Jurnal Riset Sistem Informasi Dan Teknik Informatika (JURASIK)
Volume 9, Nomor 1, Februari 2024, pp 248-256
Terakreditasi Nomor 204/E/KPT/2022 | ISSN: 2527-5771/EISSN: 2549-7839
https://tunasbangsa.ac.id/ejurnal/index.php/jurasik

Evaluasi Responsivitas dan Akurasi: Perbandingan Kinerja ChatGPT dan Google BARD dalam Menjawab Pertanyaan seputar Python

Yayan Heryanto[1], Fauziah[2], Frenda Farahdinna[3], Sigit Wijanarko[4]
[1,2]Magister Teknologi Informasi, Fakultas Teknologi Komunikasi dan Informatika, Universitas Nasional, Indonesia
[3,4]Informatika, Fakultas Teknologi Komunikasi dan Informatika, Universitas Nasional, Indonesia
Email: 2022.yayan.heryanto@student.unas.ac.id[1], fauziah@civitas.unas.ac.id[2], frenda.farahdinna@civitas.unas.ac.id[3], sigit.wijanarko@civitas.unas.ac.id[4]

Abstract

This reseach aims to evaluate the responsiveness and accuracy of two natural language processing systems, namely ChatGPT and Google BARD, in answering questions related to the Python programming language. The evaluation is conducted using the Bleu Score metric as an indicator of the accuracy of answers generated by both systems. This research involves experiments with various Python-related questions to measure the level of alignment with expected reference answers. The results indicate that the average Bleu Score for ChatGPT is 0.0088, while the average Bleu Score for Google BARD is 0.0073. Additionally, the response time for ChatGPT is recorded at 12.05 seconds, whereas Google BARD has a response time of 18.38 seconds. Although there is a small difference in accuracy, ChatGPT shows a slightly higher Bleu Score and faster response time compared to Google BARD. The conclusion of this research states that, in the context of answering questions related to the Python programming language, ChatGPT performs slightly better than Google BARD, measured in terms of answer accuracy and response time.

Keywords*: ChatGPT, Bard, Python*

Abstrak

Penelitian ini bertujuan untuk mengevaluasi responsivitas dan akurasi dua sistem pemrosesan bahasa alami, yaitu ChatGPT dan Google BARD, dalam menjawab pertanyaan seputar bahasa pemrograman Python. Evaluasi dilakukan menggunakan metrik Bleu Score sebagai indikator akurasi jawaban yang dihasilkan oleh kedua sistem. Penelitian ini melibatkan uji coba pertanyaan beragam terkait Python untuk mengukur tingkat kesesuaian jawaban dengan referensi yang diharapkan. Hasil penelitian menunjukkan bahwa rata-rata Bleu Score untuk ChatGPT adalah sebesar 0.0088, sedangkan rata-rata Bleu Score untuk Google BARD adalah 0.0073. Selain itu, waktu respon ChatGPT tercatat sebesar 12.05 detik, sedangkan Google BARD memiliki waktu respon 18.38 detik. Meskipun terdapat perbedaan kecil dalam tingkat akurasi, ChatGPT menunjukkan Bleu Score yang sedikit lebih tinggi dan waktu respon yang lebih cepat dibandingkan dengan Google BARD. Dari penelitian ini menyatakan bahwa, dalam konteks menjawab pertanyaan seputar bahasa pemrograman Python, ChatGPT memiliki kinerja yang sedikit lebih baik daripada Google BARD, diukur dari segi akurasi jawaban dan waktu respon.

Kata kunci*: ChatGPT, Bard, Python.*

1. PENDAHULUAN

Dalam lanskap pemrosesan bahasa alami yang terus berkembang, evaluasi responsivitas dan akurasi dalam sistem pemrosesan bahasa menjadi sangat penting. Jurnal ini menggali analisis perbandingan antara dua sistem terkemuka, yakni ChatGPT dan Google BARD [1], dengan fokus pada kemampuan keduanya dalam menanggapi pertanyaan terkait bahasa pemrograman Python [2].

Signifikansi dari evaluasi semacam ini terletak pada potensinya untuk mengungkap wawasan tentang efektivitas sistem-sistem ini dalam memberikan respons yang akurat dan tepat waktu dalam konteks pertanyaan pemrograman [3]. Seiring terus meningkatnya tuntutan akan pemahaman dan generasi bahasa alami yang efisien, perlunya mengawasi dan membandingkan kinerja model bahasa canggih menjadi krusial. Penelitian ini menggunakan metrik Bleu Score [4], sebuah ukuran yang sangat diakui untuk menilai kualitas teks yang dihasilkan, sebagai indikator kunci untuk mengevaluasi akurasi respons yang dihasilkan oleh ChatGPT dan Google BARD. Melalui serangkaian eksperimen yang beragam mencakup berbagai pertanyaan terkait Python, penelitian ini bertujuan untuk mengukur sejauh mana respons-respons tersebut sesuai dengan referensi yang diharapkan, membuka wawasan tentang kemampuan unik dari masing-masing sistem. Hasil penelitian, yang diungkapkan melalui Bleu Score dan waktu respons, memberikan dasar kuantitatif untuk membandingkan kinerja ChatGPT dan Google BARD [5]. Dengan menjelajahi analisis perbandingan ini, kami berupaya memberikan wawasan berharga tentang kelebihan dan kekurangan dari kedua sistem pemrosesan bahasa, khususnya dalam domain pemrograman Python. Hasil penelitian ini diharapkan dapat memberikan kontribusi dalam mendiskusikan optimalisasi sistem-sistem ini guna meningkatkan akurasi dan efisiensi dalam menanggapi pertanyaan terkait bahasa pemrograman.

2. METODOLOGI PENELITIAN

Desain penelitian yang diterapkan dalam kajian ini adalah desain eksperimental yang memanfaatkan dua kelompok perlakuan, yakni menggunakan model bahasa ChatGPT dan Google BARD [6]. Kedua kelompok tersebut dijadikan subjek eksperimen untuk tujuan menilai kinerja responsivitas dan akurasi masing-masing model terhadap pertanyaan seputar Python. Pendekatan eksperimental ini dipilih agar penelitian dapat menghasilkan pemahaman yang mendalam tentang perbandingan kinerja kedua model, dengan mengendalikan faktor-faktor yang dapat memengaruhi hasil penelitian secara lebih terkontrol. Hal ini diharapkan dapat memberikan pemahaman yang lebih komprehensif tentang kelebihan dan kelemahan dari ChatGPT dan Google BARD dalam konteks menjawab pertanyaan seputar bahasa pemrograman Python [7].

2.1. ChatGPT

ChatGPT adalah model bahasa alami yang dikembangkan oleh OpenAI menggunakan arsitektur GPT (Generative Pre-trained Transformer) [8]. Didesain untuk berinteraksi dengan pengguna dalam percakapan sehari-hari, ChatGPT mampu memahami dan menghasilkan teks berdasarkan konteks yang diberikan. Model ini telah melibatkan proses pelatihan yang luas dengan memanfaatkan sejumlah besar data teks dari berbagai sumber di internet. Dengan kecerdasannya yang bersifat umum, ChatGPT dapat digunakan untuk menjawab pertanyaan, memberikan informasi, atau bahkan untuk tujuan hiburan. Meskipun memiliki kemampuan impresif, perlu diingat bahwa ChatGPT mungkin juga menghasilkan jawaban yang tidak selalu benar atau sesuai konteks, sehingga pengguna harus

menggunakan kebijaksanaan dalam mengevaluasi dan menginterpretasi keluaran dari model ini [9].

2.2. Google Bard

Bard adalah chatbot kecerdasan buatan generatif yang dikembangkan oleh Google [10]. Awalnya berbasis pada keluarga model bahasa besar (LLMs) LaMDA, kemudian ditingkatkan menjadi PaLM, dan kemudian menjadi Gemini. Bard dikembangkan sebagai respons langsung terhadap popularitas pesat ChatGPT milik OpenAI. Bard awalnya dirilis dalam kapasitas terbatas pada Maret 2023 dengan tanggapan yang kurang antusias, sebelum kemudian diperluas ke negara-negara lain pada bulan Mei. LaMDA dikembangkan dan diumumkan pada tahun 2021 [11], tetapi tidak dirilis ke publik karena kehati-hatian. Peluncuran ChatGPT oleh OpenAI pada November 2022 dan popularitasnya yang mendadak membuat eksekutif Google terkejut dan memicu respons besar dalam beberapa bulan berikutnya [12]. Setelah menggerakkan kekuatan kerjanya, perusahaan meluncurkan Bard pada Februari 2023, dengan chatbot tersebut menjadi pusat perhatian selama tahun 2023.

2.3. Python

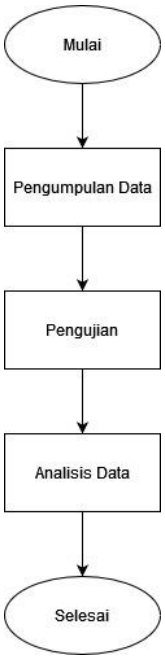
Dalam penelitian ini, pertanyaan yang diajukan kepada kedua chat AI berkaitan dengan Bahasa Pemrograman Python [13]. Pertanyaan-pertanyaan ini diperoleh dari sumber berupa website guru99.com [14]. Python adalah bahasa pemrograman yang mudah dipelajari dan digunakan [15]. Dirancang untuk membaca dan menulis kode dengan mudah, Python menggunakan indentasi sebagai cara untuk menentukan blok kode. Bahasa ini mendukung berbagai paradigma pemrograman, termasuk berorientasi objek, fungsional, dan prosedural. Python dilengkapi dengan pustaka standar yang kaya, menyediakan modul dan fungsi untuk berbagai tugas. Kelebihan Python termasuk komunitas pengembang yang besar, dokumentasi yang baik, dan dukungan untuk sistem operasi berbeda. Digunakan secara luas dalam berbagai bidang seperti pengembangan web, ilmu data, kecerdasan buatan, dan lainnya, Python terus berkembang dengan pembaruan berkala. Bersifat open source, Python menjadi pilihan populer baik untuk pemula maupun pengembang berpengalaman.

2.4. Bleu Score

Dalam penelitian ini, kami menggunakan BLEU Score di platform Python untuk mengevaluasi seberapa baik jawaban dari Chat GPT dan Google BARD sesuai dengan referensi dari website guru99.com. BLEU Score merupakan cara sederhana untuk mengukur sejauh mana jawaban kita cocok dengan jawaban referensi yang sudah ada [16]. Metode ini umum digunakan dalam penilaian sistem terjemahan otomatis, terutama terkait dengan bahasa. BLEU menghitung seberapa sering kata-kata atau kelompok kata yang muncul dalam jawaban kita juga muncul dalam jawaban referensi [17]. Semakin banyak kesamaan kata atau kelompok kata, semakin tinggi skor BLEU-nya. Skor maksimalnya adalah 1, yang berarti jawaban kita sempurna sesuai dengan referensi.

Jurnal Riset Sistem Informasi Dan Teknik Informatika (JURASIK)
Volume 9, Nomor 1, Februari 2024, pp 248-256
Terakreditasi Nomor 204/E/KPT/2022 | ISSN: 2527-5771/EISSN: 2549-7839
https://tunasbangsa.ac.id/ejurnal/index.php/jurasik

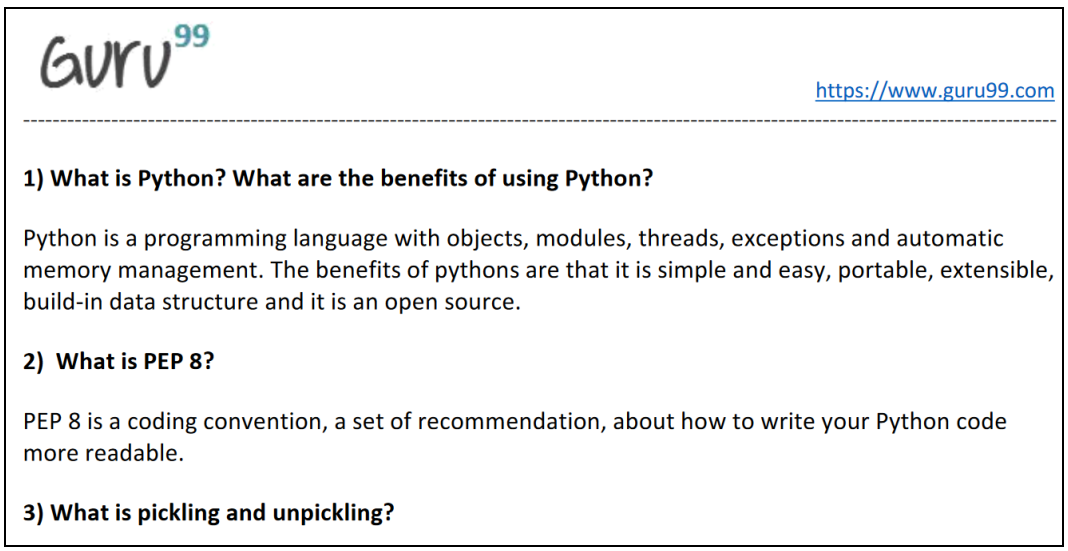
2.5. Flowchart Tahapan Proses Aplikasi



Gambar 1. Tahapan Proses Aplikasi

2.4.1. Pengumpulan Data

Penelitian ini mengawali tahapan pengumpulan data dengan merancang serangkaian pertanyaan yang terfokus pada bahasa pemrograman Python. Pertanyaan-pertanyaan ini dipilih secara cermat dari sumber tepercaya, yaitu website www.guru99.com seperti pada gambar 2, untuk memastikan keberagaman dan relevansi dalam menguji responsivitas dan akurasi dari dua model bahasa, ChatGPT dan Google BARD [18].


Guru99

https://www.guru99.com

1) What is Python? What are the benefits of using Python?

Python is a programming language with objects, modules, threads, exceptions and automatic memory management. The benefits of pythons are that it is simple and easy, portable, extensible, build-in data structure and it is an open source.

2) What is PEP 8?

PEP 8 is a coding convention, a set of recommendation, about how to write your Python code more readable.

3) What is pickling and unpickling?

Gambar 2. Website guru99, Sumber https://www.guru99.com

Jurnal Riset Sistem Informasi Dan Teknik Informatika (JURASIK)
Volume 9, Nomor 1, Februari 2024, pp 248-256
Terakreditasi Nomor 204/E/KPT/2022 | ISSN: 2527-5771/EISSN: 2549-7839
https://tunasbangsa.ac.id/ejurnal/index.php/jurasik

Dengan memanfaatkan sumber informasi yang dapat diandalkan, langkah ini bertujuan untuk membentuk landasan yang kuat dalam menguji dan membandingkan kinerja kedua model dalam menanggapi pertanyaan seputar Python.

2.4.2. Pengujian

Setelah itu, data pertanyaan dan jawaban dikumpulkan secara rinci selama tahap pengujian, di mana kedua model, ChatGPT dan Google BARD, diuji dengan serangkaian pertanyaan tersebut seperti pada tabel 1.

Tabel 1. Contoh Hasil Jawaban.

Pertanyaan	Jawaban Referensi	Chat GPT	Google Bard
What is Python? What are the benefits of using Python?	Python is a programming language with objects, modules, threads, exceptions and automatic memory management. The benefits of pythons are that it is simple and easy, portable, extensible, build-in data structure and it is an open source.	PEP 8 is the Python Enhancement Proposal that establishes the style guide for writing Python code. "PEP" stands for "Python Enhancement Proposal," and PEP 8 specifically addresses the conventions for writing readable and maintainable code in the Python programming language. Created by Guido van Rossum, Barry Warsaw, and Nick Coghlan, PEP 8 covers various aspects of code style, including indentation, spacing, naming conventions, and more. Adhering to PEP 8 helps maintain a consistent coding style across Python projects, making it easier for developers to understand and collaborate on code. Some key points covered by PEP 8 include: Indentation: Python code uses indentation to define blocks of code, and PEP 8 recommends using four spaces per indentation level. Whitespace in Expressions and Statements: PEP 8 defines rules for the use of whitespace in expressions and statements, promoting readability. Imports: Guidelines for importing modules and organizing import statements are outlined in PEP 8. Whitespace in Function and Method Calls: PEP 8 provides recommendations on how to format whitespace in function and method calls. Comments: Recommendations for writing comments in code, including their frequency and style, are covered. Naming Conventions: PEP 8 outlines conventions for naming variables, functions, classes, and modules. This includes the use of underscores, avoiding single-character names, and adhering to a consistent naming style. Programming Recommendations: PEP 8 includes various programming recommendations, such as preferring exceptions to returning None, using the join() method for string concatenation, and more. Documentation Strings (Docstrings): PEP 8 provides guidance on writing docstrings, which are used to document functions, modules, and classes in Python.	Here's a clear and concise summary of PEP 8, incorporating key takeaways and addressing noted points: PEP 8 is a style guide for Python code, designed to promote: Readability: Ease of understanding code for both the original author and other developers. Maintainability: Simplifying code updates and modifications over time. Consistency: Ensuring a uniform coding style across projects, especially in collaborative settings. Key guidelines within PEP 8 include: Indentation: Use 4 spaces per indentation level to define code blocks. Whitespace: Employ whitespace strategically around operators, expressions, and statements for clarity. Imports: Organize import statements consistently for better readability. Function and method calls: Use whitespace appropriately for visual clarity. Comments: Write clear and concise comments to explain code logic and purpose. Naming conventions: Follow consistent naming rules for variables, functions, classes, and modules to enhance code understanding. Programming recommendations: Adhere to best practices for various Python constructs, such as using join() for string concatenation and preferring exceptions to returning None. Documentation strings (docstrings): Write informative docstrings to document functions, classes, and modules effectively. Benefits of following PEP 8:

		By following PEP 8, developers can create code that is consistent, readable, and maintainable, which is especially important in collaborative projects or when contributing to open-source codebases. Many Python developers and teams adopt PEP 8 as a standard style guide to ensure a uniform coding style throughout their projects. The adoption of PEP 8 helps improve code quality, readability, and the overall development experience in the Python community.	Improved code quality: Consistent and readable code is easier to understand, maintain, and debug. Enhanced collaboration: Uniform style facilitates teamwork and understanding of shared code. Community standards: Adherence to PEP 8 aligns with common Python practices, fostering better communication and collaboration within the Python community. In essence, PEP 8 serves as a valuable guide for writing clear, consistent, and maintainable Python code, benefiting both individual developers and the broader Python community.

Waktu respons, yang mencatat seberapa cepat kedua model merespons pertanyaan, menjadi aspek yang diperhatikan untuk mengevaluasi responsivitas masing-masing. Sementara itu, akurasi jawaban diukur dengan memanfaatkan metode Blue score dan sejumlah kriteria evaluasi yang telah ditetapkan sebelumnya. Penggunaan metode Blue score dalam proses pengujian membantu mengukur sejauh mana jawaban dari kedua model sesuai dengan jawaban referensi manusia [19].

```
from nltk.translate.bleu_score import sentence_bleu, SmoothingFunction
```

Gambar 3. Fungsi sentence_bleu dari pustaka NLTK

Pertama, data dibaca dari file Excel menggunakan pustaka Pandas, dan kolom "Referensi" dan "Prediksi" diambil sebagai referensi dan prediksi teks. Selanjutnya, teks referensi dan prediksi dibagi menjadi daftar kata untuk persiapan perhitungan skor BLEU. Proses perulangan dilakukan untuk setiap pasangan referensi dan prediksi, di mana skor BLEU dihitung menggunakan fungsi sentence_bleu dari pustaka NLTK seperti pada gambar 3. Hasil skor BLEU untuk setiap pasangan disimpan dalam daftar bleu_scores. Program mencetak skor BLEU untuk setiap pasangan, dan pada akhirnya, rata-rata skor BLEU dari semua pasangan dihitung dan dicetak sebagai output program. Algoritma BLEU ini umumnya digunakan dalam evaluasi sistem terjemahan otomatis dan pengujian model generasi teks untuk mengukur sejauh mana teks prediksi mendekati teks referensi manusia, dengan memberikan perhatian pada kesamaan n-gram antara teks referensi dan prediksi.

Metode ini memberikan kerangka evaluasi yang lebih objektif terhadap kualitas jawaban. Selain itu, perhitungan waktu respons memberikan gambaran yang komprehensif tentang kecepatan tanggapan masing-masing model terhadap pertanyaan yang diajukan, menyoroti aspek responsivitas secara lebih mendalam. Setelah berhasil mengumpulkan semua data yang diperlukan, langkah selanjutnya mengarah pada tahapan analisis data. Proses analisis ini mencakup pengolahan skor Blue, evaluasi rata-rata waktu respons, dan perbandingan performa antara ChatGPT dan Google BARD. Dengan demikian, tahapan analisis data menjadi kunci

dalam menggali makna dan implikasi dari hasil penelitian ini terkait responsivitas dan akurasi kinerja kedua model dalam menanggapi pertanyaan seputar bahasa pemrograman Python.

2.4.3. Analisis

Proses analisis ini terlibat dalam perhitungan skor Blue, penilaian terhadap waktu respons rata-rata, dan perbandingan performa antara ChatGPT dan Google BARD [20]. Melalui hasil analisis ini, terungkap temuan-temuan berharga yang memberikan wawasan mendalam mengenai responsivitas dan akurasi kinerja kedua model bahasa. Perhitungan skor Blue memberikan pemahaman yang lebih terperinci mengenai sejauh mana jawaban dari kedua model cocok dengan jawaban referensi manusia, sementara evaluasi waktu respons rata-rata membantu mengidentifikasi perbedaan dalam kecepatan tanggapan keduanya terhadap pertanyaan seputar Python. Perbandingan performa secara keseluruhan memungkinkan penelitian ini untuk memberikan gambaran yang komprehensif tentang kelebihan dan kekurangan masing-masing model, menggali aspek responsivitas dan akurasi yang menjadi fokus utama penelitian ini dalam konteks menjawab pertanyaan dalam domain bahasa pemrograman Python.

3. HASIL DAN PEMBAHASAN

Dalam penelitian ini, kami mengevaluasi responsivitas dan akurasi dua sistem pemrosesan bahasa alami, yakni ChatGPT dan Google BARD, khususnya dalam menanggapi pertanyaan seputar bahasa pemrograman Python.

3.1. Responsivitas Model

Dalam penelitian ini, kami mengevaluasi responsivitas dan akurasi dua sistem pemrosesan bahasa alami, yakni ChatGPT dan Google BARD, khususnya dalam menanggapi pertanyaan seputar bahasa pemrograman Python. Responsivitas ChatGPT menonjol dengan waktu rata-rata respons yang lebih singkat, yakni 12,5 detik, dibandingkan dengan Google BARD yang mencapai 18,38 detik. Perbedaan ini memberikan indikasi kuat bahwa ChatGPT memiliki kemampuan respons yang lebih tinggi dalam menanggapi pertanyaan seputar Python.

3.2. Akurasi Jawaban

Dalam mengukur akurasi, metode Blue score digunakan. Hasilnya menunjukkan bahwa ChatGPT memiliki tingkat akurasi yang sedikit lebih tinggi dengan skor rata-rata 0,0088, sementara Google BARD mencapai 0,0073. Meskipun perbedaannya tipis, hasil ini menandakan kecenderungan ChatGPT memberikan jawaban yang lebih sesuai dengan referensi manusia.

Tabel 2. Hasil Pengujian.

AI	Rata Rata Bleu Score	Rata Rata Waktu Respon

AI	Rata Rata Bleu Score	Rata Rata Waktu Respon
Chat GPT	0.0088	12.05 Detik
Google Bard	0.0073	18.38 Detik

3.3. Implikasi dan Rekomendasi

Hasil penelitian memberikan wawasan lebih dalam tentang kinerja kedua model dalam konteks menjawab pertanyaan Python. Rekomendasi untuk peningkatan model mencakup penyesuaian parameter dan penambahan data pelatihan guna meningkatkan akurasi jawaban dan responsivitas keduanya.

4. SIMPULAN

Secara keseluruhan, dalam konteks evaluasi responsivitas dan akurasi menjawab pertanyaan Python, ChatGPT menunjukkan kinerja yang lebih unggul dibandingkan dengan Google BARD sesuai dengan jenis dan karakteristik data yang diuji coba pada penelitiam ini. Meskipun perbedaan dalam skor Blue dan waktu respons relatif kecil, hasil ini memberikan wawasan yang berharga tentang kelebihan dan kekurangan kedua model bahasa. Rekomendasi untuk pengembangan model, seperti penyesuaian parameter dan penambahan data pelatihan, menjadi panduan untuk meningkatkan kinerja keduanya dalam tugas-tugas serupa di masa depan.

DAFTAR PUSTAKA

[1] Sing, S. K., Kumar, S., Mehra, P. S., “Chat GPT & Google Bard AI: A Review”, IEEE 2023 International Conference on IoT, Communication and Automation Technology (ICICAT), ISBN 979-8-3503-0282-0.

[2] Haau-Sing (Xiaocheng) Li, Mohsen Mesgar, André Martins, Iryna Gurevych, “Python Code Generation by Asking Clarification Questions”, Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp 14287–14306.

[3] Adamson, V., Bägerfeldt, J., “Assessing the effectiveness of ChatGPT in generating Python code”, Independent thesis Basic level (degree of Bachelor), 20 credits / 30 HE credits, pp 38, 2023.

[4] Avisha Das, Rakesh M. Verma, “Can Machines Tell Stories? A Comparative Study of Deep Neural Language Models and Metrics”, IEEE Access (Volume: 8), pp 181258 – 181292, September 2020.

[5] Feriel Khennouche, Youssef Elmir, Yassine Himeur, Nabil Djebari, Abbes Amira, “Revolutionizing generative pre-traineds: Insights and challenges in deploying ChatGPT and generative chatbots for FAQs”, ScienceDirect Expert Systems with Applications, pp 246, July 2024.

[6] Anup Kumar D. Dhanvijay, Mohammed Jaffer Pinjar, Nitin Dhokane, Smita R. Sorte, Amita Kumari, Himel Monda, "Performance of Large Language Models (ChatGPT, Bing Search, and Google Bard) in Solving Case Vignettes in Physiology", DOI: 10.7759/cureus.42972, April 2023.

[7] Muhammad Usman Hadi, qasem al tashi, Rizwan Qureshi, Abbas Shah, amgad muneer, Muhammad Irfan, Anas Zafar, Muhammad Bilal Shaikh, Naveed Akhtar, Jia

Wu, Seyedali Mirjalili, Mubarak Shah., "Large Language Models: A Comprehensive Survey of its Applications, Challenges, Limitations, and Future Prospects", TechRxiv, DOI: 10.36227/techrxiv.23589741.v4, November 2023.

[8] Dodi Setiawan, Emilia Ayu Dewi Karuniawati, Saksia Imelda Janty, "Peran Chat Gpt (Generative Pre-Training Transformer) Dalam Implementasi Ditinjau Dari Dataset", INNOVATIVE: Journal Of Social Science Research, pp 9527-9539, 2023.

[9] Imtiaz Ahmed, Ayon Roy, Mashrafi Kajol, Uzma Hasan, Partha Protim Datta, Md. Rokonuzzaman Reza, "ChatGPT vs. Bard: A Comparative Study", Authorea DOI: 10.22541/au.168923529.98827844/v1, July 2023.

[10] Ethan Waisberg, Joshua Ong, Mouayad Masalkhi, Nasif Zaman, Prithul Sarker, Andrew G. Lee, Alireza Tavakkoli, "Google's AI chatbot "Bard": a side-by-side comparison with ChatGPT and its utilization in ophthalmology", DOI: 10.1038/s41433-023-02760-0, September 2023.

[11] Jurgen Rudolph, Shannon Tan, Samson Tan, "War of the chatbots: Bard, Bing Chat, ChatGPT, Ernie and beyond. The new AI gold rush and its impact on higher education", Journal of Applied Learning & Teaching Vol.6 No.1, DOI: 0.37074/jalt.2023.6.1.23, 2023.

[12] Trautman, Lawrence J. and Voss, W. Gregory and Shackelford, Scott J., "How We Learned to Stop Worrying and Love AI: Analyzing the Rapid Evolution of Generative Pre-Trained Transformer (GPT) and its Impacts on Law", Business, and Society (July 20, 2023). Available at SSRN: https://ssrn.com/abstract=4516154 or http://dx.doi.org/10.2139/ssrn.4516154.

[13] Andreas Dengel, Rupert Gehrlein, David Fernes, Sebastian Görlich, Jonas Maurer, Hai Hoang Pham, Gabriel Großmann and Niklas Dietrich genannt Eisermann, "Qualitative Research Methods for Large Language Models: Conducting Semi-Structured Interviews with ChatGPT and BARD on Computer Science Education", Informatics, DOI: 10.3390/informatics10040078, October 2023.

[14] Guru99, https://www.guru99.com/pdf/python-interview-questions-answers.pdf, diakses Januari 2024.

[15] K. R. Srinath, "Python – The Fastest Growing Programming Language", International Research Journal of Engineering and Technology (IRJET), Volume: 04 Issue: 12, Desember 2017.

[16] F.Noorbehbahani, A.A. Kardan, "The automatic assessment of free text answers using a modified BLEU algorithm", Computers & Education, Volume 56, Issue 2, pp 337-345, Februari 2011.

[17] Y. Heryanto, A. Triayudi., "Evaluating Text Quality of GPT Engine Davinci-003 and GPT Engine Davinci Generation Using BLEU Score", SAGA: Journal of Technology and Information Systems , pp 121-129, November 2023.

[18] Ishith Seth, Bryan Lim, Yi Xie, Jevan Cevik, Warren M. Rozen, Richard J. Ross, Mathew Lee, "Comparing the Efficacy of Large Language Models ChatGPT, BARD, and Bing AI in Providing Information on Rhinoplasty: An Observational Study", Aesthetic Surgery Journal Open Forum, Volume 5, 2023, September 2023.

[19] Peiyi Wang, Lei Li, Liang Chen, Feifan Song, Binghuai Lin, Yunbo Cao, Tianyu Liu, Zhifang Sui, "Making Large Language Models Better Reasoners with Alignment", DOI: 10.48550/arXiv.2309.02144, September 2023.

[20] Vagelis Plevris, George Papazafeiropoulos, George Papazafeiropoulos, "Chatbots Put to the Test in Math and Logic Problems: A Comparison and Assessment of ChatGPT-3.5, ChatGPT-4, and Google Bard", AI 2023, 4(4), pp 949-969; DOI: 10.3390/ai4040048, October 2023.