UNIVERSIDADE DE PERNAMBUCO

ESCOLA POLITÉCNICA DE PERNAMBUCO

Universidade de Pernambuco

Programa de Pós-Graduação em Engenharia da Computação

(PPGEC)

Proposta de Dissertação de Mestrado

Áreas: Engenharia de Software

Título: Critérios de Benchmark Multidimensionais de Teste de Software para Avaliação de Sistemas Conversacionais Baseados em LLMs

Orientador: Wylliams Barbosa Santos (wbs@upe.br)
Coorientador: Cleyton Mário de Oliveira Rodrigues (cleyton.rodrigues@upe.br)

Teste de software, uma atividade fundamental para a garantia da qualidade, e notoriamente conhecido por ser um processo custoso e intensivo em trabalho manual. Em particular, as tarefas de desenho e especificação de casos de teste são frequentemente citadas como repetitivas, demoradas e propensas a erro humano, podendo se tornar um gargalo que limita a agilidade das equipes de desenvolvimento [Pressman et al., 2016; Sommerville, 2019].

Sistemas baseados em inteligência artificial, especialmente chatbots baseados em Grande Modelo de Linguagem (Large Language Model, LLMs), desafiam pressupostos tradicionais de teste de software, exigindo a reavaliação de abordagens sob diferentes dimensões de teste. Adamopoulou e Moussiades [2020] destacam que a natureza não determinística desses sistemas implica que uma mesma entrada pode gerar respostas distintas, dificultando a definição de um oráculo de teste fixo. Além disso, o espaço de entradas em linguagem natural é praticamente ilimitado, o que inviabiliza a enumeração exaustiva de casos de teste e demanda o uso de amostragem, métricas estatísticas e benchmarks especializados para avaliação [Ammann et al., 2016; Kasneci et al., 2023].

Avaliar a concordância e a coerência das respostas do chatbot não é uma tarefa trivial, pois as interações são dinâmicas e dependem fortemente do contexto e da interação anterior do usuário. Por essa razão, desenvolver técnicas e métricas que permitam avaliar de forma automática e contínua é crucial para garantir a experiência do usuário (UX) e a confiança no sistema [Barbosa et al., 2022].

A literatura apresenta diversas técnicas e benchmarks utilizados para validação de testes de software, incluindo a identificação de *test smells* em processos de avaliação e garantia de qualidade. No contexto de testes aplicados a sistemas baseados em LLMs, são propostas abordagens específicas voltadas a testes funcionais, de segurança e de robustez, além de métricas relacionadas à evolução dos modelos e avaliações de aspectos éticos e de adequação das respostas. Embora algumas dessas abordagens sejam derivadas de práticas tradicionais de teste de software, sua aplicação em sistemas baseados em inteligência artificial exige adaptações, em função da natureza probabilística e do espaço aberto de entradas desses modelos. Nesse contexto, torna-se relevante a distinção entre abordagens de caixa-preta e caixa-branca, especialmente ao considerar modelos fechados, como *ChatGPT* e *Gemini*, e modelos abertos, como *Ollama* e *DeepSeek*.

Adicionalmente, abordagens como *LLM-as-a-Judge* têm sido utilizadas para automatizar parte do processo de avaliação, permitindo a análise qualitativa de respostas geradas por modelos. Um dos principais fatores para o avanço dessas práticas é a construção de benchmarks padronizados, utilizados para avaliar e comparar o desempenho de chatbots e LLMs de forma sistemática [Cuadrado et al., 2021; Salim et al., 202].

O objetivo desta proposta é desenvolver e validar um framework de qualidade de software voltado a sistemas baseados em *Large Language Models (LLMs)*, considerando técnicas, métricas e benchmarks específicos para avaliação desses sistemas.

Referências Bibliográficas

[1] **PRESSMAN, Roger S.; MAXIM, Bruce R.** *Software Engineering: A Practitioner's Approach*. 9. ed. New York: McGraw-Hill Education, 2019.

[2] **SOMMERVILLE, Ian.** *Software Engineering*. 10. ed. Boston: Pearson, 2016.

[3] **ADAMOPOULOU, E.; MOUSSIADES, L.** An overview of chatbot technology. In: *IFIP International Conference on Artificial Intelligence Applications and Innovations*. Cham: Springer, 2020. p. 373-383.

[4] **AMMANN, Paul; OFFUTT, Jeff**. *Introduction to Software Testing*. 2. ed. Cambridge: Cambridge University Press, 2016.

[5] **KASNECI, Enkelejda et al.** *ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education*. *Learning and Individual Differences*, v. 103, p. 102274, 2023. DOI: 10.1016/j.lindif.2023.102274.

[6] **BARBOSA, A.; SANTOS, W.; OLIVEIRA, M.** Avaliação de chatbots: métricas e métodos automáticos. *Revista Brasileira de Computação Aplicada*, v. 14, n. 2, p. 45-60, 2022.

[7] **CUADRADO, Jesus et al.** *Integrating Static Quality Assurance in CI Chatbot Development Workflows*. In: Proceedings of the IEEE/ACM International Conference on Automated Software Engineering Workshops, 2021.

[8] **SALIM, R. A.; et al.,** *A Benchmarking Survey: Evaluating the Accuracy and Effectiveness of Benchmark Models in Measuring the Performance of Large Language Models*. In: 6th International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS), 2024, p. 27–32. DOI: 10.1109/icimcis63449.2024.10957638