Modelos como o GPT-4o-mini e o Sabiazinho-4 representam alternativas economicamente viáveis, especialmente em projetos com grandes volumes de PRUs ou com restrições orçamentárias. Outro aspecto positivo é que a estratégia *zero-shot* apresentou desempenho comparável ao *few-shot*, indicando que não é necessária uma base de dados anotada para iniciar o uso dos modelos no contexto do MALTU. Isso reduz significativamente a barreira de entrada para pesquisadores e profissionais que desejam adotar o MALTU. Por outro lado, os principais desafios identificados estão na classificação de dimensões semanticamente ambíguas, como Usabilidade, e na padronização terminológica de dimensões abertas, como Funcionalidade, que continuam exigindo atenção cuidadosa na elaboração dos *prompts* e nos critérios de avaliação adotados. Esses achados dialogam diretamente com o GC6 do GrandIHC-BR 2025–2035 [3], que discute as implicações do uso de IA no campo da IHC. Ao demonstrar que LLMs podem apoiar uma etapa historicamente dependente de especialistas, este trabalho contribui para as ações de médio prazo voltadas à proposição de aplicações de IA em processos de design e avaliação. Ao mesmo tempo, os desafios identificados reforçam a necessidade de manter a supervisão humana como componente essencial, uma vez que a automação não elimina a interpretação especializada, mas redistribui o esforço do avaliador da classificação exaustiva para a validação crítica dos resultados.

Cruzando os custos de inferência da Tabela 1 com a acurácia obtida, o **GPT-OSS-120B** desponta como o melhor custo-benefício entre os modelos avaliados: com acurácia média de 76,7% no *few-shot*, apenas 1,1 ponto percentual abaixo do líder Kimi K2 (77,8%), seu custo de inferência (R\$0,20/R\$0,91 por milhão de tokens de entrada/saída) é aproximadamente 60 vezes menor que o do GPT-4o (R\$12,65/R\$50,60) e cerca de 15 vezes menor que o do Kimi K2 (R\$3,04/R\$12,65). Já o GPT-4o, apesar de seu custo elevado, obteve desempenho apenas mediano entre os cinco modelos (75,3% no *few-shot*), atrás inclusive de opções mais baratas como o Sabiazinho-4 e o GPT-OSS-120B, evidenciando que maior custo de inferência não se traduz necessariamente em maior acurácia na tarefa de classificação de PRUs.

A dimensão Funcionalidade apresentou os menores valores de acurácia em ambas as estratégias (33–49%), destacando-se como o principal desafio para todos os modelos avaliados. Diferentemente

das demais dimensões, que possuem vocabulário fechado e predefinido no prompt, Funcionalidade exige que o modelo identifique e nomeie livremente a funcionalidade central do comentário. Isso abre espaço para variações terminológicas entre modelos e avaliadores humanos, por exemplo, termos como “login”, “autenticação” e “acesso ao sistema” referem-se ao mesmo conceito, mas são tratados como respostas distintas no método de avaliação adotado. Esse comportamento é esperado em tarefas de extração de informação aberta e representa uma limitação inerente à comparação literal entre classificações humanas e automáticas. Para trabalhos futuros, recomenda-se o uso de grupos de sinônimos ou métricas baseadas em similaridade semântica para avaliar essa dimensão de forma mais justa.

A dimensão Usabilidade apresentou desempenho intermediário, com acurácia entre 57% e 65% nas duas estratégias. Embora possua vocabulário fechado, ao contrário de Funcionalidade, suas categorias apresentam proximidade semântica que dificulta a distinção precisa. Os casos mais frequentes de confusão ocorrem entre *falta de eficácia* e *falta de eficiência*: enquanto a primeira indica que o sistema não realiza a tarefa corretamente, a segunda indica que a realiza com esforço ou lentidão excessivos. Essa distinção é sutil e depende de uma leitura cuidadosa do comentário, o que representa um desafio mesmo para avaliadores humanos. O fato de todos os modelos terem errado nas mesmas proporções sugere que essa ambiguidade é sistemática e não específica de um modelo, indicando que o próprio critério de classificação pode se beneficiar de definições mais precisas ou de exemplos adicionais que ilustrem explicitamente a diferença entre essas categorias.

O **Sabiazinho-4** merece atenção especial por ser o único modelo do conjunto especializado em português brasileiro. Ele obteve um desempenho comparável aos modelos internacionais de maior porte, posicionando-se em terceiro lugar no few-shot (75,7%) e apresentando ganho de 1,7 pontos percentuais em relação ao zero-shot, o segundo maior ganho entre todos os modelos avaliados. Esse resultado sugere que modelos especializados em língua portuguesa se beneficiam mais de exemplos no prompt, possivelmente por terem maior sensibilidade ao contexto linguístico fornecido. Além do desempenho técnico, o Sabiazinho-4 apresenta o menor custo de inferência em moeda nacional entre os modelos avaliados (R\$1,00/R\$4,00 por milhão de tokens de entrada/saída), tornando-se uma alternativa viável e economicamente acessível para pesquisas em IHC no contexto brasileiro, especialmente em cenários com grandes volumes de PRUs a serem classificadas.

Este estudo apresenta algumas limitações que devem ser consideradas na interpretação dos resultados. A classificação humana utilizada como gabarito foi realizada por um único grupo de avaliadores, sem cálculo de concordância entre eles, o que pode introduzir subjetividade nas comparações. A temperatura dos modelos não foi fixada durante as execuções, o que significa que os resultados podem variar em execuções futuras com os mesmos prompts e comentários. Para as dimensões de saída multi-valor (Funcionalidade, Usabilidade e UX), o F1-score não reflete o desempenho real dos modelos devido à forma como a correspondência parcial é tratada no método de avaliação adotado, sendo reportada apenas a acurácia nessas dimensões. A dimensão Funcionalidade, por ser aberta, é avaliada por correspondência literal após normalização textual, o

que penaliza respostas semanticamente equivalentes mas terminologicamente distintas. Por fim, o experimento foi conduzido com 299 PRUs extraídas de um único sistema (Gov.br), o que limita a generalização dos resultados para outros domínios e contextos de uso.

7 CONCLUSÕES E TRABALHOS FUTUROS

Este trabalho investigou a integração de Grandes Modelos de Linguagem ao método MALTU, avaliando a capacidade de cinco LLMs em classificar automaticamente PRUs nas seis dimensões do método: tipo, intenção, sentimento, funcionalidade, usabilidade e UX. Para isso, foram utilizadas 299 PRUs extraídas do Gov.br, classificadas manualmente como gabarito e submetidas a dois cenários de prompt distintos: zero-shot e few-shot.

Os resultados indicam que as LLMs avaliadas são capazes de classificar PRUs com acurácia média entre 74% e 78%, com destaque para o Kimi K2 no cenário few-shot (77,8%) e o GPT-OSS-120B no cenário zero-shot (76,8%). As dimensões tipo, intenção, UX e sentimento apresentaram os melhores desempenhos, enquanto funcionalidade e usabilidade se mostraram as mais desafiadoras, refletindo a natureza aberta da primeira e a proximidade semântica entre as categorias da segunda. A comparação entre as estratégias de prompt revelou que o zero-shot apresentou desempenho igual ou superior ao few-shot na maioria dos modelos, sugerindo que prompts bem estruturados podem dispensar o uso de exemplos anotados. O Sabiazinho-4 obteve desempenho comparável aos modelos internacionais de maior porte, com o menor custo de inferência em moeda nacional, demonstrando ser uma alternativa viável para pesquisas em IHC no contexto brasileiro.

Conclui-se que LLMs representam uma alternativa promissora para automatizar a etapa de classificação do MALTU, com potencial para reduzir significativamente o tempo e o esforço necessários para avaliações textuais em larga escala, sem sacrificar de forma expressiva a qualidade analítica.

Para trabalhos futuros, recomenda-se: ampliar a base de PRUs para além das 299 utilizadas, abrangendo outros sistemas e contextos de uso; adotar múltiplos avaliadores humanos independentes com cálculo de concordância entre eles (Cohen's Kappa) para fortalecer o gabarito; investigar métricas adequadas para dimensões multi-valor, como o F1 multi-label; testar outros modelos de LLMs como Claude, Gemini e Llama; fixar a temperatura dos modelos para garantir reprodutibilidade dos experimentos; e avaliar o custo-benefício da abordagem em cenários reais de avaliação com grandes volumes de PRUs.

8 CUIDADOS ÉTICOS

Este estudo utilizou exclusivamente avaliações públicas disponibilizadas na Google Play Store, sem interação direta com os autores das avaliações, recrutamento de participantes ou coleta de dados adicionais. Em observância aos princípios da Lei Geral de Proteção de Dados Pessoais (LGPD), Lei no 13.709/2018, foram coletados e processados somente os conteúdos textuais necessários aos objetivos da pesquisa, aplicando-se o princípio da minimização de dados. Informações potencialmente identificáveis, como nomes de usuários, datas, número de curtidas e outros elementos não essenciais à análise, foram desconsideradas e não foram divulgadas. Os

dados foram tratados de forma desvinculada dos autores e apresentados de maneira agregada, sem tentativa de reidentificação ou associação entre comentários e respectivos usuários. Os arquivos utilizados no estudo foram armazenados em ambiente de acesso restrito aos pesquisadores e utilizados exclusivamente para as finalidades descritas neste trabalho. Na divulgação dos resultados, não foram reproduzidos conteúdos ou informações que pudessem permitir a identificação individual dos autores das avaliações.

ACKNOWLEDGMENTS

Neste trabalho, ferramentas de IA foram utilizadas como apoio em diferentes etapas. O ChatGPT foi utilizado para auxílio em correções gramaticais e melhoria da fluência textual. O Claude (Anthropic) foi utilizado para a geração do código Python desenvolvido para análise das métricas de avaliação, incluindo o cálculo de acurácia, precision, recall e F1-score, bem como a geração dos gráficos comparativos apresentados na seção de Resultados. Em ambos os casos, os conteúdos gerados foram revisados e validados pelos autores antes de sua utilização.

REFERÊNCIAS

- [1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems* 33 (2020), 1877–1901.
- [2] Peitong Duan, Jeremy Warner, and Björn Hartmann. 2023. Towards Generating UI Design Feedback with LLMs. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology Adjunct Proceedings (UIST '23 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 1–3. doi:10.1145/3586182.3615810
- [3] Emanuel Felipe Duarte, Paula Toledo Palomino, Taciana Pontual Falcão, Grace Lis Proença Meireles Barreto Porto, Carlos dos Santos Portela, Douglas Francisco Ribeiro, André Nascimento, Yuska Paola Costa Aguiar, Mauricio Souza, Angelita Moutin Segoria Gasparotto, and Armando Maciel Toda. 2024. GrandIHC-BR 2025–2035 - GC6: Implications of Artificial Intelligence in HCI: A Discussion on Paradigms, Ethics, and Diversity, Equity and Inclusion. In *Proceedings of the XXIII Brazilian Symposium on Human Factors in Computing Systems (IHC '24)*. ACM, Article 22, 19 pages. doi:10.1145/3702038.3702059
- [4] Lúcio França, Lisieux Andrade, and Windson Carvalho. 2025. Um Estudo Sobre o Uso de Prompts LLM para a Avaliação da Acessibilidade Web dos Portais das Prefeituras da Região dos Sertões dos Cratêus. In *Proceedings of the 31st Brazilian Symposium on Multimedia and the Web (Rio de Janeiro/RJ)*. SBC, Porto Alegre, RS, Brasil, 580–584. doi:10.5753/webmedia.2025.16064
- [5] Andrew Kean Gao. 2023. Prompt Engineering for Large Language Models: A brief guide with examples for non-technical readers. Stanford University Tech Report / Online Guide. Acedido em: Maio de 2026.
- [6] Pawanjit Singh Ghatore, Seyed Ebrahim Hosseini, Shahbaz Pervaz, Muhammad J. Iqbal, and Nabil Shaukat. 2024. Sentiment Analysis of Product Reviews Using Machine Learning and Pre-Trained LLM. *Big Data Cogn. Comput.* 8 (2024), 199. doi:10.3390/bdcc8120199
- [7] Google Play Console Help. 2024. Avaliações, notas e métricas de apps. <https://support.google.com/googleplay/android-developer/answer/9859673?hl=pt-BR>. Acessado em: 23 de maio de 2024.
- [8] Google Research. 2017. Google Colaboratory. <https://colab.research.google.com/>. Acesso em: 13 maio 2026.
- [9] Alexandra-Elena Gurita and Radu-Daniel Vatavu. 2025. When LLM-Generated Code Perpetuates User Interface Accessibility Barriers, How Can We Break the Cycle?. In *Proceedings of the 22nd International Web for All Conference (W4A '25)* (Sydney, NSW, Australia). ACM, New York, NY, USA. doi:10.1145/3744257.3744266
- [10] Nien-Lin Hsueh, Hsuen-Jen Lin, and Lien-Chi Lai. 2024. Applying Large Language Model to User Experience Testing. *Electronics* 13, 23 (24 Nov 2024), 4633. doi:10.3390/electronics13234633 Open access; published 24 November 2024.
- [11] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems* 35 (2022), 22199–22213.
- [12] Kristian Kolthoff, Felix Kretzer, Lennart Fiebig, Christian Bartelt, Alexander Maedche, and Simone Paolo Ponzetto. 2024. Zero-shot prompting approaches for llm-based graphical user interface generation. *arXiv preprint arXiv:2412.11328* (2024).
- [13] Yoonjoo Lee, Hyunjin Kim, Minji Park, and Juho Kim. 2025. Leveraging Multimodal LLM for Inspirational User Interface Search. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA. doi:10.1145/3706598.3714123
- [14] LMSYS Org. 2026. Chatbot Arena Leaderboard. <https://huggingface.co/spaces/lmarena-ai/chatbot-arena>. Acedido em: Maio de 2026.
- [15] Marília Soares Mendes. 2015. *MALTU – um modelo para avaliação da interação em sistemas sociais a partir da linguagem textual do usuário*. Ph.D. Dissertation. Universidade Federal do Ceará, Fortaleza, Brazil. Advisor(s) Castro, Miguel Franklin de. Tese (Doutorado em Ciência da Computação).
- [16] n8n GmbH. 2026. n8n: Workflow Automation Tool. <https://n8n.io>. Acedido em: 29 de maio de 2026.
- [17] Donald A. Norman. 2004. *Emotional Design: Why We Love (or Hate) Everyday Things*. Basic Books, New York, NY.
- [18] George Oliveira and Simone Barbosa. 2025. Evaluating ChatGPT to support data visualization design. In *Anais do XXIV Simpósio Brasileiro sobre Fatores Humanos em Sistemas Computacionais* (Belo Horizonte/MG). SBC, Porto Alegre, RS, Brasil, 779–797. doi:10.5753/ihc.2025.11894
- [19] Hismael Oliveira, Alexandre Arruda, and Marília Mendes. 2020. Identificação Automática de Postagens Relacionadas ao Uso Através de Modelos de Aprendizagem Profunda. In *Anais do Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)*. Salvador, 587. XVI Encontro Nacional de Inteligência Artificial e Computacional.
- [20] OpenAI. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774* (2023).
- [21] OpenAI. 2025. Introducing GPT-OSS. <https://openai.com/pt-BR/index/introducing-gpt-oss/>. Acedido em: 29 de maio de 2026.
- [22] Sinan Ozdemir. 2023. *Quick Start Guide to Large Language Models: Strategies and Best Practices for ChatGPT, Embeddings, Fine-Tuning, and Multimodal AI*. Addison-Wesley Professional.
- [23] Rock Yuren Pang, Hope Schroeder, Kynnedey Simone Smith, Solon Barocas, Ziang Xiao, Emily Tseng, and Danielle Bragg. 2025. Understanding the LLM-ification of CHI: Unpacking the Impact of LLMs at CHI through a Systematic Literature Review. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA. doi:10.1145/3706598.3713726
- [24] Daniel Mesquita Feijó Rabelo, Ribamar de Souza Martins, Isaac Santos, Paulo Henrique G. Da Silva, Kiev Gama, and Windson Viana. 2025. Breaking Barriers in Mobile Accessibility: A Study of LLM-Generated Native Android Interfaces. In *2025 IEEE/ACM 12th International Conference on Mobile Software Engineering and Systems (MOBILESoft)*. 20–31. doi:10.1109/MOBIESoft66462.2025.00010
- [25] Mohaimenul Azam Khan Raiaan, Md. Saddam Hossain Mukta, Kaniz Fatema, Nur Mohammad Fahad, Sadman Sakib, Most. Marufatul Jannat Mim, Jubair Ahmad, Mohammed Eunus Ali, and Sami Azam. 2024. A Review on Large Language Models: Architectures, Applications, Taxonomies, Open Issues and Challenges. *IEEE Access* 12 (2024), 26839–26874. doi:10.1109/ACCESS.2024.3365742
- [26] Serviços e Informações do Brasil. 2026. gov.br – Apps no Google Play. https://play.google.com/store/apps/details?id=br.gov.meugovbr&hl=pt_BR. Acesso em: 21 ago. 2026.
- [27] Cole Stryker. [n.d.]. O que é LLM (grandes modelos de linguagem)? <https://www.ibm.com/br-pt/think/topics/large-language-models>. Acedido em: 29 de maio de 2026.
- [28] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All You Need. In *Advances in Neural Information Processing Systems*, Vol. 30. <https://papers.nips.cc/paper/2017/hash/3f5ee243547dee91fb053c1c4a845aa-Abstract.html>
- [29] Carlos Ventura, Daniel Rabelo, Ticianne Darin, and Windson Viana. 2025. Accessibility Evaluation of LLM-Generated Android Code in React Native. In *Anais do XXIV Simpósio Brasileiro sobre Fatores Humanos em Sistemas Computacionais* (Belo Horizonte/MG). SBC, Porto Alegre, RS, Brasil, 1301–1322. doi:10.5753/ihc.2025.10833
- [30] Xingyu Wu, Sheng-Hao Wu, Jibin Wu, Liang Feng, and Kay Chen Tan. 2025. Evolutionary Computation in the Era of Large Language Model: Survey and Roadmap. *IEEE Transactions on Evolutionary Computation* 29, 2 (2025), 534–554. doi:10.1109/TEVC.2024.3506731
- [31] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223* (2023).