POTERS
PROCEEDINGS OF TECHNOLOGY, ENGINEERING AND COMPUTERS

POTERS: Proceedings of Technology, Engineering and Computers

Vol. 2, No. 1, pp. 40 - 47, 2026
p-ISSN. 3109 - 127X, e-ISSN. 3090 - 3602

EVALUASI KELAYAKAN JAWABAN *LARGE LANGUAGE MODELS* (LLMs) PADA BIDANG KESEHATAN MENGGUNAKAN METODE *BERTSCORE*

Listia Baene[1], Dram Renaldi[2*], Edy[3]
[1, 2] Teknik Perangkat Lunak, Fakultas Sains dan Teknologi, Universitas Buddhi Dharma
*Corresponding Author, email: dram.renaldi@ubd.ac.id

ABSTRAK

Perkembangan pesat *Large Language Models* (LLMs) seperti ChatGPT menawarkan potensi besar dalam sistem tanya jawab medis otomatis, namun evaluasi akurasi jawaban sering kali tidak mampu jika hanya menggunakan metrik leksikal konvensional. Oleh karena itu, diperlukan pendekatan evaluasi berbasis semantik menggunakan BERTScore yang mampu menangkap kedekatan makna kontekstual secara lebih mendalam dibandingkan dengan kecocokan kata. Penelitian ini bertujuan untuk mengukur tingkat kelayakan jawaban ChatGPT di bidang kesehatan dengan membandingkannya terhadap jawaban referensi pakar menggunakan algoritma BERTScore, serta mengembangkan instrumen pengujian otomatis yang efisien berbasis Python pada platform *cloud* Google Colaboratory. Metode yang digunakan meliputi eksperimen komputasional dengan memproses dataset 100 pertanyaan medis yang mencakup 10 kategori spesifik. Pustaka *Hugging Face Transformers* digunakan untuk menghitung nilai *Precision* , *Recall* , dan *F1-Score* , sementara validasi sistem dilakukan oleh tiga pakar bidang perangkat lunak dan pemrograman. Analisis data menunjukkan skor *F1* rata-rata yang tinggi pada kisaran 0,84–0,89. Namun, persentase jawaban berkualitas "Bagus" sangat bervariasi, di mana kategori anatomi dan nutrisi mencapai kinerja terbaik (80–90%), sedangkan bidang Farmakologi serta Etika kedokteran memiliki akurasi terendah (40–50%). Validasi pakar memberikan skor rata-rata 82,6%, konfirmasi alat. Implementasi BERTScore terbukti efektif sebagai metode evaluasi semantik otomatis untuk LLMs kesehatan. Dapat disimpulkan bahwa meskipun model bahasa mampu menghasilkan teks yang koheren secara semantik, akurasi substansi pada topik medis yang kompleks dan berisiko tinggi masih memerlukan verifikasi lebih lanjut, menjadikan instrumen ini penting untuk pengembangan AI medis yang aman.

Kata kunci: *BERTScore, ChatGPT, Google Colab, LLMs, Python*

I. PENDAHULUAN

Large Language Models (LLMs) merupakan model bahasa berukuran besar berbasis arsitektur transformator yang dibor menggunakan miliaran kata dari berbagai sumber teks untuk memahami konteks dan menghasilkan bahasa alami berkualitas tinggi (Fh Tondowala et al., 2023). LLMs merupakan evolusi dari pemrosesan bahasa alami berdasarkan pembelajaran mendalam yang memungkinkan pemrosesan teks paralel dan penangkapan hubungan semantik antar kata untuk mendukung berbagai tugas klasifikasi dan generasi teks dengan kinerja

Published by Universitas Buddhi Dharma
Available online at https://jurnal.ubd.ac.id/index.php/poters/index
Kota Tangerang, 07 Januari 2026.

Vol. 2, No. 1, pp. 40 - 47, 2026
p-ISSN. 3109 - 127X, e-ISSN. 3090 - 3602

tinggi (Fahrezy et al., 2025). Dengan kemampuan memahami konteks secara mendalam dan menghasilkan teks yang koheren, LLMs menjadi salah satu pencapaian paling signifikan dalam perkembangan teknologi NLP. Model ini menjadi landasan utama sistem AI generatif dan sistem tanya jawab otomatis di berbagai bidang, termasuk pendidikan, bisnis, dan khususnya kesehatan. Salah satu LLM yang saat ini populer adalah ChatGPT. ChatGPT merupakan implementasi LLM yang dikembangkan oleh OpenAI menggunakan arsitektur Generative Pre-Trained Transformer (GPT). Model ini dilatih dengan teknik pembelajaran tanpa pengawasan dan Reinforcement Learning from Human Feedback (RLHF), sehingga mampu memahami konteks dialog dan menghasilkan respons yang alami, logistik, serta relevan pada berbagai domain termasuk penjelasan medis (Felisa Widjaya et al., 2024). Penelitian-penelitian terkini menegaskan bahwa model bahasa besar (LLMs) dan model berbasis BERT memainkan peran sentral dalam evaluasi kelayakan jawaban pada sistem *Question Answering* (QA) di domain kesehatan. Integrasi LLM dengan *framework* seperti LangChain, teknik *chunking*, *embeddings*, dan *Retrieval-Augmented Generation* (RAG) telah menunjukkan kemampuan menghasilkan respons yang akurat, meskipun masih menghadapi tantangan dalam pengelolaan konteks panjang dan ambiguitas makna (Kurniawan & Triloka, 2025). Sejumlah penelitian menyoroti keunggulan metrik semantik seperti BERTScore dibandingkan metrik leksikal seperti ROUGE, karena mampu menangkap makna meskipun struktur kalimat berbeda (Hanum et al., 2024; Lubis et al., 2024). Pendekatan *prompt engineering* pada QA medis berbasis RAG juga terbukti meningkatkan kualitas jawaban semantik melalui *some-shot prompting* (Haromain et al., 2025). Di sisi lain, implementasi chatbot berbasis LLM pada platform WhatsApp maupun Telegram menampilkan relevansi arsitektur kerja *retriever–generator* dalam domain kesehatan, dengan evaluasi yang memadukan metrik tujuan (misalnya akurasi dan BERTScore) serta subyektif seperti *User Experience Questionnaire* (Hasbi et al., 2025). Temuan tambahan menampilkan bahwa penggunaan embedding BERT pada sistem kesehatan pencernaan, klasifikasi teks, dan deteksi spam menghasilkan skor F1 dan *kesetiaan semantik* yang tinggi (Amin et al., 2024; Samudra et al., 2025), memperkuat relevansi BERT sebagai representasi semantik untuk penilaian kualitas respon.

Published by Universitas Buddhi Dharma
Available online at https://jurnal.ubd.ac.id/index.php/poters/index
Kota Tangerang, 07 Januari 2026.

POTERS
PROCEEDINGS OF TECHNOLOGY, ENGINEERING AND COMPUTERS
POTERS: Proceedings of Technology, Engineering and Computers
Vol. 2, No. 1, pp. 40 - 47, 2026
p-ISSN. 3109 - 127X, e-ISSN. 3090 - 3602

Penelitian lain menunjukkan efektivitas fine-tuning model seperti BloomZ dan LLaMA2 untuk bahasa sumber daya rendah dengan evaluasi BERTScore dan ROUGE-L terhadap jawaban medis (Bui et al., 2025) serta memperkenalkan kerangka multi-metrik untuk menilai utilitas klinis LLM (Builtjes et al., 2025). Secara umum, tren penelitian mengarah pada penggabungan evaluasi berbasis kesamaan semantik menggunakan BERTScore, kesetiaan konten medis, serta pemaparan terhadap panduan klinis sebagai tolok ukur baru kelayakan jawaban LLM di domain kesehatan (Abroms et al., 2025; Dzaky et al., 2024; Kuligin et al., 2025; Shieh et al., 2024).

Penelitian ini menggunakan BERTScore sebagai metode evaluasi yang memanfaatkan representasi semantik dari model BERT untuk menghitung kesesuaian makna antara jawaban sistem dan referensi. Pendekatan ini menilai kedekatan makna melalui embedding sehingga lebih kontekstual dibandingkan kesesuaian kata permukaan (Lubis et al., 2024). Implementasi BERTScore dilakukan menggunakan Python di platform Google Colaboratory (Colab), yang menyediakan lingkungan komputasi berbasis cloud dengan pustaka seperti Hugging Face Transformers. Melalui dukungan GPU/TPU dan integrasi Google Drive, evaluasi dapat dilakukan secara efisien, terukur, dan mudah direplikasi (Ahmad Tohir et al., 2024; Octarina et al., 2025). Dengan demikian, evaluasi kelayakan jawaban ChatGPT di bidang kesehatan dapat dilakukan secara sistematis dan akurat menggunakan BERTScore. Pendekatan ini mendukung pengembangan kerangka evaluasi semantik berbasis AI yang lebih aman, andal, dan terpercaya dalam penyebaran informasi kesehatan.

II. METODOLOGI

Prosedur penelitian dimulai dengan mengunggah berkas CSV berisi data pertanyaan, jawaban ChatGPT, dan jawaban referensi ke dalam *Data Frame* menggunakan *pandas*. Selanjutnya dilakukan instalasi pustaka *bert-score* dan *matplotlib*, disusul pemanggilan fungsi BERTScore untuk menghitung nilai *Precision*, *Recall*, dan *F1Score* tiap pasangan jawaban. Nilai *F1Score* kemudian digunakan untuk memberi label kualitas (“Bagus” atau “Kurang”) berdasarkan ambang batas tertentu, mengidentifikasi lima jawaban dengan skor

Published by Universitas Buddhi Dharma
Available online at https://jurnal.ubd.ac.id/index.php/poters/index
Kota Tangerang, 07 Januari 2026.

POTERS
PROCEEDINGS OF TECHNOLOGY, ENGINEERING AND COMPUTERS
POTERS: Proceedings of Technology, Engineering and Computers
Vol. 2, No. 1, pp. 40 - 47, 2026
p-ISSN. 3109 - 127X, e-ISSN. 3090 - 3602

terendah, menghitung persentase kualitas keseluruhan, dan memvisualisasikan hasil dalam grafik serta *diagram lingkaran*. Hasil evaluasi akhir disimpan ke berkas CSV baru sebagai analisis dokumentasi yang siap diunduh. Dalam penelitian ini prosedur penggunaan software khususnya platform Google Colab dapat diilustrasikan dalam flowchart berikut,

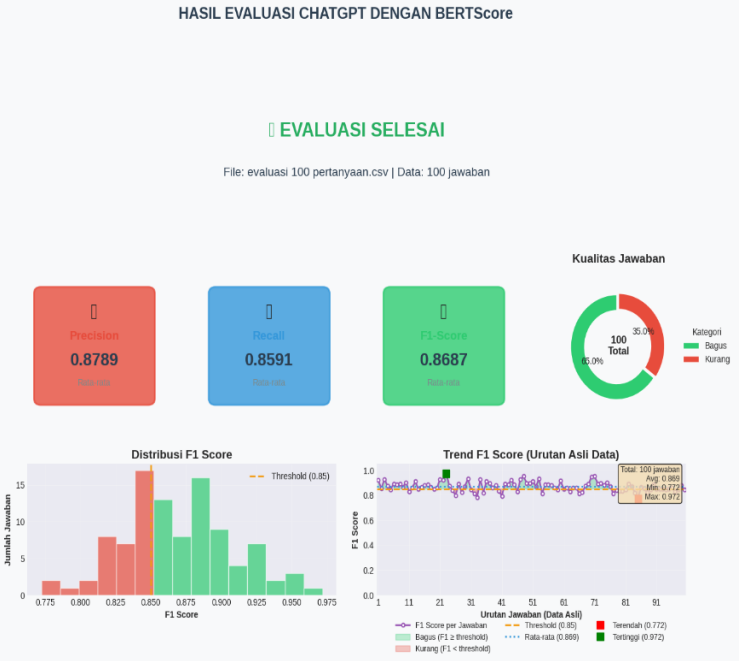
Gambar 1. Flowchart Prosedur Penggunaan Software

III. HASIL DAN PEMBAHASAN

Bagian ini menjelaskan implementasi teknis metode evaluasi kinerja *Large Language Models* (LLMs) di bidang kesehatan melalui instrumen pengembangan pengujian otomatis berbasis Python pada lingkungan *cloud* Google Colab. Aplikasi ini memproses dataset berformat CSV yang berisi pertanyaan medis, jawaban referensi, dan jawaban prediksi model untuk menghitung metrik BERTScore secara otomatis, sehingga analisis kelayakan jawaban LLM dapat dilakukan secara efisien dan terstruktur. Berikut ini adalah hasil dari 100 pertanyaan pada bidang Kesehatan:

Published by Universitas Buddhi Dharma
Available online at https://jurnal.ubd.ac.id/index.php/poters/index
Kota Tangerang, 07 Januari 2026.

POTERS
PROCEEDINGS OF TECHNOLOGY, ENGINEERING AND COMPUTERS
POTERS: Proceedings of Technology, Engineering and Computers
Vol. 2, No. 1, pp. 40 - 47, 2026
p-ISSN. 3109 - 127X, e-ISSN. 3090 - 3602



Gambar 2. Hasil Evaluasi BERTScore

Selain itu penelitian ini dilakukan uji kelayakan jawaban ChatGPT terhadap jawaban referensi yang terbagi menjadi 10 kategori bidang Kesehatan dan setiap bidangnya memiliki 10 pertanyaan. Berikut ini hasil dari uji kelayakan jawaban ChatGPT:

Tabel 1. Hasil Evaluasi BERTScore Tiap Kategori

No	Kategori	Hasil perhitungan bertscore			
		Precision	recall	F1score	Kualitas jawaban(bagus)
1	Anatomi dan Fisiologi Manusia	0.8930	0.8796	0.8861	90%
2	Penyakit dan Gangguan Umum	0.8808	0.8489	0.8645	70%
3	Mikro biologi dan Infeksi	0.8878	0.8821	0.8847	70%
4	Farmakologi	0.8636	0.8514	0.8571	50%
5	Gizi dan Nutrisi	0.8986	0.8778	0.8878	80%
6	Kesehatan Mental	0.8855	0.8747	0.8799	80%
7	Kesehatan Reproduksi	0.8673	0.8525	0.8597	60%
8	Kesehatan anak	0.8812	0.8612	0.8710	60%
9	Kesehatan Masyarakat dan Pencegahan	0.8662	0.8282	0.8467	50%
10	Etika dan Hukum dalam Kedokteran	0.8647	0.8345	0.8492	40%

POTERS
PROCEEDINGS OF TECHNOLOGY, ENGINEERING AND COMPUTERS
POTERS: Proceedings of Technology, Engineering and Computers
Vol. 2, No. 1, pp. 40 - 47, 2026
p-ISSN. 3109 - 127X, e-ISSN. 3090 - 3602

IV. SIMPULAN

Penelitian ini mengembangkan instrumen pengujian otomatis berbasis Python di Google Colab untuk memancarkan kinerja LLM bidang kesehatan menggunakan BERTScore terhadap dataset pertanyaan medis berformat CSV. Hasil pengujian 100 pertanyaan dalam 10 kategori menunjukkan skor presisi, recall, dan F1 yang tinggi (sekitar 0,84–0,89), dengan persentase kelayakan jawaban "bagus" antara 40–90%; kategori Anatomi dan Fisiologi Manusia serta Gizi dan Nutrisi mencapai kualitas jawaban terbaik (80–90%), sedangkan Farmakologi, Kesehatan Masyarakat dan Pencegahan, serta Etika dan Hukum dalam Kedokteran berada pada kisaran terendah (40–50%). Hal ini menunjukkan bahwa kedekatan semantik tidak selalu berbanding lurus dengan kualitas substansi pada domain yang lebih kompleks. Validasi tiga pakar pada aspek SQA, UI, dan struktur pengkodean Python menghasilkan skor 80%, 80%, dan 88% dengan rata-rata 82,6%, menandakan bahwa implementasi BERTScore berjalan baik dan cukup reliabel sebagai alat evaluasi jawaban LLM bidang kesehatan.

DAFTAR PUSTAKA

Abroms, L. C., Yousefi, A., Wysota, C. N., Wu, T. C. & Broniatowski, D. A. (2025). Assessing the Adherence of ChatGPT Chatbots to Public Health Guidelines for Smoking Cessation: Content Analysis. *Journal of Medical Internet Research*, *27*. https://doi.org/10.2196/66896

Ahmad Tohir, F., Irawan, B., Bahtiar, A., Kunci-Analisis Sentimen, K., Pengguna, U. & Naïve Bayes, A. (2024). *Analisis Sentimen Aplikasi ChatGPT Mobile Menggunakan Agoritma Naïve Bayes*.

Amin, M. B. M., Hakim, G., Maulana, M. T., Alwan, M. F., Anggraheni, H. S., Naufal, M. J. & Yudistira, N. (2024). Deteksi Spam Berbahasa Indonesia Berbasis Teks Menggunakan Model Bert. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, *11*(6), 1291–1302. https://doi.org/10.25126/jtiik.2024118121

Bui, N., Nguyen, G., Nguyen, N., Vo, B., Vo, L., Huynh, T., Tang, A., Tran, V. N., Huynh, T., Nguyen, H. Q. & Dinh, M. (2025). Fine-tuning large language models for improved health communication in low-resource languages. *Computer Methods and Programs in Biomedicine*, *263*. https://doi.org/10.1016/j.cmpb.2025.108655

Published by Universitas Buddhi Dharma
Available online at https://jurnal.ubd.ac.id/index.php/poters/index
Kota Tangerang, 07 Januari 2026.

Vol. 2, No. 1, pp. 40 - 47, 2026
p-ISSN. 3109 - 127X, e-ISSN. 3090 - 3602

Builtjes, L., Bosma, J., Prokop, M., Van Ginneken, B. & Hering, A. (2025). Leveraging open-source large language models for clinical information extraction in resource-constrained settings. *JAMIA Open*, *8*(5). https://doi.org/10.1093/jamiaopen/ooaf109

Dzaky, A. A., Zeniarja, J., Supriyanto, C., Fajar Shidik, G., Paramita, C., Subhiyakto, E. R. & Rakasiwi, S. (2024). Optimization Chatbot Services Based on DNN-Bert for Mental Health of University Students. In *Journal of Applied Informatics and Computing (JAIC)* (Vol. 8, Issue 1). http://jurnal.polibatam.ac.id/index.php/JAIC

Fahrezy, I., Safaat Harahap, N., Wulandari, F. & Agustian, S. (2025). 2032-Article Text-7429-1-10-20250617. *Buletin Of Information Technology (BIT)*, 173–183.

Felisa Widjaya, A., Reva Utamandarya, B., Angelica Dharmo, V. & Yola Febrianti, L. (2024). Dampak Penggunaan ChatGPT terhadap Proses Pembelajaran Mahasiswa Jurusan Business Engineering di BINUS ASO. In *Jurnal Ilmiah Sains dan Teknologi* (Vol. 2, Issue 7).

Fh Tondowala, S., A Ruagadi, H., Rynawaty Taaha, Y., Pasambaka, Y. & A Tabondo, Y. (2023). *2023) (Hal, 13-24) E-ISSN : xxxx-xxxx Pemanfaatan ChatGPT dan Model Bahasa Besar Lainnya dalam Meningkatkan Peran Dosen Perguruan Tinggi Swasta: Studi Pustaka tentang Pengembangan* (Vol. 3, Issue 2). https://publikasi.unkrit.ac.id/index.php/Pand

Hanum, A. R., Zetha, I. A., Fajrina, J. N., Wulandari, R. A., Putri, S. C., Andina, S. P. & Yudistira, N. (2024). ANALISIS KINERJA ALGORITMA KLASIFIKASI TEKS BERT DALAM MENDETEKSI BERITA HOAKS. *Jurnal Teknologi Informasi Dan Ilmu Komputer (JTIIK)*, *11*(3), 537–546. https://doi.org/10.25126/jtiik938093

Haromain, I., Munir, S., Rahmah, A., Tinggi, S., Terpadu, T., Fikri, N. & Disetujui, D. D. (2025). Analisa Prompt Engineering pada Large Language Model dengan Retrieval-Augmented Generation untuk Informasi Obat dan Vitamin. *Journal Computer Science*, *4*(2).

Hasbi, M. A., Imanda, R. & Fathan Fauzan, M. (2025). Implementasi Chatbot Berbasis Large Language Model Untuk Pencarian Skripsi Mahasiswa

Published by Universitas Buddhi Dharma
Available online at https://jurnal.ubd.ac.id/index.php/poters/index
Kota Tangerang, 07 Januari 2026.

Vol. 2, No. 1, pp. 40 - 47, 2026
p-ISSN. 3109 - 127X, e-ISSN. 3090 - 3602

Terintegrasi dengan Whatsapp. *Arcitech: Journal of Computer Science and Artificial Intelligence*, *5*(1), 148–167. https://doi.org/10.29240/arcitech.v5i1.13974

Kuligin, L., Lammert, J., Ostapenko, A., Bressem, K., Boeker, M. & Tschochohei, M. (2025). Prompt design for medical question answering with Large Language Models. *Machine Learning with Applications*, *22*, 100758. https://doi.org/10.1016/j.mlwa.2025.100758

Kurniawan, D. & Triloka, J. (2025). *Penerapan Teknologi Langchain dan LLM pada Sistem Question Answering Berbasis Chatbot Telegram: Literature Review*.

Lubis, A. T. U. BR., Harahap, N. S., Agustian, S., Irsyad, M. & Afrianty, I. (2024). Question Answering System pada Chatbot Telegram Menggunakan Large Language Models (LLM) dan Langchain (Studi Kasus UU Kesehatan). *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, *4*(3), 955–964. https://doi.org/10.57152/malcom.v4i3.1378

Octarina, S., Puspita, F. M., Yuliza, E. & Indrawati, I. (2025). PENDAMPINGAN PENGGUNAAN GOOGLE COLAB PADA PEMBELAJARAN PYTHON DAN MACHINE LEARNING BAGI DOSEN MATEMATIKA DI PALEMBANG. *Jurnal Pepadu*, *6*(1), 56–66. https://doi.org/10.29303/pepadu.v6i1.6457

Samudra, G., Turmudi Zy, A. & Ermanto. (2025). Implementasi Retrieval Augmented Generation (RAG) Dalam. *JSAI: Journal Scientific and Applied Informatics*, *8*(1). https://doi.org/10.36085

Shieh, A., Tran, B., He, G., Kumar, M., Freed, J. A. & Majety, P. (2024). Assessing ChatGPT 4.0's test performance and clinical diagnostic accuracy on USMLE STEP 2 CK and clinical case reports. *Scientific Reports*, *14*(1). https://doi.org/10.1038/s41598-024-58760-x

Published by Universitas Buddhi Dharma
Available online at https://jurnal.ubd.ac.id/index.php/poters/index
Kota Tangerang, 07 Januari 2026.