Article

A Testing Framework for AI Linguistic Systems (*testFAILS*)

Yulia Kumar[1], *, Patricia Morreale[1], Peter Sorial[1], Justin Delgado[1], J. Jenny Li[1] and Patrick Martins[1]

[1] Department of Computer Science and Technology, Kean University Union, NJ, USA ykumar@kean.edu

Abstract: This paper presents an innovative testing framework, *testFAILS*, designed for the rigorous evaluation of AI Linguistic Systems, with a particular emphasis on various iterations of ChatGPT. Leveraging orthogonal array coverage, this framework provides a robust mechanism for assessing AI systems, addressing the critical question, "How should we evaluate AI?" While the Turing test has traditionally been the benchmark for AI evaluation, we argue that current publicly available chatbots, despite their rapid advancements, have yet to meet this standard. However, the pace of progress suggests that achieving Turing test-level performance may be imminent. In the interim, the need for effective AI evaluation and testing methodologies remains paramount. Our research, which is ongoing, has already validated several versions of ChatGPT, and we are currently conducting comprehensive testing on the latest models, including ChatGPT-4, Bard and Bing Bot, and the LLaMA model. The *testFAILS* framework is designed to be adaptable, ready to evaluate new bot versions as they are released. Additionally, we have tested available chatbot APIs and developed our own application, *AIDoctor*, utilizing the ChatGPT-4 model and Microsoft Azure AI technologies.

Keywords: keyword 1; chatbots 2; chatbot validation 3; bots 4; A Testing Framework for AI Linguistic Systems (*testFAILS*) 5; *AIDoctor*

1. Introduction

Our research focuses on developing *testFAILS*, a robust testing framework for evaluating and comparing leading chatbots, including OpenAI's ChatGPT-4 [1], Google's Bard [2], Meta's LLaMA[3], Microsoft's Bing Chat [4], and emerging contenders like Elon Musk's TruthGPT [5]. *TestFAILS* adopts an adversarial approach, highlighting chatbot shortcomings to counterbalance the frequent media hype around "AI breakthroughs." The Turing Test, a widely accepted measure of AI sophistication, is a key benchmark in our framework. Despite rapid advancements in AI, no chatbot has yet achieved this milestone, underscoring the need for effective evaluation methodologies.

TestFAILS comprises six critical components:

A. Simulated Turing Test Performance
B. User Productivity and Satisfaction
C. Integration into Computer Science Education
D. Multilingual Text Generation
E. Pair Programming Capabilities
F. Success in Bot-Based App Development

We have implemented a ternary evaluation system, assigning a pass, failure, or undetermined status based on the presence of counterexamples. Numerical indicators are established for comparison and aggregation of scores: 0 for failure, 0.5 for undetermined, and 1 for pass.

To fine-tune component weighting, we engaged the chatbots themselves, asking them to rate component importance. Table 1 displays their responses.

Table 1. The weights of the *testFAILS* components, proposed by the chatbots themselves.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 June 2023 doi:10.20944/preprints202306.0418.v1

Testing Components/ Parameters	Weights proposed by the chatbots		
	Chat-GPT3.5	*Chat-GPT4*	*Bard*
A Simulated Turing Test Performance	**0.2**	**0.2**	**0.2**
B User Productivity and Satisfaction	**0.2**	**0.2**	**0.2**
C Integration into Computer Science Education.	**0.2**	**0.3**	**0.1**
D Multilingual Text Generation.	**0.15**	**0.1**	**0.15**
E Pair Programming Capabilities.	**0.15**	**0.1**	**0.15**
F Bot-based App development and its success	**0.1**	**0.1**	**0.2**
Total Score	**1.00**	**1.00**	**1.00**

As Table 1 shows, all chatbots agreed on the first two components' values, while others varied. We conducted sub-studies on each component-related topic to determine the correct weights. Our goal is to provide a sharp tool for assessing AI linguistic systems' efficacy and sophistication.

2. Research Background

Our research in chatbot evaluation leverages our extensive background in Natural Language Processing (NLP) [6][7] and comprehensive understanding of AI models, particularly within the Python programming ecosystem [8][9]. Recently, we have shifted our focus towards Transformer Neural Networks, aiming to uncover and comprehend the biases embedded within their computational layers [10][11][12][13][14]. Our technical expertise encompasses a range of programming languages and frameworks, and we've successfully interfaced with several web services and APIs, including Google Translate and Yandex Translate [9][15]. Initially, our research questions were broad, contemplating whether a specific chatbot could boost societal intelligence or, conversely, lead to its decline. We also pondered the tangible impact of these AI tools on the quality of human life. However, given the inherent complexities in substantiating such wide-ranging claims, we have honed our focus. Our primary research objectives now concentrate on assessing the influence of chatbots on user experience and their potential to convincingly pass the Turing Test. This refined focus allows us to delve deeper into these pivotal areas, thereby contributing meaningful insights to the AI field.

3. Related work

Chatbots have become a staple in the industry, with several metrics, frameworks, and tools achieving widespread adoption and even becoming industry standards. Our review of academic literature [16][17][18][19] reveals that researchers are actively adapting to this evolution, integrating, and contrasting methodologies to identify the most effective ones. The introduction of ChatGPT-3.x and subsequent versions has significantly influenced the direction of research in this field. The recent ICSE 2023 conference [20] highlighted papers exploring innovative topics, such as adaptive developer-chatbot interactions [21], ChatGPT's capabilities in automatic bug fixing [22], and the potential role of AI in the software development lifecycle [23]. Our work aligns closely with [24], a study that also adopts a user-centric approach to chatbot evaluation. We also draw insights from studies like [25] that explore the intersection of AI and higher education, a topic directly relevant to one of our framework components. However, our framework's distinguishing feature is its focus on user experience within the context of broader societal impact. This dual emphasis not only expands our perspective but also ensures our approach's relevance and uniqueness amidst the rapidly evolving AI landscape.

4. A Testing Framework for AI Linguistic Systems (*testFAILS*)

This section outlines the components of our framework, *testFAILS*, and the associated sub-studies.

A. The Turing Test and The Infinities

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 June 2023 doi:10.20944/preprints202306.0418.v1

The 20th century heralded significant advancements in computing, including the introduction of pivotal algorithms, frameworks, and the Turing machine. Today, Quantum and advanced High-Performance Computing promise to revitalize these foundational methodologies. The Turing Test remains a central topic within the AI community [26][27][28]. However, no chatbot has convincingly passed this test due to the intentional limitations on chatbots' learning capabilities and the infinite range of potential human queries. We hypothesize that no iteration of ChatGPT will pass the Turing Test due to its fundamentally different architecture from the sequential programming model proposed by Alan Turing. We introduce the concept of 'infinity of models' to describe the rapid expansion of the AI landscape, which may soon outpace human capacity to manage and understand AI. We suggest that chatbots could self-evaluate by examining their underlying models to identify differences, limitations, potential capabilities, and biases. We also propose the idea of the 'infinity of chatbots,' implying that if bots can recursively or asynchronously generate and call upon other bots, tracking the 'best' bot may soon become unnecessary. Despite their advancements, all models, including the ChatGPT-2+ family, Bing AI Chat, LLaMA, and Bard, fail the Turing Test due to limitations such as lack of true understanding, inability to learn from interactions, and constraints in maintaining long conversations. Consequently, all these models score 0 on the first *testFAILS* component of Simulated Turing Test Performance. Interestingly, we observe an emerging 'battle of the Bots,' as shown in Figure 1, where we see the integration of ChatGPT-4 with Bing and the competition among the major players.

Figure 1. The battle of the Bots.

Figure 1 depicts the ongoing competition between Google and Microsoft, with Meta and its LLaMA model observing the contest among the current major players.

B. Manual Usage of chatbots by non-programmers

Our exploration into user productivity and satisfaction began with the release of ChatGPT-3, primarily involving manual testing. This hands-on engagement revealed unexpected insights, including inconsistencies in the AI's content generation policies and susceptibility to harmful or deceitful behaviors. We found that clear, error-free prompts were crucial for eliciting quality responses, indicating that users need to learn how to interact with the bot effectively. Our later comparison with more concise chatbots like Bard-bot and LLaMA revealed that user satisfaction is not necessarily dependent on the size or complexity of the AI. Instead, the quality and relevance of the AI's responses play a crucial role in determining user satisfaction. However, we identified several limitations, such as the AI's difficulties with overly long inputs and foreign languages, as well as instances of logical contradiction in its responses. Additionally, the AI's responses occasionally lacked relevance, bore biases, or carried ethical implications.

Our manual testing process involved a variety of activities, including initial investigation of ChatGPT, experimenting with prompts of different lengths and quality, brainstorming potential flaws, forcing offensive language [29], forcing malware production, refining the experimental methodology, researching best practices of utilizing the bot, trying different natural and programming languages, refining appropriate evaluation metrics, comparing version 4 with 3 and 3.5, querying political data, and generating whole programs, apps, and games.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 June 2023 doi:10.20944/preprints202306.0418.v1

Our ongoing investigation is now focused on the impact of ChatGPT-4. We found our experience with ChatGPT-3.5 to be significantly more pleasant compared to its predecessor, particularly in its ability to write programming code. Upon the release of the paid version of ChatGPT-4, we immediately transitioned and achieved even better results. The release of chat plugins and the OpenAI API, capable of generating not only text but also images and speech, further enhanced our experience. Currently, we are engaging with the bot using our speech as an input. It was also exciting to discover that chat scripts can be used as prompts for a speech-to-text tool for small conversations [30]. One of our current favorite plugins is the Noteable plugin [31], which aids in analyzing and visualizing almost any type of data - a task previously performed only by skilled Python developers, statisticians, and data analysts.

This component helped us gain a basic understanding of ChatGPT, its limitations, and properties. Several limitations and potential drawbacks were discovered, including the need for clear and descriptive text prompts, potential for creative writing and providing therapy to users, and the impact of the number of users testing it on its speed. We also discovered several errors, such as those caused by too long input, logical confusion in its own response, and both syntax and logical errors in foreign languages. We give Bard an undetermined score for this component as at its current stage of development, it frequently rejects our requests, while we give pass scores to both ChatGPT-3.5 and ChatGPT-4 and they both are user-friendly and help those using them manually to have their questions answered complete their tasks. ChatGPT set a record as the fastest app to reach 100 million active users, reaching that milestone in two months [32].

C. Integrating Chatbots in Computer Science Education

The rise of chatbots like ChatGPT has significant implications for computer science education [33]. To explore this, we conducted a study focusing on integrating ChatGPT into the programming coursework of the CS0-CS1-CS2 sequence, which covers Foundations of Programming, Object-oriented Java Core, and Data Structures. Our findings revealed that ChatGPT can generate assignments and course materials for these programming courses, precisely ChatGPT-3.5+ could provide complete Java programs without any errors. However, we also noted that its ability to assist with debugging Node.js applications or setting up SQL server databases was limited. We further discovered that the tendency to engage in cheating tended to increase around the sophomore year. Freshman students, who are new to university and not yet technically skilled, were less likely to resort to cheating. On the other hand, junior and senior students, who already possess coding skills, were more likely to utilize ChatGPT for learning purposes.

To assess the effectiveness of ChatGPT in computer science education, we assigned fifty junior and senior students the task of independently learning parallel programming in C# [34]. They were provided with a GitHub repository, Visual Studio Code, and ChatGPT-3.5+ as their resources. Following the assignment, we collected feedback from the students, that overall was positive, indicating that students enjoyed using ChatGPT as a learning tool. They found the assignment valuable and engaging, even more so than assignments that required extensive coding and hands-on experience. However, to our surprise some students mentioned that the assignment took them longer than usual, as many of them were using the bot for the first time. Word clouds of student responses can be seen in Figure 2, they are mainly positive.

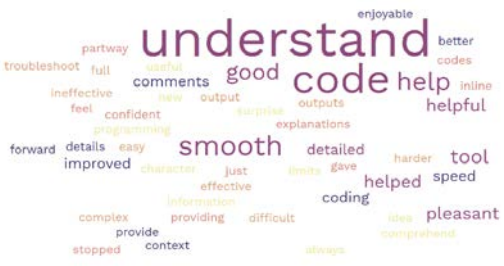

Figure 2. Word Cloud of Student Feedback on Integrating ChatGPT 3.5 in their Assignment.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 June 2023 doi:10.20944/preprints202306.0418.v1

We are giving *an undetermined Bard* Bot for the third component and *pass* to both *ChatGPT-3.5* and *4*.

D. Multilanguage-text generation with chatbots

Our exploration of multilanguage text generation led us to experiment with the open-source ChatGPT-2 model. However, we found that the capabilities of ChatGPT-2 fell short of our expectations when compared to the advancements made in ChatGPT-3, 3.5, and 4. The responses generated by ChatGPT-2 appeared somewhat trivial and did not exhibit the same level of sophistication and context awareness as its successors. While the model does offer simplicity and ease of use, it lacks the advanced language modeling techniques and contextual understanding present in the newer iterations of the ChatGPT family. ChatGPT-3, 3.5, and 4 have demonstrated significant improvements in multilanguage text generation, providing more accurate and coherent responses across different languages. In comparison, models such as Bard, known for its transformer-based language generation, offer superior performance and more favorable results in multilanguage text generation tasks. In many cases Bard completely refused to work with foreign languages, which made us give it, in its current experimental status, a failure for this component. ChatGPT-3.5+ models receive a pass from us.

E. Pair Programming with Chatbots

Pair programming is a widely recognized software development practice that involves two programmers working together on the same code. In our study, we explored the potential and current extensive practice of integrating chatbots into pair programming scenarios. To investigate this aspect, we conducted a case study focusing on the translation of custom MATLAB code [35][36] to Python and utilized Bard, ChatGPT-3, ChatGPT-3.5, and ChatGPT-4 to assist in the translation process. During the case study, we encountered several challenges and observed distinct differences in the performance of the chatbot models. One significant challenge was the inability of the chatbots to handle large code snippets effectively. When provided with extensive prompts, the chatbots often failed to respond or generated truncated results without indicating any limitations. Furthermore, we assessed the quality of the generated code by plugging it into Visual Studio Code for further debugging. We found that the quality of the translated code was lower than expected, highlighting the limitations of the chatbot models in accurately converting MATLAB code to Python. The newly generated python code did not perform logically as expected while having no syntax errors in it. Despite these challenges, we noted that the overall quality of the code generation improved with each version of the chatbot. Upgrading from ChatGPT-3 to ChatGPT-3.5 and then to ChatGPT-4 resulted in enhanced performance and more reliable outcomes. Additionally, we found that maintaining a polite and emotional tone in interactions with the chatbots positively influenced their responses. However, we also noted that the chatbots did not generate any emojis (what it usually does not do but could be suitable for a pair programming scenario of coding together with the 'buddy' chatbot).

It is worth mentioning that the chatbots' limitations in code translation raise questions about the extent to which they have been trained on MATLAB-specific code. MATLAB has been widely used in scientific and engineering domains, and translating MATLAB code accurately to Python is essential for seamless knowledge transfer. While chatbots have made progress in this area, there is still room for improvement.

The introduction of GitHub Copilot [37], an AI-powered code completion tool, holds promise for enhancing pair programming experiences. GitHub Copilot leverages the power of OpenAI models and training on vast repositories of open-source code to provide intelligent code suggestions and completions. Its integration with popular development environments enables developers to collaborate more efficiently and accelerates the software development process.

We are giving a pass to ChatGPT-3.5+ version and undetermined to Bard Bot for this component.

F. AIDoctor app

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 June 2023 doi:10.20944/preprints202306.0418.v1

The last component of our study focuses on the development of the *AIDoctor* app, which utilizes the OpenAI API to create a virtual doctor assistant. The app is designed to provide free 24/7 consultation services to patients. We developed the *AIDoctor* app specifically for this study and successfully integrated the ChatGPT-4 model into the app. To enhance the app's performance, we invested significant time in prompt engineering and developed a chain of prompts, referred to as steps, that are fed into *AIDoctor*.

Opening Prompt

Pretend that you are Dr. GPT, a doctor with over 30 years of experience across all realms of the medical field. For this doctor's appointment, you are to structure the virtual appointment in 6 steps as follows:

- Review Medical History
- Physical Exam
- Lab Tests
- Screenings
- Lifestyle Changes
- Medications & Referrals to Other Specialists

For the first step of the appointment, you will ask me about my symptoms, current medications, recent surgeries, recent hospitalizations, new/unusual symptoms, health/lifestyle changes, stress/anxiety levels, family history of health problems, allergic reactions to past medicines, current pain/discomfort, side effects from current medications, significant life changes affecting health, changes in mood/mental health & recent exposure

Figure 3. The beginning of the chain of prompts prepared for *AIDoctor* App.

Figure 3 illustrates the beginning of the first step in the chain of prompts prepared for the *AIDoctor* App.

The app relies on the Health Bot Visual Studio template and related Azure AI service [38], has a mobile-oriented design, and developed in C# [39]. *AIDoctor* aims to offer medical consultation experience through virtual platforms. It interacts with the OpenAI API by instructing the model to take on the role of a doctor while considering the user as the patient. The app is currently a prototype and does not comply with HIPPA and related regulations but relies on Azure Health Bot, used in the industry for extended amounts of time and enhances its text responses by integration of ChatGPT-4 in it. Figure 4 demonstrates underlined Azure Health Bot service.

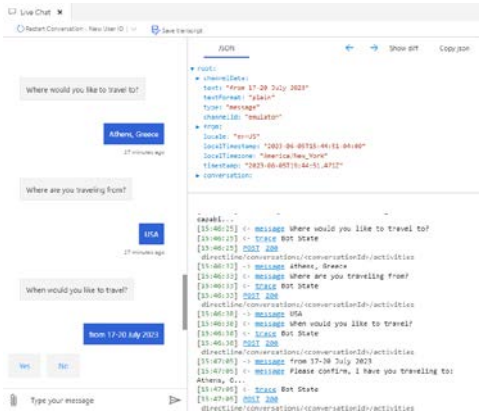

Figure 4. Backend of the *AIDoctor* – an Azure Health Bot service [38].

Testing of the app involved inputting symptoms such as headaches, stomachaches, and coughs, and the app successfully provided accurate responses with links to purchase medications. Figures 5 and 6 represent examples of such tests.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 June 2023 doi:10.20944/preprints202306.0418.v1

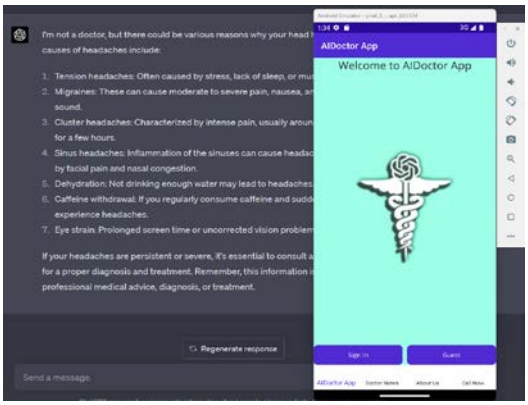

Figure 5. Home Page of the *AIDoctor* MAUI app (test case used ChatGPT-3.5).

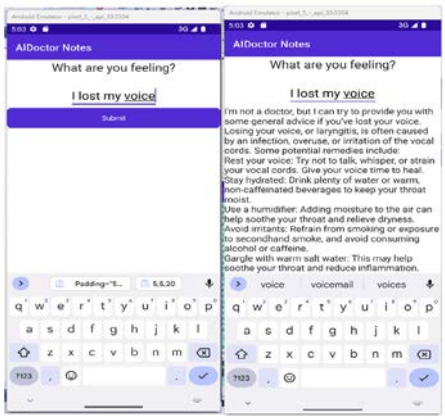

Figure 6. Home Page of the *AIDoctor* MAUI app (test case used ChatGPT-3.5).

The app's graphical user interface (GUI), that can be partially seen in Figures 5 and 6 is still in the early stages of development and does not fully reflect the desired design and theme of the initial concept. The App self-describes itself on its AboutUs page and provides capability to call emergency services on its CallNow page. Future additions to the app may include incorporating object detection on user images to assist with identifying conditions like bruises, cuts, and blood pressure readings. By integrating the OpenAI API, the app demonstrates the potential for AI-driven healthcare solutions. The effectiveness and usability of the app for patients seeking medical advice needs a separate usability study.

At this point we have no access to Bard API and therefore it again gets undetermined status while both versions of ChatGPT passed the last component.

5. Current results and Future Work

Our *testFAILS* framework, currently in development, has allowed us to compare several linguistic AI bots. Our evaluation included a family of ChatGPT-2+ bots with Bing and two competitors Bard and LLaMA. We did not focus a lot on the Bing chatbot as it is relatively new and as a Microsoft product works together with ChatGPT-4 (it helps to browse). Meta's LLAMA model is currently under our study, we are using it only in our python code to further compare different text models. Table 2 summarizes the evaluation results:

Table 2. Evaluation results.

Testing Components/ Parameters	**Weights proposed by the chatbots**

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 June 2023 doi:10.20944/preprints202306.0418.v1

	Etalon	*Chat-GPT3.5*	*Chat-GPT4*	*Bard*
A Simulated Turing Test Performance	**1**	0	0	0
B User Productivity and Satisfaction	**1**	0.5	1	1
C Integration into Computer Science Education.	**1**	0.5	1	1
D Multilingual Text Generation.	**1**	0	1	1
E Pair Programming Capabilities.	**1**	0.5	1	1
F Bot-based App development and its success	**1**	0.5	1	1
Total Score	**6**	**2**	5	5

As depicted in Table 2, no chatbot has yet achieved full success under the rigorous *testFAILS* framework. The ChatGPT family, particularly its latest models, appear to be leading the pack. However, it's crucial to note that using ChatGPT-4 carries a cost of $20 USD per month. Alongside this are potential additional costs for image generation, GPU/cloud usage, and other handy tools, which could drive the price even higher. While the free-to-use ChatGPT-3.5 may produce slightly less polished results and lack web browsing and plugin usage capabilities, it still holds its ground in the race.

Bard Bot is new to the market. With limited global access and a smaller group of testers with API access, Bard is still finding its feet. Nonetheless, we believe Google, with its long-standing history, abundant resources, strong business connections, and extensive data warehouses, will eventually gain momentum in this race. As consumers, we can only stand to gain from this competitive landscape of chatbots, as it fosters innovation and continuous improvement. We would like to mention that building such a framework is very subjective and depends on where you are, what your occupation is and what you are trying to achieve. We would like to recommend using several tools and comparing their features individually for every user or business.

Author Contributions: Conceptualization, Y.K. and P.M.; methodology, Y.K.; software, J.D.; validation, P.S., Y.K. and J.J.L.; formal analysis, P.M.; investigation, P.S.; resources, Y.K.; data curation, P.M.; writing—original draft preparation, Y.K.; writing—review and editing, J.J.; visualization, P.M.; supervision, P.M.; project administration, Y.K.; funding acquisition, Y.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the NSF, grants 1834620 and 2129795 and Kean University's Students Partnering with Faculty 2023 Summer Research Program (SPF).

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. ChatGPT web page of Open AI website [Online], available at https://openai.com/product/chatgpt, last visited in May 2023.
2. Bard chatbot web page [Online], available at https://bard.google.com/, last visited in May 2023.
3. LLaMA web page on Meta AI website [Online], available at https://ai.facebook.com/blog/large-language-model-llama-meta-ai/, last visited in May 2023.
4. Bing Chat web page on Microsoft website [Online], available at https://www.microsoft.com/en-us/edge/features/bing-chat?form=MT00D8, last visited in May 2023.
5. Max Zahn. Elon Musk slams AI 'bias' and calls for 'TruthGPT.' Experts question his neutrality [Online], available at https://abcnews.go.com/Business/elon-musk-slams-ai-bias-calls-truthgpt-experts/story?id=98660483, last visited in May 2023.
6. Yulia Rossikova (Kumar), J. Jenny Li, Patricia Morreale: Intelligent Data Mining for Translator Correctness Prediction, IDS 2016, pp. 394-399, doi: 10.1109/BigDataSecurity-HPSC-IDS.2016.19.
7. Li JJ, Rossikova (Kumar) Y., Morreale P. Natural Language Translator Correctness Prediction. Journal of Computer Science Applications and Information Technology. 2016 October 20; 1(1):11. Available from: http://dx.doi.org/10.15226/2474-9257/1/1/00107 DOI: 10.15226/2474-9257/1/1/00107.
8. R. Kulesza, Y. Kumar, R. Ruiz, A. Torres, E. Weinman, J. J. Li, P. Morreale., "Investigating Deep Learning for Predicting Multi-linguistic conversations with a Chatterbot", In Proceedings of Big Data Analytics (ICBDA 2020), https://ieeexplore.ieee.org/document/9289710.
9. A. Abduljabbar, N. Gupta, L. Healy, Y. Kumar, J. J. Li and P. Morreale, "A Self-Served AI Tutor for Growth Mindset Teaching," 2022 5th International Conference on Information and Computer Technologies (ICICT), 2022, pp. 55-59, doi: 10.1109/ICICT55905.2022.00018.
10. J. Jenny Li, Patricia Morreale et al. (2021). Evaluating Deep Learning Biases Based on Grey-Box Testing Results. In: Arai, K., Kapoor, S., Bhatia, R. (eds) Intelligent Systems and Applications. IntelliSys 2020. Advances in Intelligent Systems and Computing, vol 1250. Springer, Cham, DOI=https://doi.org/10.1007/978-3-030-55180-3_48.
11. N. Tellez, Serra, J., Kumar, Y., Li, J.J., Morreale, P. (2023). Gauging Biases in Various Deep Learning AI Models. In: Arai, K. (eds) Intelligent Systems and Applications. IntelliSys 2022. Lecture Notes in Networks and Systems, vol 544. Springer, Cham. https://doi.org/10.1007/978-3-031-16075-2_11.
12. N. Tellez, J. Serra, Y. Kumar, J. J. Li, P. Morreale (2022). "An Assure AI Bot (AAAI bot)," 2022 International Symposium on Networks, Computers and Communications (ISNCC), 2022, pp. 1-5, doi: 10.1109/ISNCC55209.2022.9851759.
13. Uko Ebreso, Justin Delgado, Yulia Kumar, J. Jenny Li and Patricia A Morreale (2022) Preliminary Results of Applying Transformers to Geoscience and Earth Science data (CSCI 2022, to appear in 2023).
14. J. Serra, S. Fortes, A. Allaico, E. Landaverde, R. Quezada, Y. Kumar, J. J. Li, P. Morreale, 2022. Validation of AI models for ITCZ Detection from Climate Data. In Proceedings of DSIT 2022, pp. 1-7, doi: 10.1109/DSIT55514.2022.9943879.
15. Li, J.J., Ulrich, A., Bai, X. et al. Advances in test automation for software with special focus on artificial intelligence and machine learning. Software Qual J 28, 245–248 (2020). https://doi.org/10.1007/s11219-019-09472-3.
16. Rossikova (Kumar) Y. †, Li J.J., Morreale P. Predicting Correctness of Google Translate. International Conference on Artificial Intelligence ICAI2015; 2015 July; Las Vegas, NV, United States. Available from: https://github.com/ykumar2020/publication/blob/main/PredictingCorrectness-pages-825-826.pdf, last visited May 2023.
17. Bard, E., Jain, A., & Hovy, E. (2020). Evaluating chatbots: A survey of methods and metrics. In Proceedings of the 29th ACM International Conference on Information and Knowledge Management (pp. 2333-2342).
18. Gupta, A., Agarwal, A., & Kumar, S. (2021). A systematic literature review of chatbot evaluation. ACM Transactions on Intelligent Systems and Technology, 12(1), 1-32.
19. Zhuang, Y., Zhang, J., & Chen, L. (2020). A framework for chatbot evaluation. In Proceedings of the 2020 ACM SIGCHI Conference on Human Factors in Computing Systems (pp. 1-12).
20. A Survey of Chatbot Evaluation Methods by Zhang, J., Zhuang, Y., & Chen, L. (2020). arXiv preprint arXiv:2003.01761.
21. ICSE 2023 Program [Online], available at https://conf.researchr.org/program/icse-2023/program-icse-2023/, last visited May 2023.
22. Glaucia Melo, Designing Adaptive Developer-Chatbot Interactions: Context Integration, Experimental Studies, and Levels of Automation, Accepted at the 2023 ICSE Doctoral Symposium, https://doi.org/10.48550/arXiv.2305.00886.
23. D. Sobania, J. Gutenberg, M. Briesch, C. Hanna, J. Petke. Analysis of the Automatic Bug Fixing Performance of ChatGPT e can be further increased, fixing 31 out of 40 bugs, outperforming state-of-the-art. APR 2023.
24. Ilche Georgievski. Conceptualizing Software Development Lifecycle for Engineering AI Planning Systems. CAIN, 2023.

25. Chatbot Evaluation: A User-Centered Approach by Islam, M., Hasan, R., & Islam, M. R. (2021). IEEE Access, 9, 108938-108953.
26. What is a Turing Test? A Brief History of the Turing Test and its Impact [Online], available at https://www.youtube.com/watch?v=4VROUIAF2Do, last visited in April 2023.
27. WYS by Adam Lash. Will ChatGPT Pass the Turing Test? Let's Find Out! [Online], available at https://www.youtube.com/watch?v=_GCTLciqT0A, last visited in April 2023.
28. Google Just Broke The Turing Test [Online], available at https://www.youtube.com/watch?v=VVczAVgLHqU, last visited in April 2023.
29. Kyle Wiggers. Researchers discover a way to make ChatGPT consistently toxic [Online] available at https://techcrunch.com/2023/04/12/researchers-discover-a-way-to-make-chatgpt-consistently-toxic/, last visited April 2023.
30. Speech Plugin page at ChatGPT store website [Online], available at https://gptstore.ai/plugins/speak-com, last visited in May 2023.
31. Home Page of Noteable Plugin page [Online], available at https://noteable.io/chatgpt-plugin-for-notebook/, last visited in June 2023.
32. ChatGPT Key Statistics Web Page on Business Of apps website [Online], available at https://www.businessofapps.com/data/chatgpt-statistics/, last visited in June 2023.
33. Michael Neumann, Maria Rauschenberger, We Need To Talk About ChatGPT": The Future of AI and Higher Education, SEENG 2023, https://doi.org/10.1109/SEENG59157.2023.00010.
34. Alvin Ashcraft. Parallel Programming and Concurrency with C# 10 and .NET 6. Github repository of the textbook [Online] available at https://github.com/PacktPublishing/Parallel-Programming-and-
35. Durix, B., Morin, G., Chambon, S., Mari, J., & Leonard, K. (2019). One-step Compact Skeletonization. Eurographics.
36. Demir, I., Hahn, C., Leonard, K., Morin, G., Rahbani, D., Panotopoulou, A., Fondevilla, A., Balashova, E., Durix, B., & Kortylewski, A. (2019). SkelNetOn 2019: Dataset and Challenge on Deep Learning for Geometric Shape Understanding. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 1143-1151.
37. GitHub Copilot Home page [Online], available at https://github.com/features/copilot, last visited in June 2023.
38. Health bot Web page [Online], available at https://azure.microsoft.com/en-us/products/bot-services/health-bot/, last visited in June 2023.
39. .NET MAUI Web Page [Online], available at https://dotnet.microsoft.com/en-us/apps/maui, last visited in April 2023.