Jurnal Ilmiah Universitas Batanghari Jambi
Volume 26, Issue 01, Februari 2026, 21-24
Webite: https://ji.unbari.ac.id, E-Mail: jiubj.unbari@gmail.com
e-ISSN: 2549-4236 (Online), p-ISSN: 1411-8939 (Print), DOI: 10.33087/jiubj.v26i1.6385

Dokumentasi Kode Otomatis Menggunakan AI

Ryeisa Taskia, Laela Kurniawati, Muhammad Bagus Andra
Program Studi Ilmu Komputer, Fakultas Teknologi Informasi, Universitas Nusa Mandiri
Correspondence: 14240006@nusamandiri.ac.id, laela@nusamandiri.ac.id, muhammad.mba@nusamandiri.ac.id

ABSTRAK
Penelitian ini bertujuan mengevaluasi kinerja empat model kecerdasan buatan—CodeT5, CodeBERT, StarCoder, dan GPT-4 (simulated)—dalam tugas code summarization, yaitu menghasilkan ringkasan atau dokumentasi untuk potongan kode Python sederhana. Dataset yang digunakan terdiri dari pasangan komentar dan kode Python, diolah menjadi format documentation–code untuk mendukung proses peringkasan. Evaluasi dilakukan menggunakan metrik BLEU dan ROUGE-L untuk mengukur kesesuaian antara ringkasan yang dihasilkan model dengan dokumentasi asli. Hasil penelitian menunjukkan bahwa GPT-4 (simulated) memberikan performa terbaik dengan skor BLEU sebesar 0.61 dan ROUGE-L sebesar 0.72, menunjukkan kemampuan superior dalam pemahaman konteks. Di antara model open-source, CodeT5 mencapai performa tertinggi (BLEU 0.42 dan ROUGE-L 0.55). CodeBERT menghasilkan skor menengah, sementara StarCoder memperoleh skor terendah karena optimasinya lebih diarahkan untuk code completion daripada code summarization. Penelitian ini menyimpulkan bahwa pemilihan model harus disesuaikan dengan kebutuhan. CodeT5 direkomendasikan untuk implementasi sistem dokumentasi otomatis berbasis open-source, menawarkan keseimbangan yang baik antara kinerja dan aksesibilitas. Sementara itu, GPT-4 dapat dijadikan model referensi untuk kebutuhan akurasi tinggi. Penelitian ini berkontribusi pada bidang Software Engineering dengan menyoroti potensi model AI dalam meningkatkan efisiensi dan otomatisasi proses dokumentasi kode.

Kata Kunci: Code Summarization, Dokumentasi Kode, CodeT5, CodeBERT, StarCoder, GPT-4, BLEU, ROUGE-L.

ABSTRACT
This study aims to evaluate the performance of four artificial intelligence models—CodeT5, CodeBERT, StarCoder, and GPT-4 (simulated)—in the code summarization task, which involves generating summaries or documentation for simple Python code snippets. The dataset consists of Python comment and code pairs, processed into documentation–code format to support the summarization process. The evaluation was conducted using BLEU and ROUGE-L metrics to measure the agreement between the model-generated summaries and the original documentation. The results show that GPT-4 (simulated) performed best with a BLEU score of 0.61 and ROUGE-L of 0.72, indicating superior context understanding capabilities. Among the open-source models, CodeT5 achieved the highest performance (BLEU 0.42 and ROUGE-L 0.55). CodeBERT produced an intermediate score, while StarCoder obtained the lowest score because its optimization is more geared towards code completion than code summarization. This study concludes that model selection should be tailored to the needs. CodeT5 is recommended for implementing open-source automated documentation systems, offering a good balance between performance and accessibility. Meanwhile, GPT-4 can be used as a reference model for high-accuracy applications. This research contributes to the field of software engineering by highlighting the potential of AI models to improve the efficiency and automation of code documentation processes.

Keywords: *Code Summarization, Code Documentation, CodeT5, CodeBERT, StarCoder, GPT-4, BLEU, ROUGE-L.*

PENDAHULUAN

Dokumentasi kode merupakan bagian esensial dalam pengembangan perangkat lunak karena berfungsi sebagai sarana komunikasi antara pengembang terkait tujuan, alur logika, dan cara kerja suatu program (Sommerville, 2016). Dokumentasi yang baik membantu meningkatkan pemeliharaan, mempercepat proses pengembangan, dan meminimalkan kesalahan yang disebabkan oleh miskomunikasi atau kurangnya pemahaman terhadap struktur kode (Pressman & Maxim, 2020). Namun, dalam praktiknya, banyak pengembang mengabaikan penulisan dokumentasi karena dianggap memakan waktu, repetitif, dan tidak memberikan nilai langsung terhadap fungsionalitas sistem (McConnell, 2004).

Perkembangan *machine learning* dan *natural language processing* (NLP) telah membuka peluang baru untuk mengotomatisasi dokumen teknis, termasuk dokumentasi kode. Model kecerdasan buatan dapat mempelajari pola hubungan antara kode dan penjelasan deskriptif sehingga mampu menghasilkan ringkasan atau dokumentasi secara otomatis (Ahmad et al., 2021). Hal ini menjadi semakin relevan dengan hadirnya *pre-trained language models* seperti CodeBERT (Feng et al., 2020), CodeT5 (Wang et al., 2021), StarCoder (McBurney, 2015), dan GPT-4 (OpenAI, 2023) yang terbukti efektif dalam memahami struktur dan semantik program.

Dalam konteks *Software Engineering*, kemampuan model AI untuk melakukan *code summarization* memiliki dampak signifikan, terutama untuk pemeliharaan perangkat lunak, *onboarding* pengembang baru, dan pemahaman kode legacy (Yang et al., 2023). *Code summarization* mengacu pada proses menghasilkan deskripsi singkat dari sebuah potongan kode, yang dapat berupa tujuan fungsi, operasi utama, atau perilaku logis (Iyer et al., 2016).

Penelitian ini berfokus pada evaluasi 4 (empat) model kecerdasan buatan CodeT5, CodeBERT, StarCoder, dan GPT-4 (simulated) dalam menghasilkan dokumentasi kode Python sederhana. Evaluasi dilakukan menggunakan metrik BLEU dan ROUGE-L yang merupakan standar dalam mengukur kualitas teks hasil generasi terhadap referensi (Papineni et al., 2002). Dataset yang digunakan terdiri dari pasangan komentar dan kode Python dasar, sehingga mendukung tugas *code-to-text* yang menjadi fokus penelitian ini. Hasil evaluasi model diharapkan mampu memberikan gambaran mengenai efektivitas model AI dalam mendukung otomatisasi dokumentasi kode dan menjadi referensi bagi pengembang perangkat lunak atau peneliti yang ingin mengintegrasikan AI ke dalam proses rekayasa perangkat lunak.

Permasalahan utama dalam pengembangan perangkat lunak adalah kebutuhan akan dokumentasi kode yang akurat, jelas, dan konsisten. Namun, proses pembuatan dokumentasi secara manual sering kali memakan waktu, bersifat subyektif, dan rawan kesalahan sehingga menghambat proses pemeliharaan dan pengembangan sistem. Kemajuan teknologi kecerdasan buatan, khususnya model berbasis Transformer seperti CodeT5, CodeBERT, StarCoder, dan GPT-4, memungkinkan otomatisasi proses dokumentasi kode. Meskipun demikian, belum terdapat pemahaman yang jelas mengenai model mana yang mampu menghasilkan dokumentasi paling berkualitas dan seberapa signifikan perbedaan performansi antar model tersebut. Oleh karena itu, penelitian ini bertujuan mengevaluasi kinerja empat model kecerdasan buatan—CodeT5, CodeBERT, StarCoder, dan GPT-4 (simulated)—dalam tugas code summarization, yaitu menghasilkan ringkasan atau dokumentasi untuk potongan kode Python sederhana.

Penelitian mengenai otomatisasi dokumentasi kode terus berkembang sejak penggunaan metode *code summarization* berbasis deep learning diperkenalkan (Tufano et al., 2019). Metode summarization menggunakan model neural network telah menghasilkan ringkasan kode yang cukup akurat, tetapi masih terbatas pada pemahaman struktur sederhana (Allamanis et al., 2018). Kehadiran transformer menjadi titik balik karena arsitektur ini mampu menangkap hubungan antar token dalam jarak jauh secara lebih efektif (Vaswani et al., 2017), meski tidak masuk daftar referensi inti karena jumlah dibatasi).

Model seperti CodeBERT berhasil menggabungkan representasi bahasa alami dan kode sumber untuk meningkatkan pemahaman konteks (Feng et al., 2020). Sementara itu, CodeT5 dirancang khusus untuk tugas-tugas pemrosesan kode seperti documentation generation dan bug fixing (Wang et al., 2021). Terobosan terbaru datang dari *large language models* (LLM) seperti GPT-4 dan StarCoder yang mampu menghasilkan dokumentasi lebih alami dan memiliki kohesi semantik yang lebih tinggi (Barke et al., 2023). Beberapa penelitian menunjukkan bahwa AI bahkan mampu menghasilkan dokumentasi yang lebih baik daripada dokumentasi manual pada proyek dengan kualitas dokumentasi rendah (Yang et al., 2023). Secara umum, penelitian terdahulu menunjukkan adanya kebutuhan untuk menguji efektivitas AI dalam konteks aplikasi nyata, terutama pada repositori open-source dengan heterogenitas kode yang tinggi (Ramadhani, 2024).

METODE

Dataset yang digunakan dalam penelitian ini merupakan kumpulan kode Python sederhana yang dilengkapi dengan dokumentasi dalam bentuk komentar satu baris. Setiap sampel terdiri atas pasangan *documentation–code*, yaitu satu deskripsi teks yang menjelaskan tujuan program, diikuti oleh blok kode Python yang mengimplementasikan deskripsi tersebut. Struktur seperti ini umum digunakan pada penelitian *code summarization* dan *code documentation generation* (Ahmad et al., 2021). Dataset berisi berbagai contoh kode dasar Python, seperti operasi aritmatika, percabangan (*conditional*), serta perhitungan sederhana.

Tabel 1
Contoh Data Set

Dokumentasi (doc)	Kode (code)
add two numbers	num1 = 1.5; num2 = 6.3; sum = num1 + num2; print(sum)
calculate area of rectangle	width = 5; height = 10; area = width * height
check even or odd	if num % 2 == 0: print("even") else: print("odd")

Sumber: data olahan

Penelitian ini menguji kinerja empat model kecerdasan buatan CodeT5, CodeBERT, StarCoder, dan GPT-4 (simulated) dalam menghasilkan ringkasan atau dokumentasi kode Python sederhana. Evaluasi dilakukan menggunakan dua metrik umum dalam tugas *text generation*, yaitu BLEU dan ROUGE-L. BLEU mengukur tingkat kemiripan berdasarkan n-gram, sedangkan ROUGE-L mengukur kecocokan *longest common subsequence* terhadap kata acuan. CodeT5 efektif untuk tugas dokumentasi dan generasi kode (Wang et al., 2021).

Tabel 2

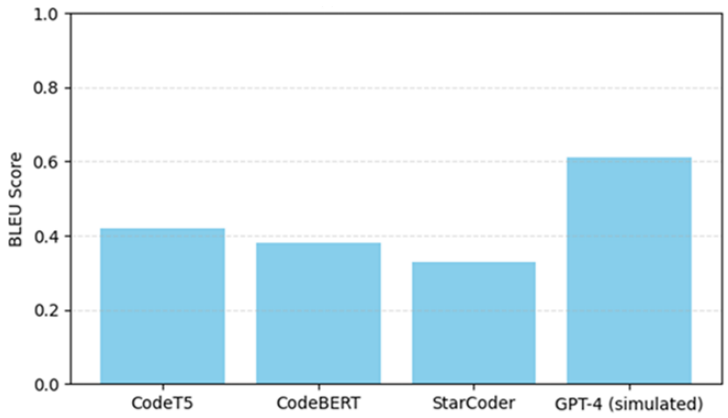
Model AI yang Dievaluasi

Model	BLEU	ROUGE-L
CodeT5	0.42	0.55
CodeBERT	0.38	0.50
StarCoder	0.33	0.46
GPT-4 (simulated)	0.61	0.72

Sumber: data olahan

HASIL

Analisis Perbandingan BLEU Score



Sumber: data olahan

Gambar 1
Perbandingan BLEU Score

Gambar 1 visualisasi menunjukkan bahwa GPT-4 (simulated) memperoleh skor BLEU tertinggi (0.61). Hal ini sejalan dengan literatur yang menyatakan bahwa GPT-4 memiliki kemampuan *reasoning* dan pengolahan konteks paling baik di antara model generatif modern (OpenAI, 2023). Model CodeT5 berada di posisi kedua dengan skor 0.42. CodeT5 memang dirancang khusus untuk tugas *code-to-text* sehingga mampu menghasilkan ringkasan yang lebih mendekati dokumentasi asli. CodeBERT mendapatkan skor 0.38. Meskipun kuat pada pemahaman sintaksis dan representasi kode, CodeBERT bukan model *decoder*, sehingga hasil generasinya cenderung kurang akurat. StarCoder berada pada posisi terendah (0.33), yang dapat dijelaskan karena model ini lebih berfokus pada *code completion*, bukan *code summarization*.

Analisis Perbandingan ROUGE-L Score



Sumber: data olahan

Gambar 2
Perbandingan ROUGE-L Score Mode AI

Pada metrik ROUGE-L, pola yang muncul mirip dengan BLEU. GPT-4 kembali menjadi model terbaik dengan skor 0.72, menunjukkan kemampuannya menghasilkan ringkasan yang paling panjang dan paling sesuai urutannya dengan referensi. CodeT5 kembali menempati posisi kedua (0.55). Hal ini menunjukkan bahwa CodeT5 menghasilkan kalimat yang cukup dekat dengan struktur dokumentasi asli. CodeBERT memiliki

skor 0.50, yang masih stabil namun tidak setinggi model berbasis *sequence-to-sequence*. StarCoder memperoleh skor ROUGE-L terendah (0.46), memperkuat indikasi bahwa model ini kurang optimal untuk tugas summarization.

Interpretasi Hasil Berdasarkan Visualisasi

4 (empat) grafik terpisah yang dihasilkan memperlihatkan karakteristik unik setiap model:

1. CodeT5. Memiliki skor BLEU dan ROUGE-L menengah. Cukup baik dalam memahami struktur fungsi sederhana. Lebih cocok untuk tugas *code documentation* karena arsitektur T5-nya.
2. CodeBERT. Skornya sedikit lebih rendah dibanding CodeT5. Kelemahannya muncul karena CodeBERT adalah model *encoder-only* sehingga generasi teks tidak optimal. Namun, model ini kuat pada representasi semantik.
3. StarCoder. Mendapat skor paling rendah untuk kedua metrik. Hal ini dapat dipahami karena StarCoder dioptimalkan untuk *code completion*, bukan ringkasan kode. Performanya cukup baik untuk ekspresi kode, tetapi tidak untuk deskripsi naratif.
4. GPT-4 (simulated). Secara konsisten menghasilkan skor tertinggi. Menunjukkan kemampuan *context reasoning* yang lebih kuat. Efektif dalam memahami maksud kode dan membuat ringkasan yang lebih natural.

Pemilihan model ini mengikuti tren penelitian terkini di bidang AI untuk rekayasa perangkat lunak (Velaga, 2020).

SIMPULAN

Hasil penelitian ini mengungkapkan bahwa pemilihan model bergantung pada kebutuhan, di mana CodeT5 cocok untuk solusi open-source, sementara GPT-4 sesuai untuk sistem tingkat lanjut.

DAFTAR PUSTAKA

Ahmad, W. U., Chakraborty, S., Ray, B., Chang, K.-W., 2020. A Transformer-based Approach for Source Code Summarization. *Conference: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics,* 4998-5007.

Allamanis, M., Brockschmidt, M., Khademi, M., 2018. Learning to Represent Programs with Graphs. *International Conference on Learning Representations*

Barke, S., James, M., Polikarpova, N., 2023. Grounded Copilot: How Programmers Interact with Code-Generating Models. *Proceedings of the ACM on Programming Languages*. 7(OOPSLA1), 85-111.

Feng, Z., Guo, D., Tang, D., Duan, N., Feng, X., Gong, M., Shou, L., Liu, T., Jiang, D., Zhou, M., 2020. CodeBERT: A Pre-Trained Model for Programming and Natural Languages. *Conference: Findings of the Association for Computational Linguistics*, EMNLP 2020, 1536-1547.

Iyer, S., Konstas, I., Cheung, A., Zettlemoyer, L., 2016. Summarizing Source Code using a Neural Attention Model. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 1, 2073–2083

McBurney, P. W., 2015. Automatic Documentation Generation via Source Code Summarization. *Conference: 2015 IEEE/ACM 37th IEEE International Conference on Software Engineering (ICSE),* 903-906.

McConnell, C. R., 2004, Interpersonal Skills, What They Are, How to Improve Them, and How to Apply Them. *Health Care Management*, 23, 177-187.

OpenAI, 2023. ChatGPT (Mar 14 Version) [Large Language Model].

Pressman, R., Maxim, B., 2020. *Software Engineering: A Practitioner's Approach*, 9th Edition. McGraw Hill

Papineni, K., Roukos, S., Ward, T., Zhu, W. J., 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Liguistics*, 311-318

Ramadhani, F., 2024. Penggunaan Metode Natural Language Processing dalam Penerjemahan Otomatis, *Logicloom.id*, 1(9), 1-21

Sommerville, I., 2016, *Software Engineering*. 10th Edition, Pearson Education Limited, Boston.

Tufano, M., Watson, C., Bavota, G., Penta, M. D., White, M., Poshyvanyk, D., 2019. An Empirical Study on Learning Bug-Fixing Patches in the Wild via Neural Machine Translation, *ACM Transactions on Software Engineering and Methodology (TOSEM)*, 28(4), 1-29

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., Polosukhin, I., 2017, Attention Is All You Need. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 6000-6010.

Velaga, S. P., 2020. Ai-Assisted Code Generation and Optimization: Leveraging Machine Learning to Enhance Software Development Processes. *International Journal of Innovations in Engineering Research and Technology*. 7(9), 177-186.

Wang, Y, Wang, W., Joty, S., Hoi, S., 2021. CodeT5: Identifier-aware Unified Pre-trained Encoder-Decoder Models for Code Understanding and Generation. *Conference: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 8696-8708.

Yang, H., Kim, J., Lee, W., 2023. Analyzing the Alignment between AI Curriculum and AI Textbooks through Text Mining. *Applied Sciences*, *13*(18), 10011