O que sabemos sobre testes em *chatbots*? Uma revisão sistemática da literatura

Gabriel Santos[3], **Williamson Silva**[2], **Pedro Henrique Dias Valle**[1,3]

[1]Universidade Federal de Juiz de Fora (UFJF) – Juiz de Fora, MG, Brasil

[2]Universidade Federal do Pampa (UNIPAMPA) – Alegrete, RS, Brasil

[3]Universidade Tecnológica Federal do Paraná (UTFPR) – Cornélio Procópio, PR, Brasil

Gabriel.Santos.BCC@gmail.com, williamsonsilva@unipampa.edu.br,

pedrohenrique.valle@ufjf.br

Abstract. *The increasing use of conversational agents (chatbots) raises complex design, implementation, and, especially, testing issues. We conducted a systematic literature review and snowballing approach to characterize which tools and methods support testing activities in this application domain. As a result, we evidenced several tools that could support testing activities in chatbots, and we realized there needed to be a consensus in the field. This work's main contribution is a characterization of state-of-the-art testing tools and methods that support the construction and validation of chatbots.*

Resumo. *O uso crescente de agentes conversacionais (chatbots) levanta questões complexas de design, implementação e, especialmente, testes. Conduzimos uma revisão sistemática da literatura e uma abordagem de snowballing para caracterizar quais ferramentas e métodos apoiam atividades de teste neste domínio de aplicação. Como resultado, evidenciamos diversas ferramentas que poderiam apoiar atividades de testes em chatbots, e percebemos que era necessário haver um consenso na área. A principal contribuição deste trabalho é a caracterização de ferramentas e métodos de teste de última geração que suportam a construção e validação de chatbots.*

1. Introdução

Agentes conversacionais são programas de computador que utilizam Inteligência Artificial (IA) e Processamento de Linguagem Natural (PLN) para interagir com usuários por meio de conversas simuladas [Souza 2022]. Nesse contexto, os *chatbots* são um tipo de agente conversacionais que podem ser receber *inputs* por meio de texto ou voz [Moraes and de Souza 2015, Nunes 2012]. Eles são capazes de compreender e responder as perguntas dos usuários dentro de um contexto relevante, simulando uma conversa com um ser humano [Velásquez 2023]. Dentre os principais domínios de aplicações de *chatbots*, destacam-se [Shawar and Atwell 2007]: educação, saúde, negócios e comércio eletrônico. Além do setor de atendimento ao cliente, eles podem também ser encontrados em aplicativos de mensagens, sites e até mesmo em dispositivos de assistência pessoal, como o *Amazon Echo* (Alexa), a Siri (Apple), a Cortana (Microsoft) e o Assistant (Google) [Guerreiro and Barros 2019].

Assim como nos sistemas de software tradicionais [Valle et al. 2020], a qualidade dos *chatbots* é algo essencial para seu sucesso. A falta de testes em *chatbots* pode resultar em respostas falsamente positivas, ou seja, situações em que o *chatbot* fornece informações incorretas ou inadequadas [Santos et al. 2020]. Essas respostas podem prejudicar a experiência do usuário e comprometer a confiabilidade da plataforma. Portanto, é crucial considerar testes abrangentes e regulares para garantir a precisão e eficácia das interações do *chatbot* [Santos et al. 2020]. Conduzir atividades de teste de software pode contribuir para avaliar a capacidade de um *chatbot* atender às necessidades dos usuário, analisando se eles conseguem compreender e responder adequadamente às solicitações dos usuários, bem como verificar possíveis falhas e limitações em seu funcionamento [Santos et al. 2020].

Neste sentido, este estudo descreve uma Revisão Sistemática da Literatura (RSL) seguindo as diretrizes propostas por Kitchenham *et al.* (2010) com o objetivo de caracterizar as principais estrategias e abordagens de testes consideradas em *chatbots*. Como resultado, apresentam-se diferentes ferramentas e estratégias que podem apoiar atividades de testes em *chatbots*. Contudo, percebe-se que ainda não se tem um consenso sobre quais as melhores práticas na área, ou seja, a área de testes em *chatbots* ainda está se consolidando. A principal contribuição deste trabalho é a caracterização de ferramentas e métodos de teste de última geração que suportam a construção e validação de *chatbots*.

2. Método de Pesquisa

Conduziu-se uma Revisão Sistemática da Literatura (RSL) seguindo as diretrizes propostas por Kitchenham *et al.* (2010). Este estudo foi realizado em três etapas: planejamento, execução e análise dos resultados. As etapas de planejamento e execução são descritas nas subseções a seguir, e a etapa de análise dos resultados é apresentada na Seção 3.

2.1. Planejamento

Foram definidas as seguintes Questões de Pesquisa (QP) para serem respondidas a partir dos estudos selecionados (Tabela 1).

Tabela 1. Questões de Pesquisa (QP)

QP	Descrição
QP_1	Quais abordagens (técnicas/critérios) têm sido utilizadas para apoiar a atividade de teste em *chatbots*?
QP_2	Como os autores automatizaram as atividades para apoiar os testes?
QP_3	Qual o nível de teste tem sido aplicado nos teste em *chatbots*?
QP_4	Quais linguagens de programação estão sendo utilizadas para apoiar os testes?
QP_5	Foi realizada uma avaliação experimental? Se sim, como foi feito?
QP_6	Quais foram os domínios analisados nos trabalhos?
QP_7	Qual a forma de interação do *chatbot* avaliado?

Para identificar os estudos analisados, definiu-se a *string de busca:* ((“testing” OR “validation” OR “verification”) AND (“software” OR “metamorphic”)) and (“chatbot” OR “chatbot testing”)). A *string* de busca foi utilizada para identificar estudos primários nas seguintes bases de dados: ACM Digital Library, Science Direct, Scopus e IEEE Xplore. A busca foi conduzida considerando o período temporal de oito anos (2015-2023) com o objetivo de assegurar que os estudos sejam pertinentes e representativos, alinhando-se assim ao panorama atual do estado da arte desta investigação. Considerou-se os últimos oito anos pois, este período representa o inicio de pesquisas relevantes sobre

o tema considerado. Foram também definidos os Critérios de Inclusão (CI) e Critérios de Exclusão (CE) para selecionar o os estudos que foram analisados, conforme descritos na Tabela 2

Tabela 2. Critérios de seleção (inclusão e exclusão) da RSL.

ID	Descrição
CI_1	Estudos disponíveis para a leitura.
CI_2	Estudos escritos em língua inglesa e portuguesa.
CI_3	Estudos que discutam teste de software em agentes conversacionais baseados em texto (*chatbot*).
CI_4	Estudos que discutam métodos, técnicas e abordagens para testar *chatbots*.
CE_1	Estudos que são livros, teses, dissertações, patentes, livros.
CE_2	Estudos curtos (2 páginas) ou incompletos, tutoriais, propostas de *workshops* ou pôsteres.

2.2. Execução

Inicialmente, na fase de busca, a *string* de busca foi aplicada nas bases descritas na subseção 2.1, resultando em 80 estudos. Em seguida, na fase de pré-seleção, procedeu-se à na leitura dos títulos e resumos dos estudos (Fase 1), resultando em 22 trabalhos selecionados. Posteriormente, na fase de seleção (Fase 2), foram lidos os textos de introdução e conclusão dos estudos pré-selecionados, resultando na escolha de 11 estudos para análise mais detalhada. Por fim, realizou-se uma leitura completa e minuciosa dos cinco estudos finais (Fase 3), os quais foram selecionados após uma análise criteriosa. Vale ressaltar que foram adotados os critérios da Tabela 2 para seleção dos estudos em cada uma das fases. Devido à limitação na quantidade de estudos primários, decidiu-se complementar a RSL com a técnica de *snowballing backward* e o *snowballing forward* [Petersen et al. 2015, Kitchenham et al. 2010].

Utilizando o *snowballing backward*, foram considerados um total de 145 estudos relacionados ao tópico de pesquisa. Na Fase 1, foram selecionados 31 estudos; na Fase 2, 16 estudos foram selecionados; e na Fase 3, sete estudos foram selecionados.Por outro lado, ao considerar o *snowballing forward* foram recuperados 16 estudos. Na Fase 1, foram selecionados cinco estudos; na Fase 2, três estudos foram selecionados; e na Fase 3, apenas um estudo foi escolhido.

3. Resultados

Como resultado, foram selecionados 13 estudos (ver Tabela 3). Ao observar a distribuição temporal, nota-se que o ano de 2019 representou o período em que a temática recebeu maior atenção. Nos anos seguintes, houve uma queda, não implica na perda de importância, mas sim em um ajuste de atenção à medida que o campo amadurece.

A Tabela 4 sumariza as respostas das QPs, destacando as abordagens utilizadas (QP_1), o nível de teste (QP_3), a linguagem de programação de apoio a condução dos testes (QP_4), a validação da abordagem (QP_5), o domínio ao qual cada estudo está associado (QP_6) e a forma de interação do *chatbot* (QP_7) .

3.1. Q_1: Abordagens (técnicas/critérios) utilizadas em teste de *chatbots*

Diversas abordagens têm sido exploradas para apoiar a atividade de teste em *chatbots*, incluindo técnicas e critérios específicos. Os principais achados indicam a utilização de abordagens como testes funcionais (E_1, E_2, E_3, E_4, E_6, E_9, E_{11} e E_{13}), unitários (E_4), metamórficos (E_8, E_{12} e E_{13}) e de integração (E_1, E_2, E_3, E_7, E_9 e E_{11}). Por exemplo, no

Tabela 3. Estudos primários selecionados

ID	Ano	Título do trabalho	Snowballing	Referência
E_1	2020	Algorithm Inspection for Chatbot Performance Evaluation	Não	[Vijayaraghavan et al. 2020]
E_2	2018	BoTest: A framework to test the quality of conversational agents using divergent input examples	Sim	[Ruane et al. 2018]
E_3	2017	Bottester: testing conversational systems with simulated users	Sim	[Vasconcelos et al. 2017]
E_4	2019	Chatbot and bullyfree chat	Sim	[Selvi et al. 2019]
E_5	2023	Chatbot Interaction with Artificial Intelligence: human data augmentation with T5 and language transformer ensemble for text classification	Sim	[Bird et al. 2023]
E_6	2019	Chatbot testing using AI planning	Sim	[Bozic et al. 2019]
E_7	2020	OggyBug: A Test Automation Tool in Chatbots	Não	[Santos et al. 2020]
E_8	2022	Ontology-based metamorphic testing for chatbots	Não	[Božić 2022]
E_9	2022	Sorry, i don't Understand: Improving Voice User Interface Testing	Não	[Guglielmi et al. 2022]
E_{10}	2019	Sustainable Test Path Generation for Chatbots using Customized Response	Não	[Padmanabhan 2019]
E_{11}	2021	Testing challenges for NLP-intensive bots	Sim	[Cabot et al. 2021]
E_{12}	2019	Testing Chatbots Using Metamorphic Relations	Sim	[Bozic and Wotawa 2019]
E_{13}	2020	Testing chatbots with Charm	Sim	[Bravo-Santos et al. 2020]

Tabela 4. Resumo das Respostas

ID	Ferramenta	Nível/Técnica	Linguagem	Validação	Domínio
E_1	-	-	-	Não	Não definido
E_2	BoTest	-	-	Sim	Entretenimento
E_3	Bottester	-	-	Sim	Finanças
E_4	Anaconda	-	Python	Não	Bem-Estar Social
E_5	Modelo de parafraseamento T5	-	Java	Sim	-
E_6	Algoritmos de planejamento em conjunto com a Planning Domain Definition Language	-	-	Sim	Hotelaria
E_7	Testes automatizados por meio de APIs, através de uma interface web.	Integração	JavaScript	Sim	Advocacia
E_8	-	Unidade	-	Sim	Hotelaria
E_9	-	-	Python	Não	Não definido
E_{10}	-	Testes funcionais	-	Sim	Educação
E_{11}	Botium, Zypnos, Chatbottest, QBox, DashBot e Botest	Unidade	-	Não	Não definido
E_{12}	-	Metamórficos	-	Sim	Hotelaria
E_{13}	Charm	-	-	Sim	Hotelaria, Alimentação

estudo E_2 foram aplicados testes funcionais e de integração para avaliar o *chatbot* ChitChatBot. Os testes funcionais verificaram se o *chatbot* executava as funções esperadas e se respondia corretamente às entradas do usuário. Os autores relatam que o ChitChatBot foi implementado utilizando o *Language Understanding Intelligent Service* (LUIS), exigindo testes de integração para verificar a a correta integração com o LUIS e a adequada interpretação e resposas às entradas dos usuários.

3.2. Q_2: Automatização das atividades de testes em *chatbots*

No estudo E_2, foi empregada a ferramenta **BoTest**, um sistema modular capaz de integrar diversas técnicas, como detecção de erros e estilo linguístico. A detecção de erros abrange divergências sintáticas, morfológicas e semânticas, implementada por meio da introdução controlada de variações na entrada de teste. Exemplos específicos incluem erros de preposição não nativa e expressões coloquiais nativas, desafiando a capacidade do agente em lidar com diferentes formas de entrada. Quanto ao estilo linguístico, a avaliação envolve a introdução de elementos que refletem nuances de estilo na linguagem, como formalidade

ou informalidade. Dentro do contexto do **BoTest**, isso foi realizado por meio da introdução de divergências no estilo coloquial nativo, incorporando expressões informais ou regionais. Essas estratégias visam testar a capacidade do agente em compreender e reagir de maneira adequada diante de variações no estilo linguístico.

No estudo E_3 foi utilizada a ferramenta Bottester para simular interações de usuários com o *chatbot* e coletar métricas relacionadas às conversas. As métricas coletadas foram: Tamanho médio da resposta - avaliando a concisão das respostas, medindo o tamanho em caracteres e palavras de cada resposta do *chatbot*; Frequência de respostas - observando a frequência com que determinadas respostas são apresentadas, indicando possíveis limitações na base de conhecimento do *chatbot*; Frequência de palavras - analisando a frequência de palavras nas respostas, destacando padrões de linguagem e possíveis limitações no vocabulário do *chatbot*; Número de respostas corretas/incorretas - avaliando a capacidade do *chatbot* em fornecer respostas corretas, comparando-as com as respostas esperadas definidas nos casos de teste; e Tempo médio de resposta - mensurando o intervalo de tempo entre a submissão da pergunta e a chegada da resposta, sendo uma métrica-chave relacionada à percepção de qualidade do serviço pelo usuário.

No estudo E_4, foi utilizada a distribuição Anaconda de Python para gerenciar o ambiente virtual e as dependências do projeto. O ambiente virtual fornecido pela Anaconda foi essencial durante os testes, garantindo um isolamento eficaz das dependências específicas. Além disso, o uso do Jupyter Notebook, integrado à distribuição Anaconda, foi fundamental para a execução interativa de trechos de código, facilitando a análise e verificação incremental do desempenho do chatbot e do algoritmo de detecção de *cyberbullying*. A combinação dessas ferramentas proporcionou um ambiente coeso e eficiente para os testes, contribuindo para a estabilidade, reprodutibilidade e gerenciamento simplificado das dependências do projeto.

No estudo E_5, foi utilizado o modelo T5 (Text-To-Text Transfer Transformer) para expandir o conjunto de treinamento com dados adicionais. O T5 é um modelo de linguagem que possui a capacidade única de abordar várias tarefas de Processamento de Linguagem Natural (PLN) como problemas de conversão de texto para texto. Essa capacidade possibilita a geração de paráfrases, essenciais para diversificar e enriquecer o conjunto de dados de treinamento. Além do T5, foram empregadas várias outras ferramentas para avaliar e aprimorar o desempenho do modelo, tais como BERT, DistilBERT, RoBERTa e XLM-RoBERTa, além de estratégias como *Logistic Regression* e *Random Forests*. Essas ferramentas foram cruciais para analisar a eficácia do modelo proposto em interações de chatbot, classificação de texto e reconhecimento de sentimentos.

No estudo E_6, foram utilizados algoritmos de planejamento com a *Planning Domain Definition Language* (PDDL), uma linguagem usada em Inteligência Artificial para descrever domínios e problemas de planejamento. A PDDL permitiu modelar ações, parâmetros e intenções do usuário para os testes funcionais em chatbots. A implementação da abordagem de teste utilizou a linguagem Java com um framework de execução de testes. A ferramenta jsoup foi utilizada para analisar as respostas do chatbot durante os testes. Os algoritmos de planejamento, especialmente o *Fast Downward Planning System*, foram essenciais para criar sequências de ações para os cenários de teste. A PDDL desempenhou um papel crucial na definição das condições iniciais, ações possíveis e condições de objetivo. Essas ferramentas possibilitaram a execução automatizada de casos de teste

abstratos, avaliando o comportamento do chatbot em várias situações e garantindo uma abordagem eficiente nos testes funcionais.

No estudo E_7, foram realizados testes automatizados utiliando APIs[1], por meio de uma interface *Web*. Os testes incluíram: Testes de Reconhecimento de Padrões, que visam evoluir a base de conhecimento dos *chatbots* e identificar conflitos entre os módulos; Testes de Variáveis de Contexto, para verificar o contexto de diálogos realizados pelos *chatbots*; e Testes de Integração, para lidar com a troca de dados com outros sistemas. Esses testes são particularmente relevantes quando os *chatbots* precisam interagir com serviços externos.

No estudo E_{11}, foram mencionadas várias ferramentas desenvolvidas pelos autores, incluindo Botium, Zypnos, Chatbottest, QBox, DashBot, Botest, entre outras. Já no estudo E_{13}, a ferramenta CHARM foi utilizada, com Botium como backend para a execução dos testes automatizados. Nos demais estudos, não foram encontradas informações sobre ferramentas específicas utilizadas para suportar a atividade de teste em *chatbots*.

3.3. Q_3: Nível de teste em *chatbots*

No estudo E_1 foram conduzidos testes de integração, enquanto nos estudos E_3 e E_{11} foram realizados testes de unidade. O estudo E_4 abordou vários níveis de teste nos *chatbots*, incluindo testes unitários para avaliar partes específicas do código, como funções de processamento de texto, algoritmos de aprendizado de máquina e detecção de *cyberbullying*. Os testes de integração verificaram a interação entre os diferentes componentes do *chatbot*, como o processamento de linguagem natural e a lógica de resposta. Os testes de aceitação foram realizados para garantir que o *chatbot* atendesse às expectativas do usuário e cumprisse os requisitos funcionais e não funcionais estabelecidos, incluindo a usabilidade, eficiência e capacidade de resposta. Por sua vez, no estudo E_{12}, foram conduzidos teste de sistema. Em relação aos demais estudos, não foram identificadas evidências sobre os níveis de teste considerados para o contexto de *chatbot*.

3.4. Q_4: Linguagens de programação utilizadas nos testes

No estudo E_7, a linguagem de programação adotada foi JavaScript. Nos estudos E_9 e E_4, a linguagem escolhida foi Python. Já no estudo E_5, a linguagem de programação utilizada foi Java. Nos demais estudos, não foi adotada uma linguagem específica, uma vez que se basearam em ferramentas já existentes.

3.5. Q_5: Avaliação experimental das abordagens

No estudo E_7, a abordagem proposta foi validada por meio da realização de testes com duas empresas distintas, com o objetivo de verificar sua aplicabilidade em domínios e em plataformas de criação/gerenciamento de *bots* diferentes. Já o estudo E_8, a abordagem foi avaliada em um Sistema de Teste (SUT) que processa entradas em linguagem natural, utilizando um caso de teste que representava o cenário de teste.

No estudo E_{10}, a validação da abordagem foi realizada por meio da execução de testes funcionais. Por sua vez, no estudo E_3, embora não tenha sido explicitamente mencionada a validação da abordagem usando o Bottester, foi conduzido um "teste de

[1] API: *Application Programming Interface*

sanidade"para o CognIA. Esse teste teve como objetivo inicial verificar a integridade e funcionalidade básica do sistema, com foco na correta resposta a todas as perguntas implementadas. Durante essa avaliação, foram contabilizadas as respostas corretas e incorretas, além da análise do tempo de resposta para identificar os tipos de perguntas com maior demanda de tempo. Essa etapa inicial de testes foi realizada para garantir o funcionamento adequado do sistema antes de prosseguir com testes mais abrangentes.

No estudo E_2, a validação do framework envolveu o ChitChatBot, um agente de conversação desenvolvido pelos autores para discussões informais. O objetivo do *framework* era testar a qualidade das conversas dos agentes usando diferentes tipos de entradas. Foram criados exemplos divergentes para cada uma das 48 sentenças corretamente classificadas, e o desempenho do ChitChatBot nesses exemplos foi avaliado em comparação com suas respostas às entradas originais. No estudo E_5, a validação foi realizada por meio de experimentos com sete modelos de classificação de texto baseados em transformadores: BERT, DistilBERT, DistilRoBERTa, RoBERTa, XLM-RoBERTa, Logistic Regression e Random Forest. Esses modelos foram treinados e avaliados em um conjunto de dados de validação.

No estudo E_6, a abordagem foi validada por meio de experimentos com um *chatbot* de reserva de hotel, usando diferentes conjuntos de teste e especificações PDDL. Os testes envolveram a execução dos casos gerados pela abordagem, abrangendo vários cenários e condições. Durante a execução, a interação foi monitorada para avaliar o comportamento do sistema. Os conjuntos de teste foram criados com três abordagens: 1) uma única informação por solicitação, com ações repetitivas e valores constantes; 2) várias informações por solicitação, incluindo solicitações subsequentes com valores válidos e não intencionais; 3) mensagens de cancelamento adicionadas às especificações anteriores. A validação incluiu a comparação das respostas do *chatbot* com os resultados esperados. Os testes foram repetidos com diferentes conjuntos e especificações PDDL para garantir a robustez da abordagem.

No estudo E_{12}, a avaliação ocorreu por meio de um *chatbot* de turismo por meio de casos de teste gerados pela técnica de teste metamórfico (MT) e *metamorphic relations* (MRs). Utilizando o algoritmo **BotMorph**, foram executados casos de teste de origem (Is) e seus resultados (Os) foram registrados. Em seguida, MRs foram aplicadas para gerar casos de teste de acompanhamento (If) e testados no *chatbot*. Comparando as saídas resultantes (Of) com os casos de origem, a validação avaliou a consistência do sistema em diversas situações. Os resultados destacaram a eficácia da abordagem em detectar comportamentos inesperados, revelando tanto o desempenho satisfatório do *chatbot* em certas condições quanto suas falhas em outras.

No estudo E_{13}, a abordagem foi validada por meio de um experimento para responder questões de pesquisa sobre detecção de problemas nos *chatbots* e melhoria da qualidade. Os testes incluíram esturdiez, coesão e precisão, com diversos operadores de mutação. Cada chatbot, incluindo Baseline, Nutrition e RoomService, foi avaliado. Os testes de esturdiez aplicaram mutações em caracteres e números, revelando falhas em todos os chatbots. Uma abordagem de correspondência difusa foi empregada para melhorar os resultados. Na coesão, sem mutações, todos os chatbots passaram. Os testes de precisão envolveram mutações em palavras e idiomas, com falhas em alguns testes para todos os chatbots. Após treinamento adicional com casos de falha, os chatbots foram retestados,

resultando em melhorias significativas na precisão.

3.6. Q_6: Domínios dos *chatbots*

Os domínios dos chatbots investigados nos estudos foram variados, abrangendo uma variedade de setores. No âmbito da pesquisa, foram analisados os seguintes domínios:

- **Hotelaria** (E_8, E_6, E_{12} e E_{13}), indicando um interesse em integrar *chatbots* nesse setor, possivelmente para melhorar serviços e interações com clientes.
- **Educação** (E_{10}), trouxe uma reflexão sobre como os *chatbots* podem ser implementados ou otimizados para melhorar processos educacionais.
- **Advocacia** (E_7), indicando uma pesquisa sobre como *chatbots* podem ser aplicados no cenário jurídico.
- **Finanças** (E_3), refletiram o interesse em compreender a viabilidade e benefícios da utilização de chatbots nesse setor específico.
- **Entretenimento** (E_2), sugerindo uma análise sobre o papel dos **chatbots** na indústria do entretenimento.
- **Bem-Estar Social** (E_4), indicando uma análise sobre como os chatbots podem contribuir para promover o bem-estar em diversas comunidades.
- **Alimentação** (E_{13}), trouxe *insights* sobre a aplicação potencial dos *chatbots* nesse ramo específico.

Por fim, os estudos E_9, E_{11} e E_1 não se limitaram a um domínio específico, sugerindo uma abordagem mais genérica na aplicação dos *chatbots*. Essa diversidade de domínios reflete a ampla gama de aplicações possíveis para os *chatbots* em diferentes setores, evidenciando a versatilidade dessa tecnologia em várias áreas de estudo e prática.

3.6.1. Q_7: Forma de interação do *chatbot* avaliado

A interação dos *chatbots* avaliados varia principalmente entre texto e voz, sendo que alguns sistemas permitem ambas as formas de interação. Os estudos E_3, E_4, E_5, E_6, E_7, E_8 e E_{12} utilizam o texto como forma de interação, enquanto o estudo E_9 utiliza voz. No entanto, nos estudos E_2, E_{10}, E_{11} e E_{13} não foi explicitada no texto a forma de interação. No estudo E_2, por exemplo, os autores mencionam o uso do LUIS da Microsoft Azure Cognitive Services para implementar o ChitChatBot, o que sugere que ocorre uma interação por meio de textos. O mesmo se aplica aos demais trabalhos mencionados. O estudo E_1 não especifica uma forma de interação específica do *chatbot* avaliado. No entanto, ele discute várias técnicas de teste e avaliação que podem ser aplicadas a diferentes tipos de *chatbots*, independentemente de sua forma de interação.

4. Discussão dos resultados

A análise dos resultados revela uma variedade de níveis de teste e ferramentas empregadas, uma prática compreensível devido à especificidade do domínio ou do *chatbot*. Essa variação reflete a natureza única de cada *chatbot*, sendo compreensível que nenhum seja totalmente idêntico ao outro em termos de requisitos e características. A diversidade nas abordagens de teste destaca a necessidade de flexibilidade e personalização para garantir a eficácia do processo de teste em consonância com as particularidades de cada aplicação.

A variedade de ferramentas identificadas para apoiar as atividades de teste reflete a falta de um consenso consolidado sobre a melhor abordagem. O uso de *APIs* em conjunto com interfaces web, simulação de interações do usuário e a coleta de métricas usando a ferramenta Bottester são apenas alguns exemplos das estratégias adotadas. A modularidade proporcionada pelo BoTest, permitindo a integração de diversas técnicas, ressalta a necessidade de flexibilidade no processo de teste. A utilização de diferentes linguagens de programação, como JavaScript, Java e Python, ressalta a adaptabilidade das abordagens de teste aos requisitos específicos de cada estudo. Essa diversidade também se reflete nos domínios analisados, que abrangem setores como Hotelaria, Educação, Advocacia, Finanças, Entretenimento, Bem-Estar Social, Alimentação, por exemplo.

A validação das abordagens adotadas nos estudos destaca a preocupação com a eficácia e a aplicabilidade prática das técnicas de teste. Desde testes funcionais até experimentos envolvendo modelos de classificação de texto, os pesquisadores buscaram garantir a confiabilidade e o desempenho dos *chatbots* em diferentes contextos. Contudo, é importante observar que, apesar dos avanços, ainda não existe um consenso sobre a "abordagem correta" para testar *chatbots*. A diversidade de ferramentas, linguagens e domínios indica a complexidade do cenário e destaca a necessidade contínua de pesquisa e desenvolvimento nesta área.

Em resumo, este estudo forneceu uma contribuição ao mapear o cenário atual de testes de *chatbots*, fornecendo *insights* valiosos para pesquisadores, desenvolvedores e profissionais envolvidos nesse campo dinâmico. O desafio agora é continuar avançando, explorando novas abordagens e estabelecendo diretrizes mais claras para o teste eficaz de *chatbots* em diferentes domínios e contextos de aplicação.

5. Limitações

Este estudo também apresenta algumas limitações que devem ser consideradas ao interpretar e generalizar os resultados. As limitações identificadas são discutidas de acordo com as diretrizes de Kitchenham *et al.* (2010), reconhecendo a importância de uma abordagem crítica na avaliação da pesquisa, sendo elas:

- **Amplitude da pesquisa.** A abrangência da pesquisa pode ser considerada como uma limitação. Embora tenhamos abordado uma variedade de estudos sobre testes de *chatbots*, é possível que algumas abordagens ou ferramentas emergentes não tenham sido incluídas. O campo de testes de *chatbots* é dinâmico conforme pode ser observado nos resultados obtidos, e novas metodologias podem surgir após a conclusão deste estudo.
- **Publicações em outros idiomas.** A limitação da pesquisa restrita a documentos em português e inglês pode ter influenciado a inclusão ou exclusão de alguns estudos relevantes em diferentes idiomas. A pesquisa em testes de *chatbots* é global, e valiosas contribuições podem ter sido publicadas em outros idiomas.
- **Disponibilidade de informações.** Algumas publicações podem não ter fornecido informações detalhadas o suficiente sobre as abordagens de teste adotadas. A falta de transparência em certos estudos pode afetar a compreensão completa das práticas de teste aplicadas.
- **Evolução tecnológica.** O campo de *chatbots* está em constante evolução, com avanços tecnológicos rápidos. As abordagens de teste identificadas neste estudo

podem se tornar obsoletas com o tempo, à medida que novas tecnologias e metodologias são desenvolvidas.

- **Viés de seleção.** A escolha de estudos incluídos pode ter introduzido um viés de seleção. Apesar dos esforços para selecionar uma amostra representativa, a subjetividade na escolha dos estudos pode influenciar os resultados e conclusões.
- **Variedade de abordagens.** A diversidade de abordagens de teste, embora seja uma característica interessante do campo, também pode representar um desafio na comparação e generalização dos resultados. A falta de padronização nas práticas de teste dificulta a identificação de melhores práticas universais.
- **Influência do contexto temporal.** Como a pesquisa foi realizada até o ano de 2023, é possível que novos desenvolvimentos tenham ocorrido na área de teste de *Chatbots* após essa data. A evolução rápida dessa tecnologia pode tornar algumas conclusões menos aplicáveis em um contexto mais recente.

Ao reconhecer essas limitações, é importante interpretar os resultados deste estudo com cautela e considerar futuras pesquisas que possam abordar essas lacunas e aprimorar a compreensão do teste de *Chatbots*.

6. Conclusões

A análise dos estudos considerado revelou uma diversidade nas abordagens e estratégias utilizadas no teste de *chatbots*. Foram aplicados testes em diferentes níveis, incluindo testes automatizados por meio de *API's*, testes de integração, testes de unidade, testes funcionais e testes metamórficos. Essa variedade reflete a natureza única de cada *chatbot*, adaptando-se aos requisitos específicos e características distintas de cada aplicação. A validação das abordagens foi uma preocupação nos estudos analisados. Os métodos de validação abrangeram desde testes com empresas distintas até experimentos com modelos de classificação de texto baseados em transformadores. Em alguns casos, foram realizados testes de sanidade para avaliar o desempenho geral do sistema de *chatbot*, enquanto em outros, a validação envolveu a criação de exemplos divergentes para avaliação de respostas em comparação com as entradas originais.

A diversidade de domínios investigados também se destacou, abrangendo setores como hotelaria, educação, assistência virtual, advocacia, finanças, entretenimento, bem-estar social e alimentação. Essa variedade de domínios evidencia a versatilidade dos *chatbots*, sugerindo sua aplicabilidade em diversos contextos e setores da economia. No que diz respeito às formas de interação, todos os estudos destacaram o uso de PLN e NLU. A interação por texto foi prevalente e um único *chatbot* disponibilizou a interação por voz, enquanto outros permitiram ambas as formas de interação. Quanto às abordagens técnicas e critérios para apoiar a atividade de teste em *chatbots*, evidenciou-se estudos que reportaram o uso de testes funcionais, unitários, metamórficos e de integração. No entanto, a área ainda está em um estágio inicial de pesquisa, e a falta de consenso consolidado destaca a predominância de testes exploratórios.

Os resultados deste trabalho contribuem para a compreensão das técnicas e ferramentas empregadas no teste de *chatbots*, preenchendo uma lacuna no estado da arte dessas aplicações. Essa contribuição oferece uma visão abrangente que pode facilitar a tomada de decisão ao escolher técnicas ou ferramentas, fornecendo um panorama mais claro e esclarecedor sobre o estado atual do teste de *chatbots*.

Referências

[Bird et al. 2023] Bird, J. J., Ekárt, A., and Faria, D. R. (2023). Chatbot interaction with artificial intelligence: human data augmentation with t5 and language transformer ensemble for text classification. *Journal of Ambient Intelligence and Humanized Computing*, 14(4):3129–3144.

[Bozic et al. 2019] Bozic, J., Tazl, O. A., and Wotawa, F. (2019). Chatbot testing using ai planning. In *2019 IEEE International Conference On Artificial Intelligence Testing (AITest)*, pages 37–44. IEEE.

[Bozic and Wotawa 2019] Bozic, J. and Wotawa, F. (2019). Testing chatbots using metamorphic relations. In *Testing Software and Systems: 31st IFIP WG 6.1 International Conference, ICTSS 2019, Paris, France, October 15–17, 2019, Proceedings 31*, pages 41–55. Springer.

[Božić 2022] Božić, J. (2022). Ontology-based metamorphic testing for chatbots. *Software Quality Journal*, 30:227–251.

[Bravo-Santos et al. 2020] Bravo-Santos, S., Guerra, E., and de Lara, J. (2020). Testing chatbots with charm. In *Quality of Information and Communications Technology: 13th International Conference, QUATIC 2020, Faro, Portugal, September 9–11, 2020, Proceedings 13*, pages 426–438. Springer.

[Cabot et al. 2021] Cabot, J., Burgueno, L., Clarisó, R., Daniel, G., Perianez-Pascual, J., and Rodriguez-Echeverria, R. (2021). Testing challenges for nlp-intensive bots. In *2021 IEEE/ACM Third International Workshop on Bots in Software Engineering (BotSE)*, pages 31–34. IEEE.

[Guerreiro and Barros 2019] Guerreiro, A. and Barros, D. M. V. (2019). Novos desafios da educação a distância: programação e uso de chatbots.

[Guglielmi et al. 2022] Guglielmi, E., Rosa, G., Scalabrino, S., Bavota, G., and Oliveto, R. (2022). Sorry, i don't understand: Improving voice user interface testing. Association for Computing Machinery.

[Kitchenham et al. 2010] Kitchenham, B., Pretorius, R., Budgen, D., Brereton, O. P., Turner, M., Niazi, M., and Linkman, S. (2010). Systematic literature reviews in software engineering–a tertiary study. *Information and software technology*, 52(8):792–805.

[Moraes and de Souza 2015] Moraes, S. M. and de Souza, L. S. (2015). Uma abordagem semiautomática para expansão e enriquecimento linguístico de bases aiml para chatbots. In *Congresso Internacional de Informática Educativa*, volume 20, pages 600–605.

[Nunes 2012] Nunes, F. O. (2012). Chatbots e mimetismo: uma conversa entre humanos, robôs e artistas. In *Proceedings of 6th International Conference on Digital Arts—ARTECH*, pages 89–96.

[Padmanabhan 2019] Padmanabhan, M. (2019). Sustainable test path generation for chatbots using customized response. *International Journal of Engineering and Advanced Technology*, 8:149–155.

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

[Ruane et al. 2018] Ruane, E., Faure, T., Smith, R., Bean, D., Carson-Berndsen, J., and Ventresque, A. (2018). Botest: a framework to test the quality of conversational agents using divergent input examples. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces Companion*, pages 1–2.

[Santos et al. 2020] Santos, M. B. D., Furtado, A. P. C., Nogueira, S. C., and Moreira, D. D. (2020). Oggybug: A test automation tool in chatbots. pages 79–87. Association for Computing Machinery.

[Selvi et al. 2019] Selvi, V., Saranya, S., Chidida, K., and Abarna, R. (2019). Chatbot and bullyfree chat. In *2019 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN)*, pages 1–5. IEEE.

[Shawar and Atwell 2007] Shawar, B. A. and Atwell, E. (2007). Chatbots: are they really useful? *Journal for Language Technology and Computational Linguistics*, 22(1):29–49.

[Souza 2022] Souza, P. H. C. (2022). Proposta de implementação de chatbot para o observatório do instituto do mar.

[Valle et al. 2020] Valle, P. H. D., Vilela, R. F., and Hernandes, E. C. M. (2020). Does gamification improve the training of software testers? a preliminary study from the industry perspective. In *Proceedings of the XIX Brazilian Symposium on Software Quality*, pages 1–10.

[Vasconcelos et al. 2017] Vasconcelos, M., Candello, H., Pinhanez, C., and dos Santos, T. (2017). Bottester: testing conversational systems with simulated users. In *Proceedings of the XVI Brazilian Symposium on Human Factors in Computing Systems*, pages 1–4.

[Velásquez 2023] Velásquez, F. R. (2023). O chatgpt na pesquisa em humanidades digitais: Oportunidades, críticas e desafios. *TEKOA*, 2(2).

[Vijayaraghavan et al. 2020] Vijayaraghavan, V., Cooper, J. B., and Leevinson, R. L. R. (2020). Algorithm inspection for chatbot performance evaluation. volume 171, pages 2267–2274. Elsevier B.V.