AL-IRSYAD
Journal of Physics Educations
https://ejurnal.stkipddipinrang.ac.id/index.php/wjpe

ANALISIS KOMPARATIF KUALITAS UMPAN BALIK CHATGPT DAN DEEPSEEK TERHADAP JAWABAN KONTEKSTUAL SISWA PADA KONSEP OPTIK FISIKA

A Comparative Analysis of the Quality of Feedback from ChatGPT and Deepsseek Regarding Students' Contextual Responses to Concepts in Physics Optics

Siti Faitul Hidayah[1,*)], Miftah[2], I Komang Werdhiana[3], Gustina[4]

[1, 2, 3, 4)]Universitas Tadulako*

[*)]sitifaitulhd.untad@gmail.com

Info Artikel:

Dikirim:
26 Mei 2026
Revisi:
1 Juni 2026
Diterima:
21 Juni 2026

Keyword:

AI Feedback, ChatGPT, Contextual Responses, DeepSeek, Physics Optics

Kata Kunci:

ChatGPT, DeepSeek, Jawaban Kontekstual, Optik Fisika, Umpan Balik AI

Abstract

This study aimed to analyze and compare the quality of feedback generated by ChatGPT 5.5 and DeepSeek V4 on students' contextual responses in physics optics. The study employed a mixed-methods approach with an explanatory sequential design. Data were collected from 20 students who answered six contextual optics questions. Each student response was subsequently evaluated by ChatGPT 5.5 and DeepSeek V4, resulting in 240 AI-generated feedback units. Feedback quality was assessed using four indicators: scientific accuracy, depth of explanation, relevance to students' responses, and language clarity. Quantitative data were analyzed using descriptive statistics, normality testing, and the Wilcoxon Signed-Rank Test, while qualitative data were analyzed through content analysis of the feedback. The results showed that ChatGPT 5.5 achieved a mean total score of 15.31 (95.68%), whereas DeepSeek V4 achieved a mean total score of 14.68 (91.77%). Both models were categorized as excellent. The Wilcoxon test revealed significant differences in scientific accuracy, depth of explanation, relevance to students' responses, and total scores, while no significant difference was found in language clarity. Qualitative analysis indicated that ChatGPT 5.5 was more consistent in providing accurate, in-depth, and contextually relevant conceptual feedback, whereas DeepSeek V4 demonstrated strong language clarity but greater variability in content quality. Therefore, the primary difference between the two models lies in the quality of feedback content rather than linguistic aspects. These findings provide empirical evidence regarding the pedagogical characteristics of feedback generated by generative AI models, showing that ChatGPT 5.5 is more effective in delivering accurate, comprehensive, and contextually relevant feedback, while DeepSeek V4 remains competitive, particularly in language clarity. The findings may serve as a reference for educators in selecting and utilizing generative AI tools to support formative assessment in physics education.

Abstrak

Penelitian ini bertujuan menganalisis dan membandingkan kualitas umpan balik ChatGPT 5.5 dan DeepSeek V4 terhadap jawaban kontekstual siswa pada konsep optik fisika. Penelitian menggunakan pendekatan campuran dengan desain explanatory sequential. Data diperoleh dari 20 siswa yang mengerjakan enam soal kontekstual optik fisika. Setiap jawaban siswa diberi umpan balik oleh ChatGPT 5.5 dan DeepSeek V4 sehingga diperoleh 240 unit umpan balik AI. Kualitas umpan balik dinilai berdasarkan empat indikator, yaitu akurasi ilmiah, kedalaman penjelasan, relevansi terhadap konteks jawaban siswa, dan kejelasan bahasa. Data kuantitatif dianalisis menggunakan statistik deskriptif, uji normalitas, dan Wilcoxon Signed-Rank Test, sedangkan data kualitatif dianalisis melalui telaah isi umpan balik. Hasil penelitian menunjukkan bahwa ChatGPT 5.5 memperoleh mean total 15,31 atau 95,68%, sedangkan DeepSeek V4 memperoleh mean total 14,68 atau 91,77%. Keduanya berada pada kategori sangat baik. Uji Wilcoxon menunjukkan perbedaan signifikan pada akurasi ilmiah, kedalaman penjelasan, relevansi terhadap konteks jawaban siswa, dan total skor, sedangkan kejelasan bahasa tidak berbeda signifikan. Analisis kualitatif menunjukkan bahwa ChatGPT 5.5 lebih stabil dalam memberikan koreksi konseptual yang akurat, mendalam, dan relevan, sedangkan DeepSeek V4 tetap kuat pada kejelasan bahasa tetapi lebih bervariasi pada substansi isi. Dengan demikian, perbedaan utama kedua model terletak pada kualitas isi umpan balik, bukan pada aspek kebahasaan. Temuan ini memberikan bukti empiris bahwa ChatGPT 5.5 lebih unggul dalam menghasilkan umpan balik yang akurat, mendalam, dan sesuai dengan konteks jawaban siswa, sedangkan DeepSeek V4 menunjukkan performa yang kompetitif terutama pada aspek kejelasan bahasa. Hasil penelitian ini dapat menjadi pertimbangan bagi pendidik dalam memilih dan memanfaatkan AI generatif sebagai pendukung asesmen formatif pada pembelajaran fisika.

I. PENDAHULUAN

Perkembangan Artificial Intelligence (AI) telah membawa perubahan dalam berbagai bidang, termasuk pendidikan. AI dimanfaatkan untuk mendukung personalisasi pembelajaran, penyediaan sumber belajar, serta pemberian umpan balik yang lebih cepat dan adaptif kepada siswa (Perera & Lankathilaka, 2023). Salah satu perkembangan penting dalam bidang ini adalah munculnya AI generatif yang mampu menghasilkan berbagai bentuk konten secara otomatis berdasarkan masukan pengguna.

Model bahasa generatif seperti ChatGPT dan DeepSeek tidak hanya mampu menghasilkan teks, tetapi juga dapat digunakan untuk memberikan penjelasan, tanggapan, dan umpan balik terhadap hasil kerja siswa. Lo (2023) menjelaskan bahwa model bahasa generatif memiliki potensi untuk mendukung pembelajaran melalui interaksi yang responsif dan mudah diakses. Namun, dalam konteks pendidikan fisika, keluaran AI yang tampak lancar dan meyakinkan belum tentu selalu tepat secara ilmiah atau sesuai dengan kebutuhan belajar siswa sehingga perlu dievaluasi secara kritis (Polverini & Gregorcic, 2024).

Salah satu bentuk pemanfaatan AI yang relevan dalam pembelajaran adalah pemberian umpan balik. Umpan balik berfungsi membantu siswa mengenali bagian jawaban yang sudah tepat, memahami kekeliruan, dan mengetahui arah perbaikan yang perlu dilakukan. Panadero dan

Lipnevich (2022) serta Carless (2022) menjelaskan bahwa umpan balik yang efektif tidak cukup hanya menyatakan benar atau salah, tetapi harus memberikan informasi yang dapat digunakan untuk memperbaiki pemahaman. Dengan demikian, kualitas umpan balik perlu dinilai dari ketepatan isi, kedalaman penjelasan, keterkaitan dengan jawaban siswa, dan keterpahaman bahasa.

Kebutuhan terhadap umpan balik yang berkualitas semakin penting dalam pembelajaran fisika. Fisika menuntut siswa tidak hanya memberikan jawaban akhir, tetapi juga menunjukkan pemahaman konsep, penalaran ilmiah, dan kemampuan menghubungkan fenomena dengan prinsip fisika yang mendasarinya. Pada materi optik, siswa sering mengalami kesulitan dalam menjelaskan fenomena seperti pembiasan, pemantulan, pembentukan bayangan, dan hubungan antara lensa, jarak benda, serta jarak bayangan. Kaltakci-Gurel (2023) dan Kunaedi et al. (2024) menunjukkan bahwa kesulitan dan miskonsepsi pada konsep optik masih sering muncul, terutama ketika siswa harus mengaitkan representasi visual dengan penjelasan konseptual.

ChatGPT dan DeepSeek merupakan dua model AI generatif yang dapat menghasilkan umpan balik tertulis. ChatGPT banyak dikaji dalam konteks pendidikan karena kemampuannya menghasilkan respons yang komunikatif, mudah dipahami, dan mampu memberikan penjelasan yang relatif rinci (Lo, 2023). Namun, beberapa penelitian menunjukkan bahwa ChatGPT masih berpotensi menghasilkan informasi yang kurang akurat atau penjelasan yang tampak meyakinkan meskipun tidak sepenuhnya sesuai dengan konsep ilmiah (Owan et al., 2023; Polverini & Gregorcic, 2024). Di sisi lain, DeepSeek mulai mendapat perhatian sebagai model yang menonjol pada efisiensi komputasi dan kemampuan penalaran (Guo et al., 2025; Jin et al., 2025). Meskipun demikian, kajian mengenai kualitas respons DeepSeek dalam konteks pendidikan masih relatif terbatas sehingga konsistensi dan kualitas umpan balik yang dihasilkannya masih memerlukan kajian lebih lanjut. Oleh karena itu, meskipun kedua model memiliki potensi sebagai pendukung pembelajaran, kualitas umpan balik yang dihasilkan tidak dapat diasumsikan setara. Respons yang jelas secara bahasa belum tentu akurat secara konsep, mendalam dalam menjelaskan alasan ilmiah, atau relevan terhadap jawaban siswa.

Kajian sebelumnya telah menunjukkan bahwa AI generatif memiliki potensi untuk mendukung pembelajaran, pemberian umpan balik, dan personalisasi proses belajar (Perera & Lankathilaka, 2023). Secara khusus, Lo (2023) melaporkan bahwa ChatGPT berpotensi dimanfaatkan sebagai tutor virtual dan pendukung pembelajaran mandiri, sedangkan Polverini dan Gregorcic (2024) menekankan pentingnya mengevaluasi keluaran model bahasa besar dalam konteks pendidikan fisika karena kualitas respons yang dihasilkan tidak selalu mencerminkan ketepatan konseptual. Selain itu, beberapa penelitian mulai membandingkan karakteristik berbagai model AI generatif pada tugas berbasis bahasa dan menunjukkan adanya perbedaan performa antar model. Namun, penelitian yang secara khusus membandingkan kualitas umpan balik ChatGPT 5.5 dan DeepSeek V4 terhadap jawaban kontekstual siswa pada konsep optik fisika berdasarkan indikator akurasi ilmiah, kedalaman penjelasan, relevansi terhadap konteks jawaban siswa, dan kejelasan bahasa masih terbatas.

Berdasarkan latar tersebut, penelitian ini bertujuan menganalisis dan membandingkan kualitas umpan balik ChatGPT 5.5 dan DeepSeek V4 terhadap jawaban kontekstual siswa pada konsep optik fisika. Perbandingan dilakukan berdasarkan empat indikator, yaitu akurasi ilmiah, kedalaman penjelasan, relevansi terhadap konteks jawaban siswa, dan kejelasan bahasa. Selain mengidentifikasi perbedaan kualitas umpan balik kedua model, penelitian ini bertujuan memberikan bukti empiris mengenai karakteristik pedagogis umpan balik AI generatif dalam konteks pembelajaran fisika. Temuan penelitian diharapkan dapat memperkaya kajian mengenai evaluasi kualitas umpan balik AI serta menjadi dasar pertimbangan dalam pemanfaatan AI untuk mendukung asesmen formatif dan pembelajaran fisika.

II. METODE PENELITIAN

Penelitian ini menggunakan pendekatan campuran dengan desain explanatory sequential. Pada tahap awal, data kuantitatif dianalisis untuk menggambarkan dan membandingkan kualitas

umpan balik kedua model AI. Pada tahap berikutnya, analisis kualitatif dilakukan untuk menjelaskan pola perbedaan skor melalui telaah isi umpan balik. Desain ini digunakan karena perbandingan kualitas umpan balik tidak cukup dijelaskan melalui angka, tetapi juga memerlukan analisis terhadap karakter koreksi, ketepatan konsep, dan relevansi isi umpan balik.

Subjek penelitian terdiri atas 20 siswa kelas XII MAS Muhammadiyah Palu yang dipilih menggunakan teknik purposive sampling. Pemilihan subjek didasarkan pada pertimbangan bahwa siswa telah mempelajari materi optik sesuai dengan kurikulum yang berlaku dan mampu memberikan jawaban tertulis terhadap soal kontekstual yang digunakan dalam penelitian. Para siswa mengerjakan enam soal kontekstual optik fisika yang selanjutnya digunakan sebagai sumber data untuk menghasilkan umpan balik dari ChatGPT 5.5 dan DeepSeek V4. Jawaban siswa terhadap setiap soal digunakan sebagai stimulus yang sama bagi ChatGPT 5.5 dan DeepSeek V4. Dengan demikian, setiap model menghasilkan 120 unit umpan balik, sehingga keseluruhan data berjumlah 240 unit umpan balik AI. Objek penelitian adalah kualitas umpan balik yang dihasilkan oleh kedua model terhadap jawaban kontekstual siswa.

Instrumen penelitian terdiri atas soal kontekstual optik fisika, prompt AI, dan rubrik penilaian kualitas umpan balik. Soal kontekstual digunakan untuk memperoleh jawaban siswa yang menuntut penalaran pada konsep optik. rompt AI disusun agar kedua model berperan sebagai pemberi umpan balik terhadap jawaban siswa, bukan sebagai penjawab soal. Untuk menjaga kesetaraan perlakuan dan validitas perbandingan, prompt yang sama digunakan pada ChatGPT 5.5 dan DeepSeek V4 tanpa perubahan redaksi, struktur instruksi, maupun informasi yang diberikan. Dengan demikian, perbedaan umpan balik yang dihasilkan dapat diinterpretasikan sebagai perbedaan karakteristik respons kedua model, bukan akibat perbedaan stimulus yang diberikan. Rubrik penilaian digunakan untuk menilai kualitas umpan balik berdasarkan empat indikator utama yang disajikan pada Tabel 1.

Tabel 1. Indikator Penilaian Kualitas Umpan Balik AI

Indikator	Fokus Penilaian
Akurasi Ilmiah	Kesesuaian umpan balik dengan konsep optik fisika yang benar.
Kedalaman Penjelasan	Kelengkapan alasan konseptual, hubungan sebab-akibat, dan dasar perbaikan
Relevansi terhadap Konteks Jawaban Siswa	Kesesuaian koreksi dengan isi, kekeliruan, dan kebutuhan
Kejelasan Bahasa	Keterpahaman, keruntutan, dan kejelasan penyampaian

Catatan: Setiap Indikator dinilai menggunakan Skala Likert, 1-4; Skor Maksimum total adalah 16.

Validasi isi dilakukan melalui expert judgment terhadap instrumen soal, prompt AI, dan rubrik penilaian untuk menilai kesesuaian isi dengan tujuan penelitian dan indikator yang diukur. Dari 18 calon butir soal yang divalidasi, 15 butir dinyatakan layak. Sebanyak 6 butir kemudian dipilih sebagai instrumen final karena telah merepresentasikan seluruh level kognitif yang menjadi fokus penelitian (C4, C5, dan C6) serta dinilai memadai untuk mengungkap variasi kualitas umpan balik AI pada konteks pembelajaran optik fisika. Instrumen soal kontekstual memperoleh nilai Cronbach's Alpha sebesar 0,710 sehingga dinyatakan reliabel. Rubrik penilaian memperoleh Cronbach's Alpha sebesar 0,688 sehingga dikategorikan cukup reliabel dengan interpretasi hati-hati karena indikator yang digunakan bersifat multidimensi.

Penilaian kualitas umpan balik dilakukan oleh dua penilai. Reliabilitas antarpenilai dianalisis menggunakan Intraclass Correlation Coefficient (ICC) dengan model Two-Way Mixed dan tipe Absolute Agreement. Karena terdapat beberapa perbedaan skor antarpenilai, skor akhir ditetapkan melalui konsensus atau penyelarasan penilaian berdasarkan rubrik dan kunci jawaban setiap soal. Skor final inilah yang digunakan dalam analisis statistik deskriptif, uji normalitas, dan uji Wilcoxon Signed-Rank Test.

Analisis data kuantitatif dilakukan dengan menghitung nilai minimum, maksimum, total skor, mean, standar deviasi, dan persentase kategori kualitas. Uji normalitas dilakukan terhadap selisih pasangan skor ChatGPT 5.5 dan DeepSeek V4. Karena seluruh data selisih tidak berdistribusi normal, uji perbandingan dilakukan menggunakan Wilcoxon Signed-Rank Test. Analisis kualitatif dilakukan melalui telaah isi umpan balik untuk menjelaskan mengapa perbedaan skor muncul pada indikator tertentu.

III. HASIL DAN PEMBAHASAN

Kualitas Umpan Balik ChatGPT 5.5 dan DeepSeek V4

Hasil analisis deskriptif menunjukkan bahwa ChatGPT 5.5 dan DeepSeek V4 sama-sama menghasilkan umpan balik dengan kategori sangat baik. Namun, capaian ChatGPT 5.5 lebih tinggi daripada DeepSeek V4 pada mean total, persentase kualitas, dan kestabilan skor. Ringkasan hasil deskriptif kedua model disajikan pada Tabel 2.

Tabel 2. Statistik Deskriptif Skor Total Kualitas Umpan Balik AI

Model AI	N	Min	Maks	Total Skor	Mean	SD	Persentase	Kategori
ChatGPT 5.5	120	11	16	1837	15,31	1,04	95,68%	Sangat Baik
DeepSeek V4	120	10	16	1762	14,68	1,29	91,77%	Sangat Baik

Catatan: Skor maksimum total = 16.

Berdasarkan Tabel 2, ChatGPT 5.5 memperoleh mean total 15,31 atau 95,68%, sedangkan DeepSeek V4 memperoleh mean total 14,68 atau 91,77%. Selisih mean sebesar 0,63 menunjukkan bahwa secara deskriptif ChatGPT 5.5 menghasilkan kualitas umpan balik yang lebih tinggi. Selain itu, standar deviasi ChatGPT 5.5 sebesar 1,04 lebih kecil daripada DeepSeek V4 sebesar 1,29. Hal ini menunjukkan bahwa kualitas umpan balik ChatGPT 5.5 lebih stabil, sedangkan DeepSeek V4 memiliki variasi skor yang lebih besar.

Capaian kedua model yang sama-sama berada pada kategori sangat baik menunjukkan bahwa AI generatif memiliki potensi sebagai pendukung pemberian umpan balik dalam pembelajaran fisika. Namun, perbedaan mean dan standar deviasi memperlihatkan bahwa kualitas umpan balik tidak dapat hanya dinilai dari kemampuan model menghasilkan teks yang lancar. Kualitas yang lebih penting terletak pada kemampuan model membaca jawaban siswa, mengenali kekeliruan konsep, dan memberikan arahan perbaikan yang sesuai.

Tabel 3. Perbandingan Kualitas Umpan Balik Berdasarkan Indikator

Indikator	Mean ChatGPT 5.5	Mean DeepSeek V4	Selisih Mean	Makna Pembeda
Akurasi ilmiah	3,70	3,53	0,17	Ketepatan konsep fisika
Kedalaman penjelasan	3,66	3,37	0,29	Kelengkapan uraian konseptual
Relevansi terhadap konteks jawaban siswa	3,96	3,79	0,17	Kesesuaian dengan jawaban siswa
Kejelasan bahasa	3,99	3,99	0,00	Tidak menjadi pembeda utama

Catatan: Skor maksimum setiap indikator = 4.

Tabel 3 menunjukkan bahwa perbedaan terbesar terdapat pada indikator kedalaman penjelasan dengan selisih mean 0,29. Temuan ini mengindikasikan bahwa ChatGPT 5.5 cenderung

lebih kuat dalam memberikan uraian konseptual yang lengkap dibandingkan DeepSeek V4. Pada indikator akurasi ilmiah dan relevansi terhadap konteks jawaban siswa, ChatGPT 5.5 juga memperoleh mean lebih tinggi, meskipun selisihnya lebih kecil. Sementara itu, kejelasan bahasa memperoleh mean yang sama, yaitu 3,99, sehingga aspek kebahasaan bukan pembeda utama antara kedua model. Temuan ini sejalan dengan Lo (2023) yang menunjukkan bahwa model bahasa generatif memiliki potensi untuk menghasilkan respons yang informatif dan mendukung pembelajaran. Namun, sebagaimana dikemukakan oleh Polverini dan Gregorcic (2024), kualitas respons AI dalam pembelajaran fisika tidak hanya ditentukan oleh kelancaran bahasa, tetapi juga oleh ketepatan konsep dan kualitas penjelasan yang diberikan. Oleh karena itu, perbedaan yang ditemukan dalam penelitian ini menunjukkan bahwa kualitas umpan balik AI lebih banyak dipengaruhi oleh substansi isi dibandingkan aspek kebahasaan semata.

Pola tersebut memperlihatkan bahwa kedua model relatif setara dalam menyampaikan umpan balik secara jelas, tetapi tidak sepenuhnya setara dalam substansi. ChatGPT 5.5 lebih konsisten dalam menjaga ketepatan konsep, mengembangkan alasan ilmiah, dan mengaitkan koreksi dengan isi jawaban siswa. Sebaliknya, DeepSeek V4 tetap mampu menghasilkan umpan balik yang jelas dan relevan, tetapi kualitasnya lebih bervariasi pada ketepatan konsep dan kedalaman uraian.

Perbandingan Berdasarkan Nomor Soal dan Level Kognitif

Perbandingan berdasarkan nomor soal menunjukkan bahwa ChatGPT 5.5 memperoleh mean lebih tinggi pada Soal 2, Soal 3, Soal 5, dan Soal 6. DeepSeek V4 sedikit lebih tinggi pada Soal 4, sedangkan pada Soal 1 kedua model memperoleh mean yang sama. Ringkasan perbandingan berdasarkan level kognitif disajikan pada Tabel 4.

Tabel 4. Perbandingan Kualitas Umpan Balik Berdasarkan Level Kognitif Soal

Level Kognitif	Nomor Soal	Mean ChatGPT 5.5	Mean DeepSeek V4	Selisih Mean	Arah Perbedaan
C4	1, 2, 3	15,50	14,85	0,65	ChatGPT 5.5 lebih tinggi
C5	4	13,85	13,90	-0,05	DeepSeek V4 sedikit lebih tinggi
C6	5, 6	15,75	14,83	0,92	ChatGPT 5.5 lebih tinggi

Berdasarkan Tabel 4, ChatGPT 5.5 memperoleh mean yang lebih tinggi pada soal dengan level kognitif C4 dan C6, sedangkan DeepSeek V4 memperoleh mean yang sedikit lebih tinggi pada level C5. Selisih mean terbesar terdapat pada level C6, yaitu sebesar 0,92, sedangkan pada level C5 selisih mean hanya sebesar 0,05. Temuan ini menunjukkan adanya perbedaan pola skor rata-rata antara kedua model pada berbagai level kognitif. Secara deskriptif, ChatGPT 5.5 cenderung memperoleh skor yang lebih tinggi pada soal yang menuntut analisis dan penyusunan solusi yang lebih kompleks, sedangkan pada level evaluatif (C5) kedua model menunjukkan capaian yang relatif berdekatan.

Temuan tersebut mengindikasikan bahwa tuntutan kognitif soal berpotensi memengaruhi kualitas umpan balik yang dihasilkan AI. Soal pada level C4 dan C6 memberikan ruang yang lebih besar bagi model untuk menguraikan konsep, menjelaskan hubungan antarkonsep, serta memberikan arahan perbaikan secara sistematis. Sebaliknya, soal pada level C5 menuntut kemampuan mengevaluasi argumen dan memberikan pertimbangan konseptual terhadap suatu pendapat, sehingga memerlukan bentuk respons yang berbeda. Oleh karena itu, analisis berdasarkan level kognitif dalam penelitian ini digunakan sebagai analisis deskriptif tambahan untuk mengidentifikasi pola kecenderungan performa ChatGPT 5.5 dan DeepSeek V4 pada

berbagai tuntutan kognitif soal, sedangkan pengujian inferensial tetap difokuskan pada indikator kualitas umpan balik dan total skor yang menjadi fokus utama penelitian.

Hasil Uji Normalitas dan Wilcoxon Signed-Rank Test

Uji normalitas dilakukan terhadap selisih pasangan skor ChatGPT 5.5 dan DeepSeek V4 pada setiap indikator dan total skor. Seluruh variabel selisih memperoleh nilai signifikansi Kolmogorov- Smirnov lebih kecil dari 0,001, sehingga data dinyatakan tidak berdistribusi normal. Oleh karena itu, perbandingan kualitas umpan balik kedua model dianalisis menggunakan Wilcoxon Signed-Rank Test.

Tabel 5. Hasil Uji Wilcoxon Signed-Rank Test

Aspek yang Diuji	Z	Asymp. Sig.	Keputusan	Keterangan
Akurasi ilmiah	-3,15	0,002	Signifikan	ChatGPT 5.5 lebih tinggi
Kedalaman penjelasan	-5,00	<0,001	Signifikan	ChatGPT 5.5 lebih tinggi
Relevansi terhadap konteks jawaban siswa	-3,36	<0,001	Signifikan	ChatGPT 5.5 lebih tinggi
Kejelasan bahasa	0,00	1,000	Tidak signifikan	Tidak berbeda
Total skor	-4,82	<0,001	Signifikan	ChatGPT 5.5 lebih tinggi

Hasil uji Wilcoxon menunjukkan bahwa terdapat perbedaan signifikan pada akurasi ilmiah, kedalaman penjelasan, relevansi terhadap konteks jawaban siswa, dan total skor. Pada keempat aspek tersebut, ChatGPT 5.5 memperoleh skor lebih tinggi daripada DeepSeek V4. Sementara itu, indikator kejelasan bahasa tidak menunjukkan perbedaan signifikan karena nilai signifikansi sebesar 1,000 lebih besar dari 0,05.

Selain signifikansi statistik, penelitian ini juga menghitung effect size menggunakan koefisien r untuk mengukur besarnya perbedaan antara ChatGPT 5.5 dan DeepSeek V4. Hasil perhitungan menunjukkan bahwa perbedaan pada indikator akurasi ilmiah memiliki effect size sebesar 0,29 (kecil–sedang), kedalaman penjelasan sebesar 0,46 (sedang), relevansi terhadap konteks jawaban siswa sebesar 0,31 (sedang), dan total skor sebesar 0,44 (sedang). Sementara itu, kejelasan bahasa memiliki effect size sebesar 0,00 yang menunjukkan tidak adanya perbedaan praktis antara kedua model pada aspek tersebut. Temuan ini menunjukkan bahwa perbedaan yang ditemukan tidak hanya signifikan secara statistik, tetapi juga memiliki makna praktis terutama pada aspek kedalaman penjelasan, relevansi terhadap konteks jawaban siswa, dan kualitas umpan balik secara keseluruhan.

Hasil ini memperkuat temuan deskriptif bahwa perbedaan utama kedua model bukan terletak pada keterbacaan atau kelancaran bahasa. Keduanya sama-sama mampu menyampaikan umpan balik dengan bahasa yang jelas. Perbedaan yang lebih menentukan justru berada pada substansi, yaitu kemampuan menjaga akurasi konsep, menyusun penjelasan yang mendalam, dan menyesuaikan koreksi dengan isi jawaban siswa.

Pembahasan Substantif

Tabel 6. Contoh Perbandingan Umpan Balik ChatGPT 5.5 dan DeepSeek V4 pada Berbagai Level Kognitif

Kasus	Level	Temuan Utama
UB-318	C4	Kedua model sama-sama mampu mengarahkan siswa untuk melengkapi proses perhitungan dan memperoleh kualitas umpan balik yang setara.
UB-420	C5	DeepSeek V4 lebih tepat dalam membaca inti evaluasi yang disampaikan siswa sehingga memperoleh skor lebih tinggi pada kasus ini.
UB-617	C6	ChatGPT 5.5 lebih mampu menghubungkan kesalahan teknis dengan konsekuensi konseptual sehingga menghasilkan umpan balik yang lebih mendalam.

Analisis kualitatif pada Tabel 6 menunjukkan bahwa perbedaan kualitas umpan balik antara ChatGPT 5.5 dan DeepSeek V4 terutama muncul pada jawaban yang memerlukan evaluasi dan koreksi konseptual yang lebih kompleks. ChatGPT 5.5 cenderung memberikan penjelasan yang lebih mendalam dengan mengaitkan kesalahan siswa pada konsep fisika yang relevan, sehingga koreksi yang diberikan tidak hanya bersifat informatif tetapi juga membantu memperkuat pemahaman konseptual. Sebaliknya, DeepSeek V4 umumnya mampu menghasilkan umpan balik yang jelas, runtut, dan mudah dipahami, namun pada beberapa kasus penjelasan yang diberikan relatif lebih ringkas sehingga elaborasi konseptual yang muncul tidak sedalam ChatGPT 5.5. Temuan ini selaras dengan hasil kuantitatif yang menunjukkan skor ChatGPT 5.5 lebih tinggi pada akurasi ilmiah, kedalaman penjelasan, relevansi terhadap konteks jawaban siswa, dan total skor, sementara kedua model menunjukkan kualitas yang relatif setara pada aspek kejelasan bahasa.

Temuan ini sejalan dengan gagasan bahwa kualitas umpan balik tidak dapat direduksi menjadi aspek kebahasaan. Boggs dan Manchón (2023) menekankan bahwa umpan balik efektif harus menunjukkan bagian yang sudah benar, bagian yang perlu diperbaiki, dan langkah perbaikan yang dapat dilakukan. Dalam pembelajaran fisika, syarat tersebut menjadi lebih ketat karena koreksi harus tepat secara ilmiah. Umpan balik yang jelas tetapi kurang akurat dapat berisiko mempertahankan miskonsepsi atau memberikan pemahaman yang tidak lengkap kepada siswa.

Analisis kualitatif terhadap contoh umpan balik memperlihatkan bahwa perbedaan kedua model terutama muncul ketika jawaban siswa memerlukan koreksi konseptual yang presisi. Pada kasus jawaban yang sudah benar tetapi belum menunjukkan proses perhitungan, kedua model dapat memberikan umpan balik yang kuat. Namun, pada kasus yang memuat kesalahan satuan atau penalaran evaluatif, perbedaan kualitas menjadi lebih terlihat. ChatGPT 5.5 cenderung lebih kuat dalam menghubungkan kesalahan teknis dengan konsekuensi konseptual, sedangkan DeepSeek V4 pada beberapa kasus lebih terbatas pada koreksi langsung.

Implikasi penelitian ini menunjukkan bahwa kualitas AI dalam pembelajaran fisika perlu dievaluasi berdasarkan kemampuannya memberikan umpan balik yang akurat, mendalam, dan sesuai dengan jawaban siswa, bukan hanya berdasarkan kejelasan bahasa yang dihasilkan. Temuan ini menunjukkan bahwa model AI yang tampak serupa dari sisi kebahasaan dapat memiliki kualitas pedagogis yang berbeda. Oleh karena itu, guru perlu lebih kritis dalam memilih dan memanfaatkan AI sebagai pendukung pemberian umpan balik, terutama pada materi fisika yang menuntut ketepatan konsep dan penalaran ilmiah.

IV. KESIMPULAN

Penelitian ini menunjukkan bahwa ChatGPT 5.5 dan DeepSeek V4 sama-sama mampu menghasilkan umpan balik berkualitas tinggi terhadap jawaban kontekstual siswa pada konsep optik fisika. Namun, ChatGPT 5.5 cenderung menghasilkan umpan balik yang lebih akurat, lebih mendalam, dan lebih relevan terhadap jawaban siswa dibandingkan DeepSeek V4. Temuan ini

menunjukkan bahwa perbedaan kualitas AI dalam konteks pembelajaran fisika lebih ditentukan oleh substansi isi umpan balik daripada aspek kebahasaan.

Penelitian ini terbatas pada satu materi fisika, yaitu optik, serta melibatkan dua model AI yang dievaluasi menggunakan seperangkat soal kontekstual tertentu. Oleh karena itu, penelitian selanjutnya dapat memperluas cakupan materi, melibatkan lebih banyak model AI generatif, serta mengeksplorasi pengaruh karakteristik soal dan strategi prompting terhadap kualitas umpan balik yang dihasilkan AI dalam konteks pembelajaran sains.

DAFTAR PUSTAKA

Boggs, J. A., & Manchón, R. M. (2023). Feedback literacy in writing research and teaching: Advancing L2 WCF research agendas. Assessing Writing, 58. https://doi.org/10.1016/j.asw.2023.100786

Carless, D. (2022). Feedback for student learning in higher education. International Encyclopedia of Education: Fourth Edition, 623-629. https://doi.org/10.1016/B978-0-12-818630-5.14066-7

Creswell, J. W., & Plano Clark, V. L. (2018). Designing and Conducting Mixed Methods Research (3rd ed.). SAGE Publications.

Etaiwi, W., & Alhijawi, B. (2025). Comparative evaluation of ChatGPT and DeepSeek across key NLP tasks: Strengths, weaknesses, and domain-specific performance. Array, 27. https://doi.org/10.1016/j.array.2025.100478

Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., Zhang, X., Yu, X.,

Wu, Y., Wu, Z. F., Gou, Z., Shao, Z., Li, Z., Gao, Z., Liu, A., et al. (2025). DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning.

Jin, I., Tangsrivimol, J. A., Darzi, E., Hassan Virk, H. U., Wang, Z., Egger, J., Hacking, S., Glicksberg, B. S., Strauss, M., & Krittanawong, C. (2025). DeepSeek vs. ChatGPT: Prospects and challenges. Frontiers in Artificial Intelligence, 8, 1-12. https://doi.org/10.3389/frai.2025.1576992

Kaltakci-Gurel, D. (2023). Exploring pre-service teachers' conceptual understanding and confidence in geometrical optics: A focus on gender and prior course achievement. Education Sciences, 13(5). https://doi.org/10.3390/educsci13050452

Kortemeyer, G. (2023). Could an artificial-intelligence agent pass an introductory physics course? Physical Review Physics Education Research, 19(1). https://doi.org/10.1103/PhysRevPhysEducRes.19.010132

Kunaedi, J., Samsudin, A., Hasanah, L., Hidiana Aminudin, A., Herliyana Dewi, F., Alfionita Umar, F., Rona Wahyu Astuti, I., & Nurul Mufida, S. (2024). Developing diagnostic test on geometrical optics (DT-GO) concept. KnE Social Sciences. https://doi.org/10.18502/kss.v9i8.15637

Lo, C. K. (2023). What is the impact of ChatGPT on education? A rapid review of the literature. Education Sciences, 13(4). https://doi.org/10.3390/educsci13040410

Panadero, E., & Lipnevich, A. A. (2022). A review of feedback models and typologies: Towards an integrative model of feedback elements. Educational Research Review, 35, 100416. https://doi.org/10.1016/j.edurev.2021.100416

Perera, P., & Lankathilaka, M. (2023). AI in higher education: A literature review of ChatGPT and guidelines for responsible implementation. International Journal of Research and Innovation in Social Science, VII(VI), 306-314. https://doi.org/10.47772/ijriss.2023.7623

Polverini, G., & Gregorcic, B. (2024). How understanding large language models can inform the use of ChatGPT in physics education. European Journal of Physics, 45(2), 1-33. https://doi.org/10.1088/1361- 6404/ad1420

Wisniewski, B., Zierer, K., & Hattie, J. (2020). The power of feedback revisited: A meta-analysis of educational feedback research. Frontiers in Psychology, 10, 1-14. https://doi.org/10.3389/fpsyg.2019.03087