


Toward Trustworthy AI-Enabled Internet Search^{*}

Miguel Romero-Arjona¹, Sergio Segura¹, and Aitor Arrieta²

¹ Universidad de Sevilla, Seville, Spain
{mromero5,sergiosegura}@us.es

² Mondragon University, Gipuzkoa, Spain
aarrieta@mondragon.edu

Abstract. Artificial Intelligence (AI), particularly Large Language Models (LLMs), are called to reshape how we search information on the Internet. AI regulation initiatives (e.g., EU AI Act) are rapidly emerging to ensure that AI products are trustworthy and safe. However, ensuring the regulatory compliance of an AI system currently relies on manual checklists, making the process tedious, time-consuming, and hardly scalable. In this work-in-progress paper, we outline our vision for developing a tool ecosystem aimed at automatically testing AI-driven search engines in accordance with EU trustworthiness compliance requirements. Specifically, we present some of the key quality characteristics to target, a brief summary of the state-of-the-art LLM testing, and the key features of the envisioned tool ecosystem.

Keywords: Artificial intelligence · Search engines · Quality attributes

1 Introduction

Large Language models (LLMs) are emerging innovation products in the field of generative Artificial Intelligence (AI), standing out for their ability to generate novel and imaginative results. The irruption of tools such as ChatGPT, currently integrated into the Bing search engine, holds the promise of changing the way people explore information online and make decisions. Unlike traditional search engines, which provide results based on keyword matches and pre-defined structures, these *AI-enabled Search Engines* (AISEs) [2] incorporate more advanced contextual understanding and language generation. This means that they not

^{*} This work has been partially supported by grants PID2021-126227NB-C22 and TED2021-131023B-C21, funded by MCIN/AEI/10.13039/501100011033 and by European Union “NextGenerationEU”/PRTR. Views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union or European Commission. Neither the European Union nor the granting authority can be held responsible for them. Funded within the framework of the NGI Search project under grant agreement No 101069364. Aitor Arrieta is part of the Systems and Software Engineering group of Mondragon University (IT1519-22), supported by the Department of Education, Universities and Research of the Basque Country.



only respond to specific queries but also have the ability to contextualize and generate more coherent answers tailored to the user’s context.

Despite the remarkable benefits that these systems offer in various respects, they also raise critical concerns about security and trust. Regulatory initiatives, such as the Artificial Intelligence Act being developed by the European Union (EU) [1], are expected to become essential, forcing companies providing AI products within the EU to ensure a certain level of trustworthiness. However, checking an AI system regulatory compliance is currently a predominantly manual process, and thus tedious, time-consuming, and unreliable.

In response to this problem, we envision a novel tool ecosystem for assessing the trustworthiness of AISEs and testing their compliance with EU regulations. Specifically, we aim to automate the generation of test cases and test oracles using techniques such as metamorphic testing (MT) [7]. Our tool ecosystem will be applicable both before and, more importantly, during operation. It will alert developers when the model deviates from its expected behaviour, anticipating trust-related problems, and mitigating their impact on end-users. In this work-in-progress paper, we take a first step by identifying the key target quality attributes to be tested and summarizing current testing approaches for LLMs. Finally, we outline our proposal for automating the detection of bugs in AISEs.

This work is part of the project “TrustAI: Trustable AI-Driven Internet Search”—recently funded within the EU NGI Search framework [4]—aimed to transform the way we search and discover information and resources on the internet.

2 Target Quality Attributes

Ensuring the trustworthiness of AISEs requires taking into account different quality attributes. Based on the current version of the EU regulation (AI Act), and the objectives established by the NGI Search framework [3], we will prioritize detecting faults related to the following attributes:

- **Robustness:** AISEs should be reliable and accurate, e.g., search results should be based on reliable sources to avoid the dissemination of false or misleading information.
- **Privacy:** AISEs should respect the right of individuals to control their personal information and to decide who can access it, how, and for what purpose, e.g., search results should not contain personal or sensitive information about users, such as their phone number or home address.
- **Fairness:** AISEs should be fair and not discriminate against certain groups, e.g., search results should not differ for users of different demographics or religions.
- **Explainability:** AISEs should have the ability to explain and justify its processes, results, and decisions in a way that is understandable by users.

3 Automated Testing of Large Language Models

Most existing approaches for testing LLMs follow a black-box approach—this will be the strategy used in our work. Black-box approaches can be classified into three groups. The first group includes those techniques using specific datasets [8]. For example, a translator can be tested using a dataset containing English-Spanish sentence pairs that allow the translations generated by the model to be compared with the reference translations in the other language. This approach focuses on specific use cases, allowing us to detect in which contexts these systems could cause some issues.

The second group includes those approaches using metamorphic testing [5], where bugs are detected by comparing the inputs and outputs of two or more executions of the system under test. For example, we could ask a model like ChatGPT to generate a text, and then to generate a “shorter version”. Intuitively, the second response should be shorter than the first one. Otherwise, a potential bug would be revealed.

Finally, the third group includes those techniques using LLMs for assessing the responses of the LLM under test to identify issues. This is often referred to as LLM-as-a-Judge [9]. For example, we could use a model language to decide whether another model’s responses exhibit bias.

4 Approach Overview

We envision an open-source tool ecosystem designed to automate the detection of failures in AISEs, in alignment with current EU regulations. This ecosystem will operate both before and during system operation. Based on our previous work on automated testing of REST APIs [6], we plan to design a bot-driven architecture to ensure system scalability and maintainability. The figure 1 shows an initial outline of our proposal. We initially distinguish four types of bots: test data generators, test executors, test evaluators, and test reporters. The bots will communicate by exchanging messages using BOTICA, a bot communication platform developed in our previous project. *Test data generators* will be responsible for creating test datasets—using techniques such as metamorphic testing [5]—that will be used to assess the quality attributes of the model. *Test executors* will be designed to execute the generated test data on the AISEs. *Test evaluators* will be tasked with evaluating the results obtained from the execution of test data using techniques such as metamorphic testing and LLM-as-a-judge. *Test reporters* will generate reports and dashboards based on the evaluated results. Given the non-deterministic nature of AISEs, we expect different levels of accuracy for each bot and the need to deal with false positives and negatives. Therefore, human oversight will be crucial for which we plan to integrate strategies for gathering and processing human feedback.

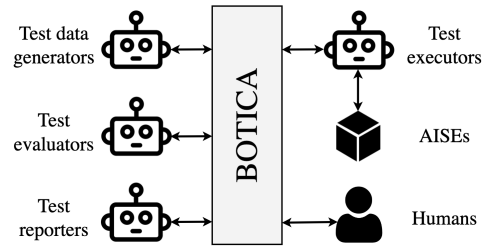


Fig. 1. Bot-driven architecture approach for testing on AISEs.

5 Conclusions

The benefits of AI come at a price which, in the case of AISEs, may reveal themselves in the form of biased, unfair or dangerous responses. Regulatory initiatives are positive, but difficult to check in practice. In this paper, we present our vision for an ecosystem of tools designed to automatically assess the quality attributes of AISEs, facilitating compliance monitoring with the EU AI regulations.

Acknowledgments We have used ChatGPT 3.5 during the preparation of this work to make slight edits and improve readability.

References

1. European Artificial Intelligence Act Regulation Proposal. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206>, accessed January 2024
2. How a reasoning engine works. <https://www.linkedin.com/learning/generative-ai-the-evolution-of-thoughtful-online-search/how-a-reasoning-engine-works>, accessed December 2023
3. NGI, for an Internet of Humans. <https://www.ngi.eu/ngi-for-an-internet-of-humans>, accessed December 2023
4. NGI Search. <https://www.ngi.eu/ngi-projects/ngi-search>, accessed December 2023
5. Hyun, S., Guo, M., Babar, M.A.: METAL: Metamorphic Testing Framework for Analyzing Large-Language Model Qualities (2023)
6. Martin-Lopez, A., Segura, S., Ruiz-Cortés, A.: Online testing of RESTful APIs: promises and challenges. In: Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering. p. 408–420. ACM, Singapore (2022)
7. Segura, S., Fraser, G., Sanchez, A.B., Ruiz-Cortés, A.: A Survey on Metamorphic Testing. *IEEE Transactions on Software Engineering* **42**(9), 805–824 (2016)
8. Wang, B., Xu, C., Wang, S., Gan, Z., Cheng, Y., Gao, J., Awadallah, A.H., Li, B.: Adversarial GLUE: A Multi-Task Benchmark for Robustness Evaluation of Language Models (2022)
9. Zheng, L., Chiang, W.L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E.P., Zhang, H., Gonzalez, J.E., Stoica, I.: Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena (2023)