PEMANFAATAN MODEL *DEEP LEARNING* (CHATGPT) DALAM DETEKSI KESALAHAN PENYELESAIAN SOAL MATEMATIKA: STUDI PERBANDINGAN PENILAIAN OTOMATIS DAN MANUAL

Dwi Lestari[1]*, Audi Alexandra Zeliyanti[2], Yonisa Aisyah Fitri[3], Annisa Maulidiyah Rahmi Lestari[4], Netriwati[5]
[1,2,3,4,5]Pendidikan Matematika FTK Universitas Islam Negeri Raden Intan Lampung
*Corresponding author**: dwilstr715@gmail.com

ABSTRACT

This study aims to analyze the effectiveness of deep learning models, particularly ChatGPT, in detecting errors in students' mathematical problem-solving processes and to compare automated assessment results with manual assessment conducted by lecturers. The research employed a quantitative descriptive-comparative approach involving 37 first-semester university students in a mathematics education program. Data were collected through written essay tests on plane geometry topics, questionnaires, and documentation. The written responses were assessed manually by lecturers and automatically by ChatGPT, focusing on conceptual, procedural, and computational errors. Data analysis used descriptive statistics and comparative analysis to examine score differences and consistency between the two assessment methods. The results show that the average score differences between manual assessment and ChatGPT assessment were relatively small, ranging from 0.4 to 4.5 points, indicating a high level of accuracy and consistency of the automated system. ChatGPT demonstrated advantages in efficiency, objectivity, and speed of assessment, while manual assessment remained superior in interpreting implicit reasoning and contextual understanding. These findings suggest that ChatGPT has strong potential as an automated assessment tool to support mathematics educators, particularly in identifying student error patterns systematically, although human judgment is still necessary for comprehensive pedagogical interpretation.

***Keywords**: deep learning, ChatGPT, mathematical error detection, automated assessment*

ABSTRAK

Penelitian ini bertujuan untuk menganalisis efektivitas model deep learning, khususnya ChatGPT, dalam mendeteksi kesalahan penyelesaian soal matematika serta membandingkan hasil penilaian otomatis dengan penilaian manual oleh dosen. Penelitian menggunakan pendekatan deskriptif kuantitatif-komparatif dengan melibatkan 37 mahasiswa semester I Program Studi Pendidikan Matematika. Data dikumpulkan melalui tes uraian pada materi bangun datar, kuesioner, dan dokumentasi. Jawaban mahasiswa dinilai secara manual oleh dosen dan secara otomatis menggunakan ChatGPT dengan fokus pada kesalahan konseptual, prosedural, dan perhitungan. Analisis data dilakukan menggunakan

statistik deskriptif dan komparatif untuk melihat perbedaan skor serta konsistensi kedua metode penilaian. Hasil penelitian menunjukkan bahwa selisih rata-rata skor antara penilaian manual dan ChatGPT relatif kecil, yaitu berkisar antara 0,4 hingga 4,5 poin, yang menunjukkan tingkat akurasi dan konsistensi ChatGPT yang tinggi. ChatGPT unggul dalam efisiensi waktu dan objektivitas penilaian, sementara penilaian manual lebih mampu menangkap konteks dan penalaran implisit mahasiswa. Dengan demikian, ChatGPT berpotensi menjadi alat bantu penilaian yang efektif dalam pembelajaran matematika, meskipun tetap perlu dikombinasikan dengan penilaian manusia untuk hasil yang lebih komprehensif.

***Kata Kunci**: deep learning, ChatGPT, kesalahan matematika, penilaian otomatis*

A. Pendahuluan

Di bidang pendidikan, penilaian hasil pembelajaran masih banyak dilakukan secara manual oleh para guru atau dosen. Cara penilaian ini memerlukan waktu yang cukup panjang, tingkat ketelitian yang tinggi, dan sering kali melibatkan unsur subjektivitas dalam memberikan nilai. Setiap orang yang menilai bisa memiliki pandangan berbeda terhadap tahapan penyelesaian tugas siswa, khususnya pada soal uraian yang memerlukan penilaian terhadap proses berpikir, bukan hanya hasil akhirnya. Hal ini berpotensi menyebabkan ketidaksesuaian dalam hasil penilaian serta menambah beban kerja para pendidik saat melakukan koreksi (Atasoy & Moslemi Nezhad Arani, 2025).

Perkembangan teknologi kecerdasan buatan (Artificial Intelligence/AI) yang begitu cepat dalam sepuluh tahun terakhir telah membawa perubahan besar di dunia pendidikan. Salah satu bidang yang paling menarik perhatian adalah deep learning (DL), yaitu teknik pembelajaran mesin yang memanfaatkan jaringan saraf tiruan dengan lapisan-lapisan dalam untuk memahami dan mengidentifikasi pola yang rumit. Sebuah penelitian berjudul "Rejecting Reduction: Clarifying the Concept of Deep Learning in Mathematics Teaching in the Era of Artificial Intelligence" menjelaskan bahwa istilah "deep learning" dalam pengajaran matematika memiliki dua arti utama: sebagai teknologi dan sebagai pendekatan pedagogi (Munfarikhatin & Natsir, 2025).

Dalam dunia pendidikan, deep learning menawarkan dua aspek utama yang saling mendukung. Pertama, dari sisi teknologi, algoritma dan model deep learning digunakan

untuk memproses data pembelajaran dalam skala besar serta mengidentifikasi pola kesalahan yang dilakukan siswa. Kedua, dari segi pedagogi, deep learning merujuk pada proses belajar yang mendalam, yaitu menghubungkan konsep-konsep, memahami secara menyeluruh, dan menerapkan pengetahuan dalam situasi kehidupan nyata. Sebuah artikel berjudul "Exploration of the Implementation of Deep Learning Approach in Teaching Mathematics in Secondary Schools" menjelaskan bagaimana guru matematika menerapkan strategi berbasis deep learning untuk meningkatkan pemahaman konseptual siswa (Syarnubi et al., 2024).

Kedua aspek teknologi dan pedagogi ini bisa diterapkan secara praktis dalam pembelajaran matematika. Dari segi teknologi, model deep learning dapat digunakan untuk mendeteksi kesalahan siswa selama proses penyelesaian soal, bukan hanya pada jawaban akhir, melainkan pada setiap langkah pemecahan. Dari segi pedagogi, kesalahan tersebut menjadi bahan untuk refleksi belajar yang berharga. Di sini, model generatif seperti ChatGPT memberikan kemampuan untuk memahami bahasa sehari-hari siswa, menilai proses berpikir mereka, dan memberikan tanggapan. Penelitian berjudul "Artificial Intelligence Integration: Error Self Reflection in Solving Integral Problems" menunjukkan bahwa pengintegrasian AI dalam pembelajaran integral bagi mahasiswa di Indonesia membantu siswa menyadari kesalahan dalam transformasi dan encoding (Aljura et al., 2025)

Kesalahan dalam matematika bukan sekadar tentang hasil yang tidak tepat, melainkan mencerminkan cara berpikir, penalaran, dan pemahaman konseptual siswa. Oleh karena itu, mendeteksi dan menganalisis kesalahan tidak hanya untuk menentukan benar atau salah, tetapi untuk mengungkap bagaimana siswa berpikir dan di mana hambatan prosedural atau konseptual terjadi. Artikel "Analysis of Student Errors in Solving Mathematical Story Problems Based on Newman's Theory in View of Student Learning Styles" membahas berbagai jenis kesalahan siswa dalam soal cerita matematika berdasarkan teori Newman, serta pengaruh gaya belajar terhadap kesalahan tersebut (Ulfa, 2024).

Penggunaan model deep learning seperti ChatGPT dalam penilaian otomatis memberikan kemudahan dalam proses evaluasi pembelajaran, termasuk penghematan waktu dan potensi objektivitas. Namun, ada pertanyaan penting mengenai validitas dan keandalan penilaian tersebut. Penilaian manual oleh dosen unggul dalam memahami konteks setempat, intuisi pedagogis, dan nuansa bahasa siswa, sedangkan sistem otomatis sangat tergantung pada data pelatihan dan algoritma. Artikel "ChatGPT: A Reliable Assistant for the Evaluation of Students' Written Texts?" membandingkan tingkat akurasi antara penilai manusia dan sistem ChatGPT dalam menilai teks tulisan siswa (Atasoy & Moslemi Nezhad Arani, 2025).

Di lingkungan pendidikan tinggi, terutama program studi pendidikan matematika, evaluasi tidak hanya melihat jawaban akhir mahasiswa, tetapi juga menilai cara berpikir mereka dalam menyelesaikan soal langkah demi langkah serta alasan di baliknya. Untuk membangun kompetensi profesional calon guru, deteksi kesalahan otomatis dapat mengungkap pola kesalahan konseptual, prosedural, atau teknis dengan lebih cepat dan terstruktur. Artikel "Chatgpt Assisted Teachers in Improving Formative Assessment" membahas bagaimana ChatGPT membantu guru merancang penilaian formatif dan memberikan umpan balik otomatis (Zhao et al., 2025).

Keunggulan ChatGPT terletak pada kemampuannya memahami konteks bahasa alami, bukan hanya angka atau simbol, sehingga dapat menilai penalaran matematis yang diungkapkan dalam bentuk narasi atau langkah-langkah deskriptif oleh mahasiswa. Dengan cara ini, model ini memungkinkan evaluasi yang lebih menyeluruh dan kontekstual, berbeda dari sistem penilaian berbasis aturan atau hanya numerik. Kajian "Distilling ChatGPT for Explainable Automated Student Answer Assessment " menunjukkan bahwa ChatGPT bisa disempurnakan untuk memberikan penjelasan atau dibalik penilaian otomatis (Li et al., 2023).

Meskipun demikian, penerapan AI dan deep learning di pendidikan masih dihadapkan pada berbagai tantangan. Kebanyakan penelitian sebelumnya lebih fokus pada hasil akhir pembelajaran, seperti skor atau jawaban benar, tetapi belum banyak

yang menganalisis proses berpikir siswa yang menyebabkan kesalahan. Selain itu, ada perbedaan antara penilaian otomatis dan manual dalam hal akurasi, konsistensi, dan kemampuan interpretasi. Artikel "ChatGPT and Its Impact on Students Assessment Practices in the Higher Educational Sector: A Systematic Review" menemukan bahwa penggunaan ChatGPT dalam praktik penilaian di pendidikan tinggi masih belum seragam (Ofusori & Hendradi, 2025).

Tantangan lainnya termasuk keterbatasan dataset pelatihan yang representatif dan belum adanya standar untuk melatih model AI agar bisa menggeneralisasi dengan baik. Selain itu, aspek keterjelasan model atau explainability—yaitu mengapa model membuat keputusan tertentu—sangat penting agar hasil deteksi kesalahan bisa digunakan secara pedagogis, bukan hanya sebagai angka atau label. Guru dan dosen perlu memahami alasan di balik evaluasi otomatis untuk merancang perbaikan yang tepat. Artikel "The Application of Machine Learning Algorithms in Analyzing Students' Conceptual Error Patterns in Science Learning" menjelaskan penggunaan algoritma pembelajaran mesin untuk mendeteksi pola kesalahan konseptual siswa di bidang sains (Hooshyar et al., 2024).

Dari sisi lokal dan kontekstual, penerapan ChatGPT dalam pembelajaran matematika di Indonesia masih menghadapi hambatan, seperti bias model terhadap data pelatihan yang sering berbahasa Inggris atau konteks Barat, serta kesulitan memahami konteks lokal, bahasa Indonesia, atau cara penulisan mahasiswa Indonesia. Oleh karena itu, penting untuk menguji kemampuan adaptasi model deep learning generatif dalam lingkungan akademik Indonesia. Artikel "Harnessing ChatGPT for Effective Assessment and Feedback in Education" menampilkan penggunaan ChatGPT untuk umpan balik dan penilaian di pendidikan umum, memberikan gambaran tantangan dan peluang (Lu et al., 2023).

Berdasarkan berbagai masalah tersebut, penelitian ini dilakukan untuk menguji akurasi, konsistensi, dan relevansi model deep learning generatif seperti ChatGPT dalam mendeteksi dan menilai kesalahan selama proses penyelesaian soal matematika mahasiswa, serta

membandingkannya dengan penilaian manual dosen. Penelitian ini tidak hanya menilai jawaban akhir, tetapi juga mempelajari pola kesalahan konseptual, prosedural, dan perhitungan yang muncul dalam proses penyelesaian. Pendekatan ini diharapkan memberikan gambaran empiris tentang efektivitas ChatGPT sebagai sistem penilaian otomatis yang objektif, efisien, dan mampu menyesuaikan dengan karakteristik kesalahan mahasiswa. Sebagai dasar, artikel "ChatGPT in the Classroom: Evaluating its Role in Fostering Critical Evaluation Skills" menjelaskan bagaimana ChatGPT digunakan dalam pembelajaran dan evaluasi kritis (Oates & Johnson, 2025).

Akhirnya, dari berbagai hasil penelitian sebelumnya, dapat disimpulkan bahwa pemanfaatan deep learning bukan hanya sebagai alat teknologi, tetapi juga sebagai pendekatan yang bisa memperdalam kualitas berpikir matematis siswa, dengan potensi yang besar. Dengan demikian, penelitian ini berusaha menjembatani kesenjangan antara penilaian manual dan otomatis melalui studi perbandingan yang menggunakan ChatGPT. Temuan yang dihasilkan diharapkan memberikan saran praktis bagi dosen dan lembaga pendidikan untuk merancang sistem penilaian yang lebih adaptif, reflektif, dan bermakna dalam pengembangan kompetensi matematika mahasiswa. Artikel "The Dawn of ChatGPT: Transformation in Science Assessment" menekankan bahwa transformasi penilaian oleh AI memerlukan integrasi pedagogis yang kuat agar tidak hanya fokus pada efisiensi saja (Li et al., 2023).

B. Metode Penelitian

Penelitian ini menerapkan pendekatan deskriptif kuantitatif-komparatif, yang bertujuan untuk menggambarkan dan menganalisis kesalahan mahasiswa saat menyelesaikan soal matematika, serta membandingkan hasil penilaian manual dengan penilaian otomatis melalui model deep learning seperti ChatGPT. Pendekatan ini dipilih karena data yang dikumpulkan meliputi hasil tes berupa angka, skor dari ChatGPT, dan jawaban kuesioner, yang kemudian diolah secara statistik deskriptif dan komparatif untuk memahami jenis kesalahan mahasiswa serta tingkat

ketepatan ChatGPT dalam mendeteksinya (Altamimi et al., 2025).

Penelitian ini berlangsung pada tanggal 29–30 September 2025 di Universitas Islam Negeri Raden Intan Lampung. Populasi penelitian mencakup semua mahasiswa semester I di universitas tersebut untuk tahun ajaran 2025/2026. Sampel terdiri dari 37 mahasiswa semester I yang dipilih menggunakan teknik convenience sampling, yaitu metode pengambilan sampel berdasarkan kemudahan akses dan kesediaan responden untuk ikut serta. Teknik ini diterapkan karena tidak semua mahasiswa dalam populasi bisa dijangkau peneliti, sehingga hanya sebagian yang hadir dan bersedia ambil bagian dalam pengumpulan data.

Instrumen penelitian disiapkan sendiri oleh peneliti dan terdiri dari tes tertulis, kuesioner, serta dokumentasi. Tes tertulis berupa lima soal uraian yang dibuat berdasarkan indikator pemahaman konsep dan keterampilan prosedural pada materi bangun datar, seperti menghitung luas dan keliling persegi, persegi panjang, segitiga, serta lingkaran. Tes ini bertujuan mengidentifikasi kesalahan mahasiswa, termasuk kesalahan konsep, prosedural, dan perhitungan.

Kuesioner disusun untuk mengetahui faktor-faktor yang mungkin menyebabkan kesalahan mahasiswa, seperti pemahaman konsep, ketelitian, motivasi belajar, dan kebiasaan belajar. Kuesioner ini mencakup 15 pernyataan dengan skala Likert lima tingkat, yaitu Sangat Setuju (SS), Setuju (S), Ragu-ragu (RR), Tidak Setuju (TS), dan Sangat Tidak Setuju (STS). Instrumen ini dikembangkan secara mandiri oleh peneliti dengan mempertimbangkan kesesuaian isi dan kejelasan bahasa agar mudah dipahami responden. Sebelum digunakan, instrumen ditinjau ulang oleh peneliti untuk memastikan relevansinya dengan tujuan penelitian, dan dilakukan uji coba kecil pada beberapa responden untuk mengecek keterbacaan, kejelasan, serta konsistensi setiap butir sebelum diterapkan dalam penelitian utama.

Instrumen dokumentasi berfungsi sebagai data tambahan, mencakup hasil kerja mahasiswa, foto kegiatan penelitian, dan catatan lapangan selama proses. Dokumentasi ini membantu memperkuat data dari tes dan

kuesioner, serta memberikan bukti autentik tentang pelaksanaan penelitian.

Pengumpulan data dilakukan dalam tiga tahap utama: pelaksanaan tes tertulis, pengisian kuesioner, dan pengumpulan dokumentasi. Tes diberikan terlebih dahulu untuk mengukur kemampuan mahasiswa dalam menyelesaikan soal matematika pada materi bangun datar. Setelah itu, mahasiswa mengisi kuesioner untuk mengungkap faktor-faktor penyebab kesalahan mereka. Seluruh kegiatan penelitian didokumentasikan dengan baik untuk menjaga keabsahan data yang diperoleh (Faseeh et al., 2024)

Data yang terkumpul dianalisis menggunakan statistik deskriptif dan komparatif. Hasil tes dianalisis dengan dua cara: (1) Menghitung jumlah dan persentase jenis kesalahan mahasiswa (analisis deskriptif); dan (2) Membandingkan skor penilaian manual dengan skor otomatis dari ChatGPT melalui perhitungan selisih rata-rata dan koefisien korelasi untuk menilai akurasi serta konsistensi model. Hasil kuesioner dianalisis dengan menghitung skor rata-rata dan persentase dari setiap indikator untuk melihat kecenderungan respon mahasiswa terhadap faktor penyebab kesalahan. Semua hasil analisis kemudian diinterpretasikan secara menyeluruh untuk memberikan gambaran lengkap tentang bentuk kesalahan mahasiswa dan faktor-faktor yang memengaruhinya (Bewersdorff et al., 2023).

Prosedur penelitian mencakup beberapa langkah, yaitu: (1) menyusun instrumen seperti tes tertulis dan kuesioner, (2) melaksanakan tes serta pengisian kuesioner kepada mahasiswa, (3) mengumpulkan dokumentasi kegiatan, (4) menganalisis data dengan statistik deskriptif, serta (5) menyimpulkan hasil berdasarkan data yang diperoleh (Yunianto et al., 2024).

C.Hasil Penelitian dan Pembahasan

Hasil Penelitian

Penelitian ini melibatkan data hasil pekerjaan mahasiswa serta pengisian kuesioner yang digunakan sebagai dasar analisis.

1. Uji Validitas

Uji validitas dilakukan untuk mengetahui sejauh mana butir-butir pertanyaan dalam instrumen penelitian mampu mengukur apa yang seharusnya diukur. Butir pernyataan dikatakan valid apabila nilai r hitung >

r tabel pada taraf signifikansi 5%. Berdasarkan hasil perhitungan terhadap 15 item pernyataan yang diujikan kepada 37 responden, seluruh item memiliki nilai r hitung lebih besar dari r tabel (0,325), sehingga dinyatakan valid dan dapat digunakan dalam penelitian.

Table 1. Case Processing Summary

Case Processing Summary

		N	%
Cases	Valid	37	100.0
	Excluded[a]	0	.0
	Total	37	100.0

Keterangan:

Berdasarkan hasil output SPSS pada tabel di atas, terdapat 37 responden (100%) yang memiliki data lengkap untuk semua variabel yang dianalisis. Tidak ada data yang dikeluarkan (excluded = 0%), sehingga dapat disimpulkan bahwa tidak terdapat data yang hilang (missing data). Dengan demikian, seluruh 37 responden disertakan dalam analisis reliabilitas.

2. Uji Reliabilitas

Uji reliabilitas digunakan untuk mengetahui tingkat konsistensi internal instrumen penelitian. Uji ini dilakukan menggunakan metode Cronbach's Alpha pada program SPSS.

Table 2. Reliability Statistics

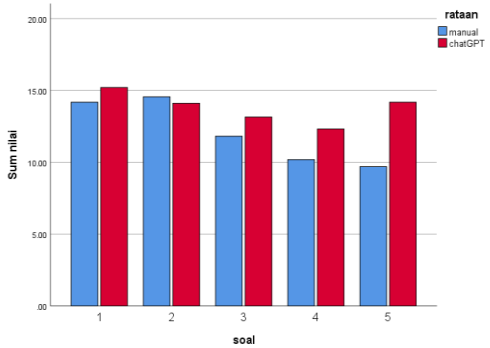
Reliability Statistics

Cronbach's Alpha	N of Items
.970	15

Keterangan:

Analisis reliabilitas dilakukan terhadap 15 butir pernyataan yang membentuk satu skala instrumen. Nilai Cronbach's Alpha sebesar 0,970 menunjukkan tingkat konsistensi internal yang sangat tinggi. Menurut Nunnally (1978), nilai Cronbach's Alpha di atas 0,70 menunjukkan reliabilitas yang baik, dan nilai di atas 0,90 menunjukkan reliabilitas yang sangat tinggi. Dengan demikian, instrumen penelitian ini dinyatakan reliabel, karena seluruh item pertanyaan secara konsisten mengukur konstruk yang sama.

Diagram 1. Diagram batang perbandingan hasil analisis manual dan ChatGPT



Keterangan:

Berdasarkan hasil perbandingan pada Diagram 1, terlihat bahwa nilai rata-

rata hasil analisis ChatGPT dan penilaian manual pada setiap soal menunjukkan pola yang hampir sama. Perbedaan nilai antara kedua metode relatif kecil, berkisar antara 0,4 hingga 4,5 poin. ChatGPT cenderung memberikan nilai sedikit lebih tinggi pada beberapa soal, khususnya soal ke-1, ke-3, dan ke-5, sedangkan pada soal ke-2 nilai manual sedikit lebih unggul. Hal ini menunjukkan bahwa ChatGPT mampu melakukan penilaian dengan tingkat akurasi dan konsistensi yang tinggi, mendekati hasil penilaian manual oleh dosen. Dengan demikian, diagram ini menggambarkan bahwa sistem penilaian otomatis berbasis Deep Learning (ChatGPT) memiliki potensi yang baik dalam membantu proses evaluasi hasil belajar matematika, terutama dalam hal efisiensi waktu dan objektivitas penilaian.

Pembahasan

Keterbaruan penelitian ini terletak pada analisis kesalahan langkah demi langkah dalam penyelesaian soal matematika menggunakan ChatGPT, bukan hanya penilaian hasil akhir seperti yang dilakukan penelitian sebelumnya (Altamimi et al., 2025; Bewersdorff et al., 2023). Penelitian ini juga menunjukkan perbandingan langsung antara penilaian manual dan otomatis di kalangan mahasiswa Pendidikan Matematika di Indonesia, yang belum banyak diteliti (Ofusori & Hendradi,2025). Hasilnya menunjukkan selisih skor yang kecil, antara 0,4 hingga 4,5 poin, dengan korelasi yang tinggi, yang membuktikan ketepatan ChatGPT dalam evaluasi cepat, sesuai dengan temuan Testolin (2024) tentang kemampuan numerik model deep learning.

Hasil penelitian mengonfirmasi bahwa ada korelasi yang sangat kuat antara hasil penilaian otomatis dengan ChatGPT dan hasil penilaian manual. Ini didukung oleh selisih nilai rata-rata yang sangat kecil, berkisar 0,4 sampai 4,5. Temuan ini menunjukkan bahwa ChatGPT memiliki ketepatan dan konsistensi tinggi dalam memberikan skor, hampir sama dengan penilai manusia (Testolin, 2024).

Tingginya ketepatan model deep learning generatif ini disebabkan oleh jenis kesalahan yang dievaluasi, yaitu kesalahan prosedural dan perhitungan, yang bersifat jelas dan tertulis. Karena kesalahan ini bisa dideteksi secara tekstual, ChatGPT mampu melakukannya dengan tepat, sehingga akurasinya mendekati

penilai manual dalam sebagian besar kasus (Pujawati et al., 2025).

ChatGPT menunjukkan keunggulan dalam hal kecepatan, objektivitas, efisiensi, dan konsistensi, sehingga menjadi alat bantu penting bagi pendidik. Namun, penelitian ini juga menemukan batasan model, terutama dalam memahami konteks dan logika penyelesaian siswa yang tidak tertulis jelas. Penilai manusia masih lebih unggul karena bisa menangkap niat dan penalaran implisit, seperti dalam kasus di mana penilai manual memberikan skor 87 karena menghargai pemahaman konseptual, sedangkan ChatGPT hanya fokus pada kesalahan langkah prosedural dengan skor 85 (Niemi et al., 2022).

Temuan tentang ketepatan model bahasa dalam penalaran matematis ini sejalan dengan penelitian Lu et al. (2023). Di lain pihak, batasan model dalam memahami penalaran implisit menegaskan kembali kesenjangan penelitian sebelumnya tentang rendahnya tingkat interpretabilitas model deep learning dalam aspek pedagogis (Zhang et al., 2020).

Secara keseluruhan, penelitian ini memiliki implikasi praktis yang penting. ChatGPT berpotensi besar sebagai alat bantu yang efisien dan objektif dalam proses evaluasi bagi pendidik. Secara teoretis, penelitian ini memberikan bukti empiris tentang keandalan model deep learning generatif seperti ChatGPT dalam penilaian pendidikan matematika, sambil memetakan batasannya dalam interpretasi penalaran implisit. Dengan demikian, penelitian ini berkontribusi pada pengembangan paradigma pembelajaran yang didukung data dan pemahaman mendalam tentang cara siswa belajar (García-Varela et al., 2025).

D. Kesimpulan

Berdasarkan hasil penelitian dan pembahasan, model *deep learning* yang diimplementasikan melalui ChatGPT menunjukkan tingkat akurasi dan konsistensi yang tinggi dalam menilai hasil penyelesaian soal matematika materi bangun datar, yang dibuktikan dengan selisih skor yang sangat kecil (berkisar antara 0 hingga 7 poin) jika dibandingkan dengan hasil penilaian manual.Secara komparatif, ChatGPT memiliki keunggulan signifikan dalam hal efisiensi waktu dan objektivitas penilaian, menjadikannya alat bantu yang potensial bagi pendidik untuk mengurangi beban kerja evaluasi.

Meskipun demikian, penelitian ini menegaskan bahwa penilaian manual masih memiliki keunggulan dalam menginterpretasi konteks dan proses berpikir siswa yang tidak diekspresikan secara eksplisit dalam tulisan. Oleh karena itu, penelitian ini menyimpulkan bahwa ChatGPT memiliki potensi besar sebagai alat bantu yang efisien bagi pendidik dalam proses evaluasi pembelajaran matematika, khususnya dalam mendeteksi kesalahan konseptual maupun prosedural secara cepat dan sistematis.

DAFTAR PUSTAKA

Jurnal :

Aljura, A. N., Retnawati, H., Zulnaidi, H., & Mbazumutima, V. (2025). Understanding High School Students' Errors in solving Mathematics Problems: A Phenomenological Research. *Indonesian Journal on Learning and Advanced Education (IJOLAE)*, *7*(1), 154–178. https://doi.org/10.23917/ijolae.v7i1.24005

Altamimi, M., Altameemi, Y., Alkhalil, A., Mansour, R. F., Abdelrhman, M., Ahmed, I., Ahmad, A., & Alogali, A. (2025). A deep learning model for automated marking of students' assessments in a learning management system (LMS). *International Journal of Advanced and Applied Sciences*, *12*(10), 1–10. https://doi.org/10.21833/ijaas.2025.10.001

Atasoy, A., & Moslemi Nezhad Arani, S. (2025). ChatGPT: A reliable assistant for the evaluation of students' written texts? In *Education and Information Technologies* (Vol. 30, Issue 14). Springer US. https://doi.org/10.1007/s10639-025-13553-1

Bewersdorff, A., Seßler, K., Baur, A., Kasneci, E., & Nerdel, C. (2023). Assessing student errors in experimentation using artificial intelligence and large language models: A comparative study with human raters. *Computers and Education: Artificial Intelligence*, *5*. https://doi.org/10.1016/j.caeai.2023.100177

Faseeh, M., Jaleel, A., Iqbal, N., Ghani, A., Abdusalomov, A., Mehmood, A., & Cho, Y. I. (2024). Hybrid Approach to Automated Essay Scoring: Integrating Deep Learning Embeddings with Handcrafted Linguistic Features for Improved Accuracy. *Mathematics*, *12*(21). https://doi.org/10.3390/math12213416

García-Varela, F., Nussbaum, M., Mendoza, M., Martínez-Troncoso, C., & Bekerman, Z. (2025). ChatGPT as a Stable and Fair Tool for Automated Essay Scoring. *Education Sciences*, *15*(8). https://doi.org/10.3390/educsci15080946

Hooshyar, D., Azevedo, R., & Yang, Y. (2024). Augmenting Deep Neural Networks with Symbolic Educational Knowledge: Towards Trustworthy and Interpretable AI for Education. *Machine Learning*

and Knowledge Extraction, *6*(1), 593–618. https://doi.org/10.3390/make6010028

Li, J., Gui, L., Zhou, Y., West, D., Aloisi, C., & He, Y. (2023). Distilling ChatGPT for Explainable Automated Student Answer Assessment. *Findings of the Association for Computational Linguistics: EMNLP 2023*, 6007–6026. https://doi.org/10.18653/v1/2023.findings-emnlp.399

Lu, P., Qiu, L., Yu, W., Welleck, S., & Chang, K. W. (2023). A Survey of Deep Learning for Mathematical Reasoning. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, *1*, 14605–14631. https://doi.org/10.18653/v1/2023.acl-long.817

Munfarikhatin, A., & Natsir, I. (2025). Rejecting Reduction: Clarifying the Concept of Deep Learning in Mathematics Teaching in the Era of Artificial Intelligence. *J Statistika: Jurnal Ilmiah Teori Dan Aplikasi Statistika*, *18*(1), 930–936. https://doi.org/10.36456/jstat.vol18.no1.a10570

Niemi, H., Pea, R. D., & Lu, Y. (2022). AI in Learning: Designing the Future. In *AI in Learning: Designing the Future*. https://doi.org/10.1007/978-3-031-09687-7

Oates, A., & Johnson, D. (2025). ChatGPT in the Classroom: Evaluating its Role in Fostering Critical Evaluation Skills. *International Journal of Artificial Intelligence in Education*. https://doi.org/10.1007/s40593-024-00452-8

Ofusori, L. O., & Hendradi, R. (2025). ChatGPT and Its Impact on Students Assessment Practices in the Higher Educational Sector: A Systematic Review. *Journal of Information Systems Engineering and Business Intelligence*, *11*(1), 65–78. https://doi.org/10.20473/jisebi.11.1.65-78

Pujawati, F., Azkia, M. N., & Susilawati, W. (2025). Exploration of the Implementation of Deep Learning Approach in Teaching Mathematics in Secondary Schools. *Unnes Journal of Mathematics Education*, *14*(2), 98–105. https://doi.org/10.15294/ujme.v14i2.27374

Syarnubi, Efriani, A., Pranita, S., Zulhijra, Anggara, B., Alimron, Maryamah, & Rohmadi. (2024). An analysis of student errors in solving HOTS mathematics problems based on the newman procedure. *AIP Conference Proceedings*, *3058*(1), 321–332. https://doi.org/10.1063/5.0201077

Testolin, A. (2024). Can Neural Networks Do Arithmetic? A Survey on the Elementary Numerical Skills of State-of-the-Art Deep Learning Models. *Applied Sciences (Switzerland)*, *14*(2). https://doi.org/10.3390/app14020744

Ulfa, S. M. (2024). Analysis of Student Errors in Solving Mathematical Story Problems Based on Newman's Theory in View of Student Learning Styles. *Journal of Mathematical Pedagogy (JoMP)*, *4*(2), 97–105. https://doi.org/10.26740/jomp.v4n2.p97-105

Yunianto, W., Lavicza, Z., Kastner-Hauler, O., & Houghton, T. (2024). Investigating the use of

ChatGPT to solve a GeoGebra based mathematics+computational thinking task in a geometry topic. *Journal on Mathematics Education*, *15*(3), 1027–1052. https://doi.org/10.22342/jme.v15i3.pp1027-1052

Zhang, H., Li, L. H., Meng, T., Chang, K., & Broeck, G. Van Den. (2020). *On the Paradox of Learning to Reason from Data*.

Zhao, C., Silva, M., & Poulsen, S. (2025). Autograding Mathematical Induction Proofs with Natural Language Processing. *International Journal of Artificial Intelligence in Education*. https://doi.org/10.1007/s40593-025-00498-2