BAB V
KESIMPULAN DAN SARAN

A. Kesimpulan

Berdasarkan hasil penelitian yang telah dilakukan mengenai evaluasi performa ChatGPT dan Gemini AI dalam menjawab pertanyaan akademik menggunakan metode *Cosine Similarity*, diperoleh beberapa kesimpulan sebagai berikut:

1. Berdasarkan pengukuran menggunakan Sentence-BERT (SBERT) dan Cosine Similarity, diperoleh rata-rata nilai similarity sebesar 0,7879 untuk ChatGPT dan 0,7824 untuk Gemini AI. Hasil ini menunjukkan bahwa pada dataset yang digunakan dalam penelitian ini, kedua model memiliki kedekatan semantik yang relatif baik terhadap jawaban rujukan (ground truth).
2. Secara deskriptif, ChatGPT memperoleh nilai rata-rata yang sedikit lebih tinggi dibandingkan Gemini AI dengan selisih sebesar 0,0055. Namun, selisih tersebut relatif kecil sehingga menunjukkan bahwa pada dataset penelitian ini performa kedua model cenderung setara dalam menghasilkan jawaban akademik pada materi Data Mining.
3. Berdasarkan kategori pertanyaan, Gemini AI memperoleh nilai rata-rata lebih tinggi pada kategori definisi, sedangkan ChatGPT memperoleh nilai lebih tinggi pada kategori konsep dan analisis. Perbedaan terbesar ditemukan pada kategori analisis, yang menunjukkan bahwa pada dataset penelitian ini ChatGPT memiliki performa yang lebih baik dalam menjawab pertanyaan yang memerlukan penalaran lebih mendalam.
4. ChatGPT menunjukkan performa yang lebih konsisten dibandingkan Gemini AI, ditunjukkan oleh rentang nilai similarity yang lebih kecil. Hal ini mengindikasikan bahwa kualitas jawaban ChatGPT lebih stabil pada berbagai jenis pertanyaan yang digunakan dalam penelitian.
5. Hasil uji Mann–Whitney menghasilkan nilai p-value sebesar 0,651 (> 0,05), sehingga tidak terdapat perbedaan yang signifikan secara statistik antara performa ChatGPT dan Gemini AI. Dengan demikian, kedua model

dapat dianggap memiliki kemampuan yang relatif setara dalam menghasilkan jawaban akademik yang sesuai secara semantik dengan jawaban rujukan.

B. Saran

Berdasarkan hasil penelitian yang telah dilakukan, disarankan bagi penelitian selanjutnya untuk menggunakan jumlah data yang lebih banyak serta variasi pertanyaan yang lebih luas agar hasil evaluasi menjadi lebih komprehensif. Selain itu, penggunaan model *embedding* lain atau metode pengukuran kemiripan yang berbeda dapat dipertimbangkan untuk memperoleh hasil perbandingan yang lebih beragam. Penelitian selanjutnya juga dapat mengeksplorasi aspek lain seperti kualitas struktur jawaban, kedalaman analisis, serta kemampuan model dalam memahami konteks yang lebih kompleks. Di samping itu, penting untuk menjaga konsistensi dalam proses pengambilan data dari model AI agar hasil yang diperoleh lebih valid dan dapat dibandingkan secara adil.

DAFTAR PUSTAKA

Akhmad, D. M. (2020). Pengukuran Kemiripan Semantik Berbasis Graph Pada Gene Ontology. *Computatio : Journal of Computer Science and Information Systems*, *4*(2), 102. https://doi.org/10.24912/computatio.v4i2.9678

Amalia, E. L., Jumadi, A. J., Mashudi, I. A., & Wibowo, D. W. (2021). Analisis Metode Cosine Similarity Pada Aplikasi Ujian Online Otomatis (Studi Kasus JTI POLINEMA). *Jurnal Teknologi Informasi dan Ilmu Komputer*, *8*(2), 343–348. https://doi.org/10.25126/jtiik.2021824356

Balaka, M. Y. (2022). *Metode Penelitian Kuantitatif, Kualitatif, Dan R&D. Bandung: CV Alfabeta*. *Metodologi Penelitian Pendidikan Kualitatif* (Vol. 1).

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). Step-by-step data mining guide.

Dehaene, H., Neve, J. De, & Rosseel, Y. (2021). A Wilcoxon – Mann – Whitney Test for Latent Variables. *frontiers in psychology*, *12*(November). https://doi.org/10.3389/fpsyg.2021.754898

Estremera, M. L., Ann, M., & Sarmiento, M. (2024). Content Validity and Reliability of Questionnaires : Trends , Prospects and Innovation in the Digital Research Epoch. *ASEAN Innovative and Transformative Educational Journal*, *1*(1), 1–10.

Fergus, S., Botha, M., & Ostovar, M. (2023). Evaluating Academic Answers Generated Using ChatGPT. *Journal of Chemical Education*, *100*(4), 1672–1675. https://doi.org/10.1021/acs.jchemed.3c00087

Firdaus, T., Sholeha, S. A., & Jannah, M. (2024). Comparison of ChatGPT and Gemini AI in Answering Higher-Order Thinking Skill Biology Questions : Accuracy and Evaluation. *International Journal of Science Education and Teaching*, *3*(3), 126–138. https://doi.org/http://doi.org/10.14456/ijset.2024.11

Fuseini, I. (2024). A critical review of data mining in education on the levels and aspects of education. *Quality Education for All*, *1*(2), 41–59. https://doi.org/10.1108/QEA-01-2024-0006

Gilson, A., Safranek, C. W., Huang, T., Socrates, V., Chi, L., Taylor, R. A., & Chartash, D. (2023). How Does ChatGPT Perform on the United States Medical Licensing Examination? The Implications of Large Language Models for Medical Education and Knowledge Assessment. *JMIR Medical Education*, *9*, 1–9. https://doi.org/10.2196/45312

Gusdevi, H., Setyanto, A., & Utami, E. (2025). Cosine Similarity-Based Evidences Selection for Fact Verification Using SBERT on the FEVER Dataset.

COGITO Smart Journal –, *11*(1), 52–66.

Hasanah, U., & Mutiara, D. A. (2019). Perbandingan Metode Cosine Similarity dan Jaccard Similarity untuk Penilaian Otomatis Jawaban Pendek. In *Seminar Nasional Sistem Informasi dan Teknik Informatika (SENSITIF) 2019* (hal. 1255–1263).

Hidayatullah, H. T. (2025). Literasi AI Sebagai Sarana Katalis Bagi Konselor Sekolah Masa Depan : Meningkatkan Layanan Bimbingan dan Konseling. *Jurnal Consulenza:Jurnal Bimbingan Konseling dan Psikologi*, 55–65. Diambil dari http://ejurnal.uij.ac.id/index.php/CONS

Holis, R. M., Eko, P., Utomo, P., & Hutabarat, B. F. (2025). Semantic FAQ Chatbot Using SBERT (Sentence-BERT) and Cosine Similarity for Academic Services. *Brilliance: Research of Artificial Intelligence*, *5*(2), 915–922. https://doi.org/https://doi.org/10.47709/brilliance.v5i2.7027 Semantic

Huber, S., Wiemer, H., Schneider, D., & Ihlenfeldt, S. (2019). DMME : Data mining methodology for engineering applications – a holistic extension to the CRISP-DM model. *Procedia CIRP*, *79*(March), 403–408. https://doi.org/10.1016/j.procir.2019.02.106

Jonathan, E., Ugrasena, A. A., Theo, A., & Ursia, W. (2025). Comparative Analysis Based on Task-Oriented Evaluation on ChatGPT-4 and Gemini (2 . 5 Flash). In *Proceedings of the 3rd INCONITBIS 2025* (hal. 115–126).

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., … Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, *103*. https://doi.org/10.1016/j.lindif.2023.102274

Konda, B. (2024). Explore Data Mining (DM) Techniques that Data Scientists Adopt in IT.

Koto, F., Rahimi, A., Lau, J. H., & Baldwin, T. (2020). IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP. *arXiv*, 757–770. https://doi.org/10.18653/v1/2020.coling-main.66

Laddha, M. D., Lokare, V. T., Kiwelekar, A. W., & Netak, L. D. (2021). Classifications of the summative assessment for revised bloom's taxonomy by using deep learning. *International Journal of Engineering Trends and Technology*, *69*(3), 211–218. https://doi.org/10.14445/22315381/IJETT-V69I3P232

Laskar, M. T. R., Bari, M. S., Rahman, M., Bhuiyan, M. A. H., Joty, S., & Huang, J. X. (2023). A Systematic Study and Comprehensive Evaluation of ChatGPT on Benchmark Datasets. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics* (hal. 431–469).

https://doi.org/10.18653/v1/2023.findings-acl.29

LearnLM Team, Modi, A., Veerubhotla, A. S., Rysbek, A., Huber, A., Anand, A., … Assael, Y. (2025). Evaluating Gemini in an arena for learning. Diambil dari http://arxiv.org/abs/2505.24477

Martínez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernández-Orallo, J., Kull, M., Lachiche, N., … Flach, P. (2019). CRISP-DM Twenty Years Later : From Data Mining Processes to Data Science Trajectories.

Masuwai, A., Zulkifli, H., & Hamzah, M. I. (2024). Evaluation of content validity and face validity of secondary school Islamic education teacher self-assessment instrument. *Cogent Education*, *11*(1). https://doi.org/10.1080/2331186X.2024.2308410

Mustika, Ardila, Y., Manuhutu, A., Ahmad, N., Hasbi, I., Guntoro, … Ernawati, I. (2021). *Data Mining dan Aplikasinya*. *Penerbit Widina*. Diambil dari https://repository.penerbitwidina.com/uk/publications/351768/data-mining-dan-aplikasinya

Nuraeni, R., Fadilah, S., Arpiandi, Z., Manajemen, P., Islam, P., Sunan, U. I. N., & Djati, G. (2024). Pemanfaatan Gemini AI untuk Proses Pembelajaran Pada Lembaga Pendidikan Islam Masa Kini. *Gunung Djati Conference Series*, *45*, 121–128.

Peyton, K., & Unnikrishnan, S. (2023). Results in Engineering A comparison of chatbot platforms with the state-of-the-art sentence BERT for answering online student FAQs. *Results in Engineering*, *17*(October 2022), 100856. https://doi.org/10.1016/j.rineng.2022.100856

Prakoso, A. A., & Nugroho, F. R. (2025). Pemanfaatan Gemini AI sebagai Sarana Pemberi Informasi di Bidang Pendidikan. *jurnal wawasan pengembangan pendidikan*, *13*(02), 67–76.

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *arXiv*.

Riyani, A., & naf'an, muhammad burhanudin, A. (2019). Penerapan Cosine Similarity dan Pembobotan TF-IDF untuk Mendeteksi Kemiripan Dokumen. *Seminar Hasil Pengabdian Masyarakat 2019*, *2*(1), 49–54.

Salman. (2025). Sistem Pendeteksi Kemiripan Judul Skripsi Berbasis Web Menggunakan NLP dan Word Embeddings. *jurnal pendidikan indonesia: Teori, Penelitian dan Inovasi*, *5*(5). https://doi.org/10.59818/jpi.v5i5.1964

Santoso, J. T. (2023). *Kecerdasan buatan (Artificial intelligence)*.

Sato, K. (2024). Language Models' Approaches to Explaining.

Strzalkowski, P., Strzalkowska, A., Chhablani, J., Pfau, K., Errera, M. H., Roth, M., … Guthoff, R. (2024). Evaluation of the accuracy and readability of ChatGPT-4 and Google Gemini in providing information on retinal detachment: a multicenter expert comparative study. *International Journal of Retina and Vitreous*, *10*(1), 1–11. https://doi.org/10.1186/s40942-024-00579-9

Taşyürek, M., Adıgüzel, Ö., Gündoğar, M., Goncharuk-Khomyn, M., & Ortaç, H. (2025). Comparative evaluation of the responses from ChatGPT-5, Gemini 2.5 Flash, and DeepSeek-V3.1 chatbots to patient inquiries about endodontic treatment in terms of accuracy, understandability, and readability. *International Dental Research*, *15*(Advanced Online), 1–13. https://doi.org/10.5577/intdentres.662

Ulimaz, A., Cahyono, D., Dhaniswara, E., Arifudin, O., & Agus Rukiyanto, B. (2024). Analisis Dampak Kolaborasi Pemanfaatan Artificial Intelligences (AI) Dan Kecerdasan Manusia Terhadap Dunia Pendidikan Di Indonesia. *INNOVATIVE: Journal Of Social Science Research*, *4(3)*, 9312–9319.

Yüksel, Y., Gür, Ö. E., Bedel, C., Selvi, F., Zortuk, Ö., & Yıldız, G. (2024). Comparison of Chat GPT and Gemini in ENT Evaluation Questions. *Sarcouncil Journal of Medical Sciences*, *3*(12), 17–22.

Zhang, P., Wang, J., Hu, X., Wang, X., Fan, X., Chi, W., & Yang, W. (2026). Comparative performance of GPT-4, GPT-o3, GPT-5, Gemini-3-Flash, and DeepSeek-R1 in ophthalmology question answering. *Frontiers in Cell and Developmental Biology*, (January), 1–12. https://doi.org/10.3389/fcell.2026.1744389