JPFS 9 (2) 2026 374-392

Jurnal Pendidikan Fisika dan Sains (JPFS)

Journal homepage: http:// journal.unucirebon.ac.id/index.php/jpfs

Representasi Berbasis Kecerdasan Artifisial dalam Pembelajaran Fisika: *Systematic Literature Review*

Fitria Arifiyanti*[1], Soeharto[2]

[1]Pendidikan Fisika, FPMIPA, Universitas Pendidikan Indonesia
[2]Pusat Riset Pendidikan, Badan Riset dan Inovasi Nasional

*E-mail: fitriaarifiyanti@upi.edu

DOI: https://doi.org/10.52188/jpfs.v9i2.2566

Accepted: 23 Oktober 2026 Approved: 21 November 2026 Published: 23 November 2026

ABSTRAK

Kecerdasan artifisial (artificial intelligence/AI) generatif kini mampu membaca, menafsirkan, dan menghasilkan representasi fisika seperti grafik, diagram benda bebas, simulasi, dan visualisasi realitas tertambah, tetapi belum ada kajian sistematis yang menempatkan representasi sebagai fokus analisis. Kajian ini bertujuan memetakan peran, kinerja, dan dampak AI terhadap representasi dalam pembelajaran fisika. Kajian literatur sistematis dilakukan mengikuti pedoman Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 terhadap 15 artikel bereputasi dari basis data Scopus yang terbit pada 2016–2026, kemudian dianalisis melalui analisis isi deduktif dan induktif. Hasil kajian menunjukkan empat peran AI, yaitu sebagai pembuat, penafsir, mediator, dan penilai representasi. Temuan utama adalah kesenjangan antara penalaran verbal yang relatif kuat dan interpretasi visual yang lemah: pada satu studi benchmark TUG-K (60 kali pengerjaan ChatGPT-4), penalaran benar muncul pada 69,7% respons, tetapi pembacaan grafik yang benar hanya 30,9%. Pada studi benchmark terbesar yang menguji 15 model multimodal, akurasi soal bergambar berkisar antara 21,0% dan 81,5%, sedangkan dua studi evaluasi diagram tidak menemukan satu pun diagram benda bebas buatan AI yang sepenuhnya benar. Representasi menjadi lebih andal ketika keluaran AI dibatasi oleh kerangka fisika atau dimediasi guru. Bukti dampak pada hasil belajar masih terbatas pada dua studi berskala kecil karena sebagian besar studi menguji kemampuan AI, bukan efektivitas pembelajaran. Kajian ini merekomendasikan pergeseran riset menuju pengembangan kompetensi representasional dan kemampuan kritis siswa terhadap keluaran AI.

Kata kunci: Kecerdasan Artifisial Generatif; Kompetensi Representasional; Multirepresentasi; Pendidikan Fisika; Kajian Literatur Sistematis

ABSTRACT

Generative artificial intelligence (AI) can now read, interpret, and generate physics representations such as graphs, free-body diagrams, simulations, and augmented-reality visualizations, yet no systematic review has treated representation as its unit of analysis. This review maps the roles, performance, and impact of AI on representations in physics learning. Following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 guideline, 15 peer-reviewed Scopus-indexed articles published in 2016–2026 were analyzed through deductive and inductive content analysis. Four roles of AI emerged: generator, interpreter, mediator, and assessor of representations. The central finding is a gap between relatively strong verbal reasoning and weak visual interpretation: in a TUG-K benchmark study (60 ChatGPT-4 runs), correct reasoning appeared in 69.7% of responses, whereas correct graph reading appeared in only 30.9%. In the largest benchmark of 15 multimodal models, accuracy on image-based items ranged from 21.0% to 81.5%, and two diagram-evaluation studies found no fully correct AI-generated free-body diagram. Representations were more reliable when AI output was constrained by a physics framework or mediated by teachers. Evidence of impact on learning outcomes remains limited to two small-scale studies, as most studies tested AI capability rather tahan

learning effectiveness. This review recommends shifting research toward developing students' representational competence and their critical evaluation of AI output.

Keywords: generative artificial intelligence; representational competence; multiple representations; physics education; systematic literature review

@2026 Pendidikan Fisika FKIP Universitas Nahdlatul Ulama Cirebon

PENDAHULUAN

Fisika adalah disiplin yang bekerja melalui representasi. Sebuah balok di bidang miring diterjemahkan menjadi diagram benda bebas, gerak sebuah mobil dinyatakan sebagai grafik posisi–waktu, dan medan magnet digambarkan dengan garis-garis gaya. Van Heuvelen (1991) menegaskan bahwa belajar berpikir seperti fisikawan berarti belajar bergerak lincah di antara representasi verbal, piktorial, grafis, dan matematis. Kemampuan ini kemudian dikenal sebagai kompetensi representasional (Kozma & Russell, 1997; Kohl & Finkelstein, 2005). Kerangka DeFT dari Ainsworth (2006) menjelaskan bahwa multirepresentasi dapat berfungsi melengkapi informasi, membatasi interpretasi, dan membangun pemahaman yang lebih dalam, asalkan siswa mampu menerjemahkan dan mengoordinasikan representasi tersebut. Di sisi lain, penelitian pendidikan fisika selama tiga dekade menunjukkan bahwa siswa kerap memaknai grafik sebagai gambar harfiah, mengacaukan tinggi grafik dengan kemiringannya, dan menggambar gaya pada titik yang keliru (Beichner, 1994; Rosengrant, Van Heuvelen, & Etkina, 2009).

Sejak akhir 2022, kehadiran ChatGPT dan model sejenis mengubah lanskap pendidikan secara cepat (Kasneci et al., 2023). Perubahan yang paling relevan bagi fisika terjadi ketika model bahasa memperoleh kemampuan multimodal. Pada Oktober 2023, ChatGPT-4 mulai dapat menerima masukan gambar (Polverini & Gregorcic, 2024). Generasi berikutnya, seperti GPT-4o, GPT-5, Gemini, dan Claude, tidak hanya menafsirkan gambar tetapi juga menghasilkan diagram, menulis kode simulasi, dan memberi umpan balik atas grafik buatan siswa. Artinya, kecerdasan artifisial (*artificial intelligence*, AI) kini tidak lagi sekadar menjawab soal; AI ikut memproduksi dan menafsirkan representasi yang menjadi bahan baku penalaran fisika. Pergeseran ini membawa dua kemungkinan yang berlawanan. AI dapat memperluas akses terhadap visualisasi berkualitas, misalnya bagi guru tanpa perangkat lunak berbayar (Ramkorun, 2024; Mortimer, 2026). Namun AI juga dapat menyebarkan miskonsepsi melalui diagram yang tampak meyakinkan tetapi keliru secara fisis (Fulminar, 2026a, 2026b).

Kajian sistematis mengenai AI dalam pendidikan sudah cukup banyak, tetapi umumnya berfokus pada pendidikan tinggi secara umum (Zawacki-Richter, Marín, Bond, & Gouverneur, 2019) pada penggunaan ChatGPT lintas jenjang (Zhang & Tur, 2024), atau pada AI di pendidikan K-12 secara umum (Martin, Zhuang, & Schaefer, 2024). Kajian-kajian tersebut membahas peluang, ancaman, dan integritas akademik, tetapi belum menempatkan **representasi** sebagai unit analisis. Padahal, dalam fisika, persoalannya bukan hanya apakah AI menjawab benar, melainkan apakah AI "melihat" grafik dengan benar, menggambar gaya pada titik tangkap yang tepat, dan membantu siswa menghubungkan representasi yang berbeda. Di bidang fisika, kajian yang tersedia umumnya berupa tinjauan naratif tentang kemampuan AI mengerjakan soal dan ujian fisika dari jenjang sekolah hingga universitas (Yeadon & Hardy, 2024) atau studi tunggal tentang kinerja AI pada mata kuliah fisika dan dalam penyusunan tugas fisika oleh calon guru (Kortemeyer, 2023; Küchemann et al., 2023). Untuk memastikan kesenjangan tersebut, penulis melakukan penelusuran pendahuluan pada Scopus dengan menggabungkan kata kunci "review" atau "systematic review", "artificial intelligence" atau "ChatGPT", dan "physics education" untuk terbitan 2016–2026. Penelusuran itu menemukan kajian AI dalam pendidikan secara umum dan beberapa tinjauan naratif di bidang fisika, tetapi tidak menemukan kajian sistematis yang menjadikan representasi sebagai unit analisis. Dengan demikian, kesenjangan yang dijawab artikel ini tidak hanya terletak pada belum adanya pemetaan peran AI, tetapi juga pada belum adanya sintesis tentang bagaimana peran-peran tersebut mendukung atau menghambat kompetensi representasional siswa, yaitu kemampuan menerjemahkan, mengoordinasikan, dan mengevaluasi representasi fisika.

Dengan mengikuti kerangka *Preferred Reporting Items for Systematic Reviews and Meta-Analyses* (PRISMA) 2020 (Page et al., 2021), kajian ini dipandu oleh empat pertanyaan penelitian. Pertanyaan pertama (RQ1) menanyakan bagaimana karakteristik studi tentang representasi berbasis AI dalam pembelajaran fisika, dilihat dari tahun, sumber publikasi, negara, jenjang pendidikan, desain

penelitian, dan sistem AI yang digunakan. Pertanyaan kedua (RQ2) menanyakan jenis representasi fisika apa yang dihasilkan, ditafsirkan, atau dimediasi oleh AI, dan pada topik fisika apa. Pertanyaan ketiga (RQ3) menanyakan bagaimana kinerja AI dalam menangani representasi fisika serta dampaknya terhadap hasil dan proses belajar siswa. Pertanyaan keempat (RQ4) menanyakan keterbatasan, tantangan, dan arah penelitian selanjutnya bagi pemanfaatan representasi berbasis AI dalam pembelajaran fisika.

Kontribusi artikel ini ada tiga. Pertama, artikel ini menawarkan tipologi empat peran AI terhadap representasi fisika, yaitu pembuat, penafsir, mediator, dan penilai, sebagai lensa analisis yang dikaitkan dengan fungsi multirepresentasi dalam kerangka DeFT, sehingga setiap peran dapat dibaca dari sisi kontribusi dan hambatannya terhadap kompetensi representasional siswa. Kedua, artikel ini mensintesis bukti kuantitatif mengenai kesenjangan antara penalaran verbal dan kemampuan visual AI. Ketiga, artikel ini menyajikan analisis kritis terhadap kualitas metodologis studi primer berbasis MMAT 2018 beserta agenda riset yang dapat ditindaklanjuti oleh peneliti dan guru fisika, termasuk di Indonesia. Dengan demikian, kebaruan kajian ini terletak pada pergeseran pertanyaan dari "seberapa akurat AI menangani representasi fisika" menjadi "bagaimana AI dapat mendukung atau justru menghambat perkembangan kompetensi representasional siswa".

METODE

Penelitian ini merupakan *systematic literature review* (SLR) dengan pendekatan kualitatif yang dilengkapi deskripsi kuantitatif sederhana. SLR dipilih karena mampu mensintesis temuan dari berbagai studi secara transparan dan dapat direplikasi (Kitchenham, 2004). Kajian ini disusun dan dilaporkan berdasarkan PRISMA 2020 (Page et al., 2021). Sesuai pedoman tersebut, bagian metode ini melaporkan kriteria kelayakan, sumber informasi, strategi pencarian, proses seleksi, proses pengumpulan data, penilaian kualitas studi, dan metode sintesis. Alur kerjanya mengadaptasi tahapan yang digunakan Zhang dan Tur (2024), yaitu merumuskan pertanyaan penelitian, melakukan penelusuran awal, menetapkan strategi pencarian, menetapkan kriteria kelayakan, mencari pada basis data, menyaring judul dan abstrak, menelusuri dan menilai teks lengkap, mengekstraksi data dan menilai kualitas, lalu mensintesis temuan.

Strategi Pencarian

Pencarian dilakukan dengan strategi Boolean yang menggabungkan tiga kelompok kata kunci, yaitu teknologi AI, representasi, dan pendidikan fisika. String pencarian yang digunakan adalah sebagai berikut.

("artificial intelligence" OR "generative AI" OR "ChatGPT" OR "large language model" OR "multimodal large language model") AND ("representation" OR "graph" OR "diagram" OR "visualization" OR "simulation" OR "free-body diagram") AND ("physics education" OR "physics learning" OR "physics teaching")

Penelusuran dilakukan pada basis data Scopus pada 30 September 2026 dengan menerapkan string di atas pada kolom TITLE-ABS-KEY. Scopus dipilih karena mengindeks jurnal bereputasi di bidang fisika dan pendidikan fisika secara luas, menyediakan fitur penyaringan yang terstandar, dan memenuhi syarat sebagai sistem pencarian utama untuk kajian sistematis (Gusenbauer & Haddaway, 2020). Hasil pencarian kemudian dibatasi dengan filter bidang subjek *Social Sciences*, jenis dokumen artikel, bahasa Inggris, jenis sumber jurnal, tahap publikasi final, dan rentang tahun 2016–2026. Rentang sepuluh tahun dipilih agar kajian juga dapat menangkap studi AI dalam pembelajaran fisika sebelum hadirnya model bahasa generatif. Filter diterapkan secara berurutan melalui menu Refine results pada Scopus, yaitu Subject area = Social Sciences, Document type = Article, Language = English, Source type = Journal, Publication stage = Final, dan Publication year = 2016–2026. Bidang Social Sciences dipilih karena kajian ini berfokus pada dimensi pedagogis representasi berbasis AI, yaitu bagaimana AI mendukung proses belajar, pembelajaran, dan asesmen fisika, bukan pada pengembangan teknis model AI atau riset fisika murni. Dalam sistem klasifikasi Scopus (All Science Journal Classification), kategori Education berada di bawah bidang Social Sciences, dan jurnal pendidikan fisika utama, seperti Physical Review Physics Education Research, Physics Education, dan The Physics Teacher, terklasifikasi dalam kategori tersebut. Filter ini sekaligus menyisihkan rekaman yang memuat kata kunci AI dan fisika tetapi berasal dari riset teknis di luar konteks pendidikan. Penulis menyadari bahwa pembatasan ini berpotensi tidak menjaring artikel pendidikan fisika yang terbit di jurnal yang hanya diklasifikasikan pada bidang Physics and Astronomy; konsekuensi tersebut dibahas pada bagian Keterbatasan Kajian.

Kriteria Kelayakan

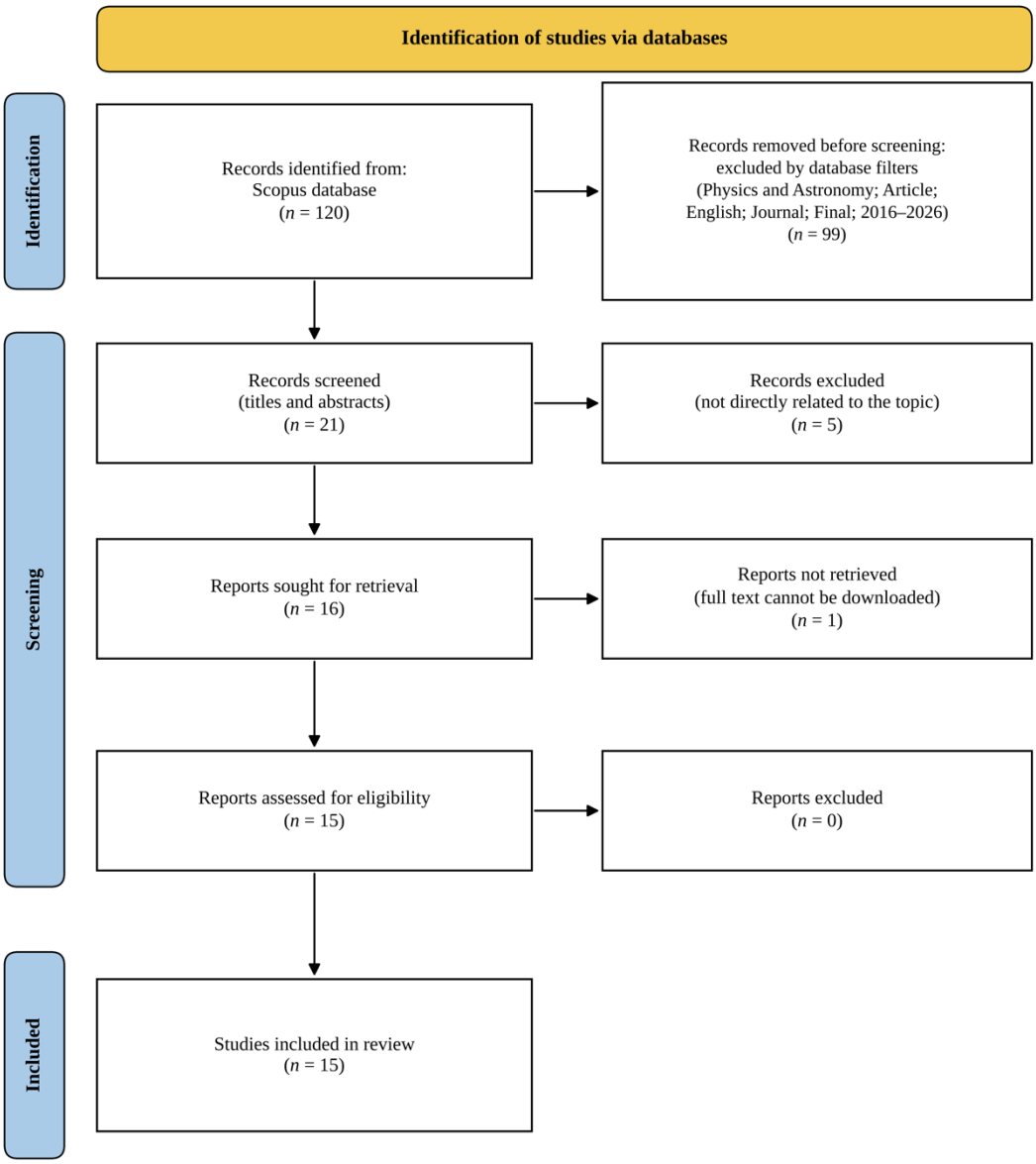
Kriteria inklusi dan eksklusi disusun berdasarkan delapan aspek, yaitu waktu terbit, bidang subjek, jenis publikasi, bahasa, fokus AI, fokus representasi, konteks, dan ketersediaan teks lengkap. Rincian kriteria tersebut disajikan pada Tabel 1.

Tabel 1. Kriteria Inklusi dan Eksklusi

Aspek	Kriteria inklusi	Kriteria eksklusi
Waktu terbit	2016–2026	Sebelum 2016
Bidang subjek	Social Sciences (klasifikasi Scopus)	Bidang subjek lain
Jenis publikasi	Artikel (document type: article) pada jurnal (source type: journal) dengan tahap publikasi final	Prosiding, bab buku, artikel review, editorial, artikel in press
Bahasa	Bahasa Inggris	Selain bahasa Inggris
Fokus AI	AI generatif, LLM, atau MLLM	AI non-generatif tanpa komponen bahasa atau visual
Fokus representasi	AI menghasilkan, menafsirkan, memediasi, atau menilai representasi fisika	AI hanya menjawab soal teks tanpa unsur representasi
Konteks	Pembelajaran atau asesmen fisika di sekolah menengah atau perguruan tinggi	Bidang selain fisika
Ketersediaan	Teks lengkap dapat diakses	Teks lengkap tidak tersedia

Seleksi Studi

Proses seleksi studi dilaporkan menggunakan diagram alir PRISMA 2020 untuk kajian sistematis baru yang hanya menelusuri basis data (Gambar 1). Pada tahap identifikasi, penelusuran di Scopus menghasilkan 120 rekaman (*records*). Sebanyak 99 rekaman dikeluarkan sebelum penyaringan melalui filter basis data, yaitu bidang subjek, jenis dokumen, bahasa, jenis sumber, tahap publikasi, dan tahun terbit. Penghapusan duplikasi tidak diperlukan karena hanya satu basis data yang digunakan. Pada tahap penyaringan, judul dan abstrak dari 21 rekaman ditelaah berdasarkan kriteria pada Tabel 1, dan 5 rekaman dikeluarkan karena tidak berada dalam konteks pembelajaran atau asesmen fisika. Kelima rekaman tersebut menggunakan AI generatif untuk keperluan riset atau aplikasi di bidang lain, yaitu perancangan pelarut baterai (n = 1), penemuan dan karakterisasi material kristalin (n = 1), pemantauan banjir berbasis geoinformatika (n = 1), simulasi pertanian presisi (n = 1), dan visualisasi catatan aurora historis (n = 1). Rekaman-rekaman ini terjaring karena memuat kata kunci AI generatif, visualisasi atau grafik, dan fisika, tetapi tidak memenuhi kriteria konteks pada Tabel 1. Sebanyak 16 laporan (*reports*) kemudian dicari teks lengkapnya, tetapi 1 laporan tidak dapat diperoleh karena teks lengkapnya tidak dapat diunduh. Sebanyak 15 laporan dinilai kelayakannya berdasarkan teks lengkap, dan seluruhnya memenuhi kriteria. Dengan demikian, pada tahap inklusi terdapat 15 studi yang dianalisis dalam sintesis ini dengan kode S1–S15.



Gambar 1. Diagram Alir PRISMA 2020 Proses Seleksi Studi

Ekstraksi Data dan Penilaian Kualitas

Data diekstraksi menggunakan lembar ekstraksi terstruktur yang memuat identitas artikel, negara, jenjang, desain penelitian, ukuran sampel, sistem AI dan versinya, jenis representasi, topik fisika, peran AI, instrumen, temuan utama, dan keterbatasan yang dinyatakan penulis. Kualitas metodologis setiap studi dinilai dengan *Mixed Methods Appraisal Tool* (MMAT) versi 2018 (Hong et al., 2018), yang sesuai untuk korpus berdesain campuran. Selain itu, penulis mencatat empat indikator yang relevan khusus untuk penelitian AI, yaitu pelaporan versi model, pelaporan *prompt*, jumlah pengulangan untuk mengatasi variabilitas keluaran, dan keberadaan penilai independen.

Ekstraksi data, pengodean, dan penilaian MMAT dilakukan oleh dua penulis secara independen. Sebelum pengodean penuh, kedua penilai melakukan kalibrasi pada tiga artikel untuk menyamakan pemahaman terhadap definisi setiap kategori. Setiap studi dinilai menggunakan kategori MMAT yang sesuai dengan desainnya, dan setiap kriteria diberi nilai ya, tidak, atau tidak dapat ditentukan. Mengikuti anjuran Hong et al. (2018), skor total tidak digunakan untuk mengeluarkan studi; jumlah kriteria yang terpenuhi digunakan untuk menggambarkan kekuatan bukti, yaitu tinggi (4–5 kriteria), sedang (3 kriteria), dan rendah ($\leq$ 2 kriteria). Studi berupa demonstrasi atau contoh kerja yang tidak memiliki pertanyaan penelitian empiris dan data yang dikumpulkan secara sistematis tidak lolos pertanyaan penyaringan MMAT, tetapi tetap dipertahankan sebagai bukti ilustratif dengan bobot paling

rendah dalam sintesis. Perbedaan penilaian antarpenilai dibahas hingga tercapai konsensus, dan bila konsensus tidak tercapai, penulis ketiga bertindak sebagai penengah (O'Connor & Joffe, 2020).

Analisis Data

Analisis dilakukan melalui analisis isi dalam dua tahap. Tahap deduktif menggunakan kategori yang diturunkan dari teori multirepresentasi (Ainsworth, 2006), yaitu jenis representasi dan fungsi representasi. Tahap induktif mengikuti analisis tematik Braun dan Clarke (2006) untuk menemukan pola peran AI, pola kesalahan, dan dampak pembelajaran. Dari tahap induktif ini muncul tipologi empat peran AI yang digunakan untuk mengorganisasi bagian Hasil. Data kuantitatif yang dilaporkan studi primer, seperti persentase akurasi dan skor tes, disajikan secara deskriptif. Sesuai butir metode sintesis pada PRISMA 2020, sintesis dilakukan secara naratif dan tematik; meta-analisis tidak dilakukan karena desain dan ukuran hasil antarstudi tidak sebanding. Karena itu, penilaian heterogenitas statistik, bias pelaporan, dan tingkat kepastian bukti secara kuantitatif juga tidak diterapkan.

Tipologi peran AI dikembangkan melalui tiga langkah. Pertama, setiap episode interaksi antara AI dan representasi dalam studi primer dikodekan secara terbuka berdasarkan tindakan AI terhadap representasi, misalnya menggambar diagram, menulis kode grafik, membaca grafik, memberi umpan balik, atau menilai laporan. Kedua, kode-kode tersebut dikelompokkan berdasarkan arah hubungan antara AI dan representasi, yaitu AI menghasilkan representasi baru (pembuat), AI menerima representasi sebagai masukan dan menarik makna darinya (penafsir), AI berada di antara siswa dan representasi untuk mendukung proses belajar (mediator), atau AI mengevaluasi representasi yang dibuat siswa (penilai). Ketiga, setiap studi dialokasikan pada peran yang paling dominan sesuai fokus pertanyaan penelitiannya, sedangkan peran sekunder dicatat secara terpisah. Proses ini dilakukan oleh dua penilai yang sama, dan perbedaan alokasi diselesaikan melalui diskusi.

HASIL

Karakteristik Studi (RQ1)

Meskipun rentang pencarian mencakup 2016–2026, seluruh 15 artikel yang memenuhi kriteria terbit dalam tiga tahun terakhir, yaitu dua artikel pada 2024, empat artikel pada 2025, dan sembilan artikel pada 2026. Tidak ada artikel sebelum 2024 yang lolos seleksi, yang menunjukkan bahwa riset representasi berbasis AI dalam pembelajaran fisika baru tumbuh setelah model bahasa generatif dan multimodal tersedia luas. Dari sisi sumber, *Physics Education* mendominasi dengan 10 artikel, diikuti *Physical Review Physics Education Research* (3 artikel), *European Journal of Physics* (1 artikel), dan *EPJ Quantum Technology* (1 artikel). Ringkasan seluruh studi disajikan pada Tabel 2.

Tabel 2. Ringkasan 15 Studi yang Dianalisis

Kode	Penulis (tahun), negara	Jenjang dan partisipan	Desain	Sistem AI	Representasi dan topik	Temuan kunci
S1	Fulminar (2026a), Filipina	Perguruan tinggi; 3 dosen penilai	Deskriptif-evaluatif	ChatGPT (GPT-5)	Diagram gerak dan diagram benda bebas; kinematika, dinamika	Tidak ada diagram bebas kesalahan; kesalahan meningkat seiring kompleksitas soal
S2	Gaurina & Štibi (2026), Kroasia	SMK; 22 siswa	Intervensi kelas pra–pasca tanpa kontrol	ChatGPT Team	Grafik x–t, v–t, a–t; kinematika	Rerata 2,91/13 menjadi 5,50/13; rutinitas pengecekan keluaran AI
S3	Ramkorun (2024), Amerika Serikat	Fisika dasar universitas	Demonstrasi pedagogis	ChatGPT 3.5	Skrip Python/Octave untuk grafik gerak 1-D	Grafik berhasil dibuat; perhitungan kemiringan dan luas keliru
S4	Paul et al. (2026), Amerika Serikat	Soal AP Physics 1 & 2 (2015–2025)	Benchmark, 3 penilai ahli	ChatGPT 4.1 mini, Gemini 2.5 Flash, Claude 4.0 Sonnet, DeepSeek R1	Diagram, grafik, rangkaian; mekanika hingga fisika modern	Skor 82–92%; galat pada diagram, grafik, arah vektor, kaidah tangan kanan
S5	Kahaleh & Lopez (2025), Lebanon	SMA; 12 dan 25 instruktur	Metode campuran dua fase	ChatGPT, Gemini, Copilot	Penjelasan dialogis; miskonsepsi gaya pada elevator	ChatGPT terbaik; semua model lemah dalam alat bantu visual
S6	Sachithanand (2026), Amerika Serikat	SMA; tanpa partisipan	Studi pilot, rubrik akurasi	GPT-4o-mini (terbatas) vs Claude Opus 4.6 (bebas)	Simulasi jatuh bebas dan gerak peluru; butir FCI	Sistem terbatas memenuhi semua kriteria; sistem bebas sekali gagal total
S7	Coban et al. (2025), Jerman	Universitas; 21 kelompok	Eksperimen dalam subjek + pelacakan mata	ChatGPT + AR (HoloLens 2)	Visualisasi AR kriptografi kuantum (BB84)	Umpan balik meningkatkan skor (d = 1,19) dan mengarahkan atensi visual
S8	Polverini et al. (2025), Swedia	Butir BEMA; pembanding	Benchmark kuantitatif-kualitatif	ChatGPT-4, ChatGPT-4o	Diagram rangkaian, medan, vektor; listrik-magnet	50,2% dan 59,8% vs mahasiswa 53,4%; kesulitan spasial

Kode	Penulis (tahun), negara	Jenjang dan partisipan	Desain	Sistem AI	Representasi dan topik	Temuan kunci
		12.214 mahasiswa				
S9	Polverini & Gregorcic (2025), Swedia	102 butir dari 4 inventori	Benchmark kinerja dan biaya	15 MLLM (OpenAI, Google, Anthropic)	Grafik, rangkaian, optika geometri, kuantum	Akurasi 21,0–81,5%; model mahal tidak selalu lebih baik
S10	Garg (2026), Uni Emirat Arab	SMA; 19 siswa	Studi kasus kelas	AI generatif (tidak disebutkan)	Laboratorium virtual difraksi, transformator, RLC	Pengembangan cepat dengan guru sebagai pengendali; persepsi positif (4,16–4,42)
S11	Mortimer (2026), Inggris	Guru KS4–5	Contoh kerja pengembangan	ChatGPT, DeepSeek	Antarmuka akuisisi data Arduino; osilasi pegas	Konstanta pegas statis dan dinamis konsisten (19,94 dan 20,44 N/m)
S12	Fulminar (2026b), Filipina	Perguruan tinggi; 3 dosen penilai	Deskriptif-evaluatif	ChatGPT (GPT-5), Gemini 3, Copilot	Diagram benda bebas; dinamika	Tidak satu pun diagram memenuhi standar
S13	Abreu et al. (2026), Uruguay	Universitas; 57 laporan praktikum	Perbandingan penilaian	GPT-5.4 melalui API	Laporan praktikum berisi tabel, persamaan, grafik	Korelasi peringkat lemah ($\rho = 0,38$); galat OCR persamaan dan grafik
S14	Klassen & Friege (2026), Jerman	Universitas; 11 pemula, 12 ahli	Pelacakan mata, think-aloud, uji AI	ChatGPT-4, DeepSeek-R1	Diagram ilmiah dan grafik	Ahli 86%, pemula 70%, ChatGPT-4 50%, DeepSeek-R1 10%
S15	Polverini & Gregorcic (2024), Swedia	Butir TUG-K; 60 kali pengerjaan	Benchmark kuantitatif-kualitatif	ChatGPT-4	Grafik kinematika	Skor 41,7%; penalaran benar 69,7%, interpretasi visual benar 30,9%

Secara geografis, studi tersebar di sembilan negara. Amerika Serikat dan Swedia masing-masing menyumbang tiga studi, Filipina dan Jerman masing-masing dua studi, serta Kroasia, Lebanon, Uni Emirat Arab, Inggris, dan Uruguay masing-masing satu studi. Tidak ada studi dari Asia Tenggara selain Filipina, termasuk Indonesia. Tiga studi Swedia berasal dari kelompok peneliti yang sama di Uppsala University (S8, S9, S15).

Dari sisi jenjang, enam studi berlatar sekolah menengah (S2, S4, S5, S6, S10, S11), enam studi berlatar perguruan tinggi (S1, S3, S7, S12, S13, S14), dan tiga studi berupa *benchmark* yang tidak terikat pada kelas tertentu (S8, S9, S15). Dari sisi desain, delapan studi mengevaluasi keluaran AI itu sendiri, baik melalui *benchmark* (S4, S8, S9, S15) maupun melalui penilaian ahli atau rubrik (S1, S5, S6, S12). Tiga studi bersifat desain dan demonstrasi (S3, S10, S11). Hanya dua studi yang mengukur hasil belajar siswa setelah intervensi (S2, S7), satu studi menguji AI sebagai penilai (S13), dan satu studi membandingkan AI dengan manusia ahli dan pemula (S14).

Dari sisi sistem AI, model OpenAI (ChatGPT dan variannya) digunakan pada 14 dari 15 studi, sedangkan satu studi (S10) tidak menyebutkan model yang dipakai. Enam studi membandingkan lebih dari satu model (S4, S5, S6, S9, S12, S14). Versi model yang diuji bergerak cepat dari ChatGPT 3.5 (S3) hingga GPT-5.4 (S13) dalam kurun dua tahun.

Hasil penilaian kualitas metodologis dengan MMAT 2018 disajikan pada Tabel 3. Lima studi tergolong berkekuatan bukti tinggi (S4, S8, S9, S13, S15), tiga studi sedang (S1, S7, S14), lima studi rendah (S2, S5, S6, S10, S12), dan dua studi (S3, S11) tidak lolos pertanyaan penyaringan MMAT karena berupa demonstrasi atau contoh kerja. Pola ini penting untuk membaca sintesis berikutnya: bukti terkuat berasal dari studi benchmark yang menguji kemampuan AI menafsirkan representasi, sedangkan bukti tentang dampak terhadap pembelajaran hanya berasal dari studi berkekuatan bukti sedang (S7) dan rendah (S2).

Tabel 3. Hasil Penilaian Kualitas Metodologis Studi Primer dengan MMAT 2018

Kode	Kategori MMAT	Kriteria terpenuhi (dari 5)	Catatan utama penilaian	Kekuatan bukti
S1	1. Kualitatif	3	Lima soal; konsensus tiga dosen tanpa koefisien reliabilitas	Sedang
S2	3. Kuantitatif non-acak	2	Pra–pasca tanpa kelompok kontrol; instrumen belum divalidasi setara	Rendah
S3	Tidak lolos penyaringan	–	Demonstrasi pedagogis tanpa pertanyaan penelitian empiris	Ilustratif
S4	4. Kuantitatif deskriptif	4	Tiga penilai independen (ICC 0,909–0,935); prompt tunggal	Tinggi
S5	5. Metode campuran	2	Inkonsistensi jumlah partisipan; skala tiga titik	Rendah
S6	4. Kuantitatif deskriptif	2	Studi pilot; penilaian oleh peneliti sendiri; soal terbatas	Rendah

Kode	Kategori MMAT	Kriteria terpenuhi (dari 5)	Catatan utama penilaian	Kekuatan bukti
S7	3. Kuantitatif non-acak	3	Desain dalam subjek; sampel kecil dan sukarela; sebagian $\alpha = 0{,}10$	Sedang
S8	4. Kuantitatif deskriptif	4	Pembanding 12.214 mahasiswa; analisis kualitatif galat	Tinggi
S9	4. Kuantitatif deskriptif	5	15 model; sepuluh kali pengulangan per butir; analisis biaya	Tinggi
S10	1. Kualitatif	2	Studi kasus; persepsi Likert tanpa ukuran hasil belajar	Rendah
S11	Tidak lolos penyaringan	–	Contoh kerja pengembangan tanpa data sistematis	Ilustratif
S12	1. Kualitatif	2	Satu soal; konsensus tanpa koefisien reliabilitas	Rendah
S13	4. Kuantitatif deskriptif	4	Rubrik tetap pada 57 laporan; galat bersumber dari OCR	Tinggi
S14	5. Metode campuran	3	Sampel kecil; kesesuaian modalitas model tidak dijelaskan	Sedang
S15	4. Kuantitatif deskriptif	5	60 kali pengerjaan; $\kappa = 0{,}809$ dan 0,967	Tinggi

Catatan. Kategori MMAT: 1 = kualitatif; 3 = kuantitatif non-acak; 4 = kuantitatif deskriptif; 5 = metode campuran. Kekuatan bukti: tinggi = 4–5 kriteria terpenuhi; sedang = 3 kriteria; rendah = ≤ 2 kriteria. Ilustratif = tidak lolos pertanyaan penyaringan MMAT.

Jenis Representasi dan Topik Fisika (RQ2)

Representasi yang paling sering dikaji adalah **grafik kinematika** (S2, S3, S4, S9, S15) dan **diagram gaya**, termasuk diagram benda bebas dan diagram gerak (S1, S4, S6, S12). Kelompok berikutnya adalah diagram listrik-magnet seperti rangkaian, garis medan, dan vektor arus (S4, S8, S9), simulasi interaktif dan laboratorium virtual (S6, S10), serta kode dan antarmuka data eksperimen (S3, S11). Representasi yang lebih jarang dikaji adalah visualisasi realitas tertambah (S7), representasi verbal-dialogis untuk mengatasi miskonsepsi (S5), laporan praktikum multimodal (S13), dan diagram ilmiah umum (S14).

Dari sisi topik, mekanika jelas dominan karena sepuluh studi menyentuh kinematika atau dinamika. Listrik-magnet muncul pada empat studi (S4, S8, S9, S10), gelombang dan optika pada dua studi (S9, S10), dan fisika kuantum pada dua studi (S7, S9). Termodinamika, fisika modern di luar kuantum, dan astrofisika hampir tidak tersentuh. Pola ini sejalan dengan ketersediaan instrumen baku seperti TUG-K (Beichner, 1994), FCI (Hestenes, Wells, & Swackhamer, 1992), dan BEMA (Ding, Chabay, Sherwood, & Beichner, 2006).

Jika jenis representasi dipetakan terhadap peran AI, terlihat pola yang lebih tajam. Representasi statis dan terstandar, seperti grafik kinematika dan diagram rangkaian, terutama dikaji dalam peran penafsir melalui benchmark (S4, S8, S9, S15), sedangkan diagram gaya terutama dikaji dalam peran pembuat (S1, S12). Representasi dinamis dan interaktif, seperti simulasi, laboratorium virtual, dan realitas tertambah, lebih banyak muncul pada peran pembuat dan mediator (S6, S7, S10). Artinya, kemampuan AI membaca representasi diuji pada representasi yang paling mapan secara instrumen dan dengan studi berkekuatan bukti tinggi, sedangkan kemampuan AI menghasilkan representasi dinamis lebih banyak dilaporkan melalui studi kasus dan demonstrasi berkekuatan bukti rendah. Distribusi topik juga menunjukkan bahwa topik di luar mekanika umumnya hanya muncul dalam benchmark (S8, S9), bukan dalam intervensi kelas, sehingga bukti dampak representasi berbasis AI terhadap pembelajaran listrik-magnet, optika, dan fisika kuantum baru berasal dari satu studi realitas tertambah (S7).

Analisis induktif terhadap cara AI berinteraksi dengan representasi menghasilkan empat peran yang diringkas pada Tabel 4. Pengelompokan didasarkan pada arah hubungan antara AI dan representasi sebagaimana dijelaskan pada bagian Analisis Data. Peran ini tidak saling meniadakan. Beberapa studi memuat lebih dari satu peran, misalnya S2 yang menempatkan AI sebagai mediator tetapi juga memuat episode AI sebagai penafsir grafik, S4 yang menguji AI sebagai penafsir sekaligus pembuat grafik, dan S5 yang menilai AI sebagai mediator dialog sekaligus mencatat kelemahannya dalam membuat alat bantu visual. Pada Tabel 4, setiap studi dicantumkan pada peran utamanya. Peran pembuat dan penafsir paling banyak dikaji, sedangkan bukti dampak belajar baru muncul pada peran mediator.

Tabel 4. Peran AI dalam Representasi pada Pembelajaran Fisika

No.	Peran AI	Fungsi Utama	Temuan Utama dan Keterbatasan	Studi
1	**AI sebagai pembuat representasi**	Menghasilkan diagram, grafik, kode, dan simulasi.	AI dapat mempercepat pembuatan berbagai bentuk representasi. Namun, keluaran visual dapat tampak meyakinkan meskipun masih	S1, S3, S6, S10, S11, S12

			mengandung kesalahan konseptual atau ketidaktepatan representasi.	
2	**AI sebagai penafsir representasi**	Membaca dan menafsirkan grafik, rangkaian, serta soal berbasis gambar.	Kemampuan penalaran AI relatif kuat, tetapi interpretasi visual masih menjadi kelemahan, terutama ketika representasi bersifat kompleks, ambigu, atau memerlukan pemahaman konteks fisika.	S4, S8, S9, S14, S15
3	**AI sebagai mediator pembelajaran**	Memberikan umpan balik, mendukung dialog pembelajaran, dan memfasilitasi rutinitas pengecekan.	AI dapat mendukung proses belajar secara efektif ketika guru tetap memegang kendali konseptual dan pedagogis. Perannya lebih tepat sebagai pendamping pembelajaran daripada pengganti keputusan guru.	S2, S5, S7
4	**AI sebagai penilai produk siswa**	Menilai laporan praktikum yang memuat tabel dan grafik.	Penilaian otomatis belum setara dengan penilaian dosen ($\rho = 0{,}38$), sehingga hasil evaluasi AI masih memerlukan verifikasi dan pertimbangan manusia.	S13

Catatan. S = kode studi yang dianalisis dalam SLR; ρ = koefisien korelasi.

Kinerja AI dan Dampaknya terhadap Pembelajaran (RQ3)

AI sebagai pembuat representasi

Enam studi menempatkan AI sebagai produsen representasi. Ketika GPT-5 diminta menggambar diagram untuk lima soal mekanika, panel tiga dosen menemukan bahwa tidak satu pun diagram bebas dari masalah representasional (S1). Kesalahan yang berulang meliputi vektor yang hilang atau salah arah, ketiadaan sistem koordinat dan konvensi tanda, gaya gesek statis yang berlawanan arah, gaya normal yang digambar lebih pendek daripada gaya berat padahal keduanya setara, serta penambahan objek yang tidak ada dalam soal. Kesalahan juga meningkat seiring kompleksitas soal. Studi lanjutan yang membandingkan ChatGPT (GPT-5), Gemini 3, dan Copilot untuk satu soal diagram benda bebas menunjukkan pola serupa, yaitu titik tangkap gaya keliru, arah gaya salah, dan besaran non-gaya ikut digambar sebagai gaya (S12). Copilot sedikit lebih koheren, tetapi tidak ada yang memenuhi standar konseptual.

Pola berbeda muncul ketika AI menulis kode yang menghasilkan representasi. ChatGPT 3.5 mampu menulis skrip Python dan GNU Octave untuk grafik jarak–waktu dan kecepatan–waktu, lengkap dengan garis singgung dan poligon untuk menghitung luas (S3). Namun, ketika diminta menghitung, model menghasilkan kecepatan 5 m/s alih-alih 6 m/s pada $t = 3$ s, serta luas 42,5 mil alih-alih 37 mil karena gagal mendeteksi titik potong grafik dengan sumbu waktu. Pada S11, guru tanpa pelatihan pemrograman membangun antarmuka grafis untuk eksperimen osilasi pegas berbasis Arduino menggunakan ChatGPT dan DeepSeek melalui *vibe coding*. Konstanta pegas statis ($19{,}94 \pm 0{,}13$ N/m) dan dinamis ($20{,}44 \pm 0{,}14$ N/m) yang dihasilkan konsisten satu sama lain. Perlu dicatat bahwa temuan pada peran pembuat sebagian besar berasal dari studi berkekuatan bukti rendah atau bersifat ilustratif (S3, S6, S10, S11, S12; Tabel 3), sehingga perlu dibaca dengan hati-hati.

Pada level simulasi, S6 membandingkan dua arsitektur. Pada sistem terbatas, GPT-4o-mini hanya mengekstraksi parameter soal FCI ke dalam format JSON, sedangkan simulasi fisisnya dijalankan oleh kerangka kode yang sudah dikunci. Pada sistem bebas, Claude Opus 4.6 menulis seluruh simulasi sendiri. Sistem terbatas memenuhi seluruh kriteria rubrik, sedangkan sistem bebas lebih bervariasi dan mengalami satu kegagalan total. Pada S10, guru menggunakan AI generatif sebagai mitra *prototyping* untuk laboratorium virtual difraksi celah tunggal, transformator, dan rangkaian RLC, sementara guru tetap memverifikasi isi fisika dan menyesuaikan alat bagi siswanya.

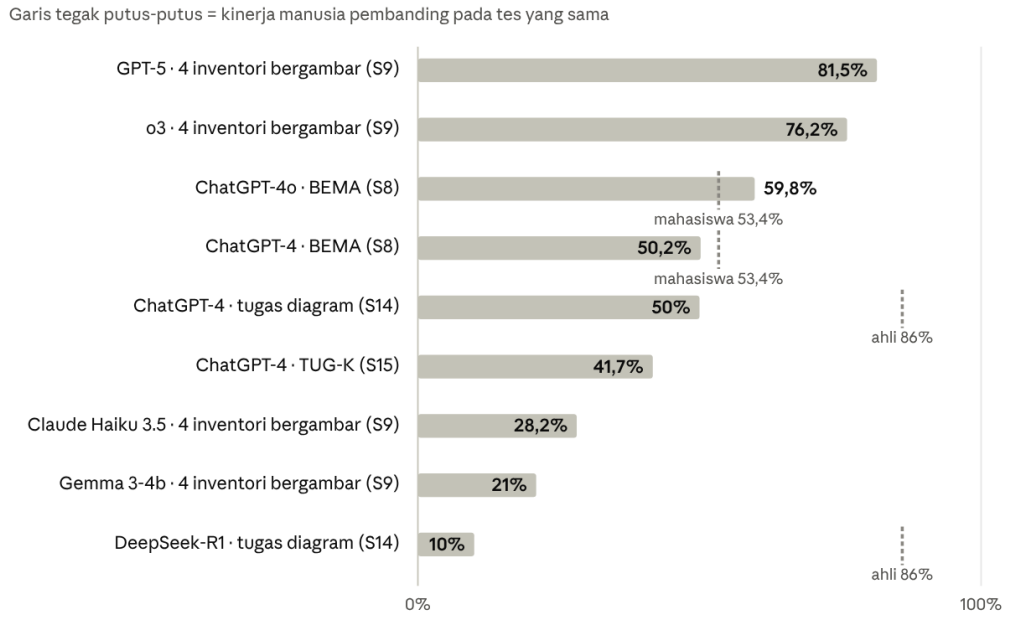
AI sebagai penafsir representasi

Lima studi menguji kemampuan AI membaca representasi visual. ChatGPT-4 yang dijalankan sebanyak 60 kali pada TUG-K memperoleh rerata 41,7% (S15). Penalaran yang benar muncul pada 69,7% respons, sedangkan interpretasi visual yang benar hanya muncul pada 30,9% respons. Di antara jawaban yang benar pun, 46,8% mengandung pembacaan grafik yang keliru. Pada butir menentukan kecepatan dari grafik posisi, kinerja ChatGPT-4 berada 34,4 poin persentase di bawah siswa.

Pada BEMA (S8), ChatGPT-4o (59,8%) mengungguli ChatGPT-4 (50,2%) dan rerata 12.214 mahasiswa (53,4%). Meski demikian, analisis kualitatif menemukan tiga jenis kesulitan, yaitu

interpretasi visual, pemilihan hukum fisika yang tepat, dan koordinasi spasial antara representasi dan prinsip fisika, dengan kaidah tangan kanan sebagai titik lemah yang paling menonjol. Pada soal uraian AP Physics (S4), empat model mencapai skor 82–92%, tetapi galat berulang muncul pada pembacaan diagram dan grafik, konstruksi grafik, arah vektor, topologi rangkaian, dan kaidah tangan kanan. Peringkat model juga tidak stabil dari tahun ke tahun.

Studi *benchmark* terbesar (S9) menguji 15 MLLM pada 102 butir dari empat inventori konsep. Akurasi berkisar antara 21,0% (Gemma 3-4b) dan 81,5% (GPT-5). Kinerja rata-rata pada BEMA (66,7%) dan TUG-K (64,4%) lebih tinggi daripada pada inventori kuantum QMVI (44,4%) dan optika geometri FTGOT (35,1%). GPT-5 mencapai 81,5% dengan biaya US$2,18 per putaran *benchmark*, sedangkan beberapa model yang lebih murah mendekati kinerjanya. Pada S14, ahli berhasil menyelesaikan 86% tugas diagram, pemula 70%, ChatGPT-4 50%, dan DeepSeek-R1 hanya 10%. Data pelacakan mata menunjukkan bahwa ahli memakai pemindaian terstruktur berbasis skema. Gambar 2 merangkum kinerja tersebut.



Gambar 2. Akurasi AI pada Tugas Fisika Bergambar (Data dari S8, S9, S14, dan S15)

Perbandingan pada Gambar 2 bersifat ilustratif dan tidak dimaksudkan sebagai perbandingan langsung, karena setiap studi menggunakan instrumen, versi model, jumlah pengulangan, dan prosedur pemberian prompt yang berbeda. Untuk S9 hanya ditampilkan dua model teratas dan dua model terbawah dari 15 model yang diuji. Berbeda dengan peran lain, temuan pada peran penafsir didukung oleh studi berkekuatan bukti tinggi (S4, S8, S9, S15), sehingga kesenjangan antara penalaran dan interpretasi visual merupakan temuan yang paling kokoh dalam korpus ini.

AI sebagai mediator pembelajaran representasional

Tiga studi menempatkan AI di antara siswa dan representasi. Dalam eksperimen kriptografi kuantum berbasis realitas tertambah menggunakan HoloLens 2, 21 kelompok mahasiswa menerima umpan balik ChatGPT pada sebagian soal dan tidak pada soal lainnya (S7). Skor jawaban dengan umpan balik secara signifikan lebih tinggi dengan efek besar ($t = 2{,}721$; $p = 0{,}014$; $d = 1{,}189$). Data pelacakan mata menunjukkan bahwa umpan balik pada soal konseptual mengarahkan atensi ke visualisasi virtual model kuantum, sedangkan umpan balik pada soal prosedural mengarahkan atensi ke perangkat eksperimen nyata.

Pada S2, rutinitas pengecekan terpandu guru diterapkan dalam lima lokakarya 90 menit bagi 22 siswa SMK. Siswa menggunakan ChatGPT untuk penjelasan dan perbandingan, lalu menguji klaim AI terhadap sumbu, satuan, kemiringan, dan luas di bawah grafik. Rerata skor naik dari 2,91/13 (22,4%) menjadi 5,50/13 (42,3%). Dalam salah satu episode, AI salah membaca garis lurus menurun sebagai memiliki garis singgung horizontal dan menyimpulkan kecepatan nol, kemudian guru meminta siswa menghitung kemiringan dari dua titik. Pada S5, instruktur menilai dialog ChatGPT, Gemini, dan Copilot dalam menangani miskonsepsi bahwa elevator naik membutuhkan gaya kabel lebih besar daripada gaya berat. ChatGPT dinilai paling baik dalam akurasi, personalisasi, dan efektivitas pedagogis, sedangkan

Copilot paling lemah. Semua model dinilai lemah dalam memanfaatkan alat bantu visual seperti diagram dan simulasi. Kedua studi yang mengukur hasil belajar (S2 dan S7) masing-masing berkekuatan bukti rendah dan sedang (Tabel 3), sehingga temuan pada peran mediator lebih tepat dibaca sebagai indikasi awal daripada bukti efektivitas.

AI sebagai penilai produk representasional

Satu studi menguji AI sebagai penilai (S13). GPT-5.4 melalui API digunakan untuk menilai 57 laporan praktikum fisika dasar dengan rubrik tetap. Korelasi peringkat dengan penilaian dosen lemah ($\rho = 0{,}38$) dengan galat absolut rata-rata 1,01 poin. Sebanyak 13% umpan balik pada butir tujuan tergolong masuk akal tetapi dangkal, dan sebagian lainnya tidak valid. Sumber galat terutama berasal dari proses ekstraksi teks dan OCR, khususnya pada persamaan, gambar, dan grafik.

Keterbatasan, Tantangan, dan Arah Penelitian Lanjutan (RQ4)

Seluruh studi melaporkan keterbatasan atau tantangan dalam pemanfaatan representasi berbasis AI. Hasil pengodean terhadap bagian keterbatasan, diskusi, dan simpulan setiap studi menghasilkan enam kategori tantangan yang disajikan pada Tabel 5.

Tabel 5. Kategori Tantangan yang Dilaporkan Studi Primer

Kategori tantangan	Uraian	Studi
Keandalan AI pada representasi visual	Kesalahan arah vektor, titik tangkap gaya, pembacaan grafik, topologi rangkaian, dan kaidah tangan kanan	S1, S4, S8, S9, S12, S14, S15
Ketergantungan pada mediasi guru atau ahli	Keluaran AI harus disaring, diverifikasi, atau dipandu guru; AI belum dapat menggantikan penilaian manusia	S1, S2, S10, S12, S13
Keterbatasan desain penelitian	Sampel kecil, tanpa kelompok kontrol, cakupan soal sempit, penilaian oleh peneliti sendiri	S2, S5, S6, S7, S10
Cepatnya perubahan versi model	Temuan terikat pada versi model tertentu; peringkat model tidak stabil	S3, S4, S8, S9
Etika, aksesibilitas, dan biaya	Privasi data, bias, akses yang setara, kelayakan bagi siswa berpenglihatan rendah, biaya penggunaan model	S5, S8, S9, S15
Kendala teknis pemrosesan	Kegagalan ekstraksi informasi dari soal bergambar; keterbatasan OCR pada persamaan dan grafik	S6, S13

Tantangan yang paling sering dilaporkan adalah rendahnya keandalan AI dalam menangani representasi visual, yang muncul pada tujuh studi. Kesalahan yang dilaporkan meliputi arah vektor, titik tangkap gaya, pembacaan kemiringan dan titik potong grafik, topologi rangkaian, serta kaidah tangan kanan (S1, S4, S8, S9, S12, S14, S15). Tantangan kedua adalah ketergantungan pada mediasi guru atau ahli. Lima studi menyatakan bahwa keluaran AI tidak dapat digunakan secara mandiri dan harus disaring, diverifikasi, atau dipandu oleh guru (S1, S2, S10, S12, S13). Tantangan ketiga berasal dari desain penelitian itu sendiri. Penulis S2, S5, S6, S7, dan S10 secara eksplisit menyebut ukuran sampel yang kecil, ketiadaan kelompok kontrol, ruang lingkup soal yang sempit, atau penilaian oleh peneliti sendiri sebagai batasan generalisasi temuan.

Tiga kategori lainnya dilaporkan lebih jarang. Empat studi menekankan bahwa temuan berlaku untuk versi model tertentu dan dapat berubah pada versi berikutnya (S3, S4, S8, S9). Empat studi menyinggung isu etika, aksesibilitas, dan biaya, seperti privasi data, bias model, akses yang setara, kelayakan AI sebagai alat bantu bagi siswa berpenglihatan rendah, dan biaya penggunaan model (S5, S8, S9, S15). Dua studi melaporkan kendala teknis pemrosesan, yaitu kegagalan ekstraksi informasi dari soal bergambar (S6) dan keterbatasan OCR dalam membaca persamaan, gambar, dan grafik pada laporan praktikum (S13).

Arah penelitian lanjutan yang diusulkan oleh studi primer dapat dikelompokkan ke dalam lima rekomendasi. Pertama, memperbesar sampel dan memperluas desain penelitian, termasuk studi dengan populasi yang lebih besar dan pengukuran dampak jangka panjang terhadap retensi (S5, S7). Kedua, memperluas cakupan topik dan representasi, misalnya ke gerak dua dimensi dan animasi (S3), bidang fisika dan STEM lain (S5), serta konversi soal konseptual apa pun menjadi simulasi (S6). Ketiga, mengintegrasikan AI dengan alat visual khusus melalui model hibrida yang menggabungkan model bahasa dengan pembuat diagram dan simulasi (S5, S6) atau dengan lingkungan realitas tertambah dan virtual (S7). Keempat, memanfaatkan kesalahan AI secara pedagogis, yaitu melibatkan siswa dalam mengkritik dan memperbaiki keluaran AI (S1, S2, S3) serta merancang asesmen yang sulit dikerjakan

oleh *chatbot* (S8). Kelima, membandingkan versi model yang lebih baru dan mempertimbangkan biaya agar pemilihan model didasarkan pada bukti (S3, S9).

PEMBAHASAN

Karakteristik Studi dan Kualitas Basis Bukti (RQ1)

Karakteristik studi menunjukkan bahwa riset representasi berbasis AI dalam pembelajaran fisika merupakan bidang yang sangat muda dan berkembang cepat. Tidak adanya studi sebelum 2024, meskipun rentang pencarian dimulai sejak 2016, menandakan bahwa kemunculan model multimodal, khususnya kemampuan ChatGPT-4 menerima masukan gambar sejak Oktober 2023 (Polverini & Gregorcic, 2024), menjadi pemicu utama bidang ini. Dominasi model OpenAI pada 14 dari 15 studi mencerminkan popularitas ChatGPT, tetapi juga berarti bahwa pengetahuan tentang representasi berbasis AI masih sangat bergantung pada satu penyedia model. Variasi kinerja lebih dari 60 poin persentase antarmodel pada S9 menegaskan bahwa generalisasi dari satu model ke model lain tidak dapat diasumsikan. Karakteristik desain studi juga menentukan seberapa kuat kesimpulan yang dapat ditarik, sehingga kualitas basis bukti dinilai secara kritis berikut ini.

Kekuatan dan Kelemahan Metodologis

Basis bukti tentang representasi berbasis AI dalam fisika kuat dalam menjawab apa yang dapat dan tidak dapat dilakukan AI, tetapi masih lemah dalam menjawab apa dampaknya bagi pembelajaran siswa. Dari sisi kekuatan, beberapa studi menunjukkan ketelitian metodologis yang patut dicontoh. S4 menggunakan tiga penilai independen dari tiga institusi dengan reliabilitas antarpenilai yang sangat tinggi pada AP Physics 1 (ICC 0,909–0,935) dan menyediakan kode analisis secara terbuka. S15 menjalankan ChatGPT sebanyak 60 kali untuk menangkap variabilitas keluaran dan melaporkan reliabilitas pengodean penalaran ($\kappa = 0{,}809$) dan penglihatan ($\kappa = 0{,}967$). S9 mengulang setiap butir sepuluh kali dan menambahkan dimensi biaya. S7 menonjol karena mengukur bukan hanya hasil belajar, tetapi juga proses kognitif melalui data pelacakan mata. Isu metodologis utama beserta konsekuensinya dirangkum pada Tabel 6, sedangkan hasil MMAT per studi disajikan pada Tabel 3.

Tabel 6. Isu Metodologis dalam Studi Primer dan Konsekuensinya

Isu metodologis	Studi terdampak	Konsekuensi terhadap kesimpulan
Siswa tidak dilibatkan sebagai subjek belajar	S1, S3, S4, S6, S8, S9, S11, S12, S15	Bukti berbicara tentang kemampuan AI, bukan tentang belajar siswa
Desain pra–pasca tanpa kelompok kontrol; perangkat soal tidak divalidasi setara	S2	Kenaikan skor tidak dapat diatribusikan pada AI
Sampel kecil dan sukarela; sebagian analisis atensi memakai $\alpha = 0{,}10$	S7	Efek besar menjanjikan, tetapi rawan bias seleksi dan galat tipe I
Persepsi siswa melalui skala Likert tanpa ukuran hasil belajar	S10	Rentan efek kebaruan; tidak menunjukkan pemahaman konsep
Jumlah soal atau kasus sangat terbatas (1–5 soal)	S1, S3, S6, S12	Generalisasi pola kesalahan belum terjamin
Penilaian oleh peneliti sendiri atau konsensus tanpa koefisien reliabilitas	S1, S6, S12, S14	Subjektivitas penilaian tidak dapat diukur
Inkonsistensi jumlah partisipan (abstrak 32; hasil 12 dan 25; simpulan 25) dan skala tiga titik	S5	Mengurangi keterlacakan dan kepercayaan terhadap peringkat model
Kesesuaian modalitas model dengan tugas tidak dijelaskan	S14	Skor 10% mungkin mencerminkan ketidaksesuaian modalitas
Prompt tunggal dan pengambilan satu kali	S1, S4, S12	Kinerja dapat berbeda dalam interaksi kelas yang berulang
Validitas temporal rendah karena model cepat berganti	Semua studi, terutama S3	Temuan adalah potret sesaat dan cepat usang

Tiga persoalan perlu digarisbawahi. Pertama, terdapat **ketimpangan antara studi kemampuan dan studi pembelajaran**. Sembilan dari 15 studi tidak melibatkan siswa sebagai subjek belajar, dan hanya dua studi (S2, S7) yang mengukur hasil belajar. Dari dua studi tersebut, satu tidak memiliki kelompok kontrol dan satu lagi berukuran kecil, sehingga klaim bahwa representasi berbasis AI meningkatkan pemahaman belum didukung bukti kausal yang kuat. Hasil MMAT memperkuat kesimpulan ini: seluruh studi berkekuatan bukti tinggi merupakan benchmark kemampuan AI, dan tidak ada studi berkualitas tinggi yang mengukur hasil belajar. Sebagai pembanding, meta-analisis lintas disiplin melaporkan efek positif ChatGPT terhadap kinerja akademik, tetapi juga menekankan perlunya

studi eksperimental yang lebih ketat dan berjangka lebih panjang (Deng et al., 2025). Dalam konteks representasi fisika, studi semacam itu bahkan belum tersedia, sehingga setiap klaim tentang efektivitas AI dalam pembelajaran fisika perlu dibatasi pada temuan yang masih bersifat awal.

Kedua, terdapat **masalah validitas temporal**. Versi model yang diuji berubah dari ChatGPT 3.5 hingga GPT-5.4 dalam dua tahun. Perbaikan dari ChatGPT-4 ke ChatGPT-4o pada BEMA (S8) dan dominasi GPT-5 pada S9 menunjukkan bahwa angka akurasi dalam satu studi bisa usang sebelum studi itu terbit. Karena itu, pola kesalahan kualitatif, seperti kesulitan koordinasi spasial dan kaidah tangan kanan, lebih berharga sebagai temuan jangka panjang daripada angka akurasi.

Ketiga, terdapat **ketidakselarasan instrumen dengan pertanyaan**. S2 secara eksplisit memiliki dua tujuan, yaitu memahami grafik dan menilai keluaran AI, tetapi hanya tujuan pertama yang diukur, padahal kemampuan menilai keluaran AI justru merupakan kontribusi khas intervensi tersebut. S13 menyimpulkan bahwa AI belum setara dengan dosen, tetapi galatnya banyak bersumber dari ekstraksi teks dan OCR, sehingga yang diukur sebagian adalah kualitas alur pemrosesan dokumen, bukan kemampuan model menilai.

Analisis kritis: bias dalam korpus

Korpus juga memperlihatkan beberapa bias struktural. Sepuluh dari 15 artikel terbit di satu jurnal, *Physics Education*, yang banyak memuat laporan praktik singkat, sehingga desain deskriptif mendominasi. Konsentrasi ini sebagian juga merupakan konsekuensi filter bidang subjek *Social Sciences* yang menjaring jurnal pendidikan fisika berorientasi praktik, sementara artikel yang terbit di jurnal fisika umum di luar kategori tersebut tidak terjaring. Tiga studi tentang kemampuan visual AI berasal dari satu kelompok riset (S8, S9, S15), sehingga temuan bernalar tanpa melihat perlu direplikasi oleh kelompok lain. Secara geografis, tidak ada studi dari Indonesia, Afrika, atau Amerika Latin selain Uruguay. Bahasa interaksi dengan AI juga jarang dilaporkan secara eksplisit, padahal kinerja model bahasa dalam bahasa selain Inggris dapat berbeda. Isu etika, seperti privasi data siswa ketika laporan dikirim ke API pihak ketiga, hanya disinggung secara singkat.

Jenis Representasi dan Topik Fisika (RQ2)

Sebaran representasi dan topik dalam korpus lebih mencerminkan ketersediaan instrumen pengukuran daripada kebutuhan kurikulum. Grafik kinematika, diagram gaya, dan diagram listrik-magnet mendominasi karena tersedia inventori baku seperti TUG-K, FCI, dan BEMA yang memudahkan *benchmarking* dan pembandingan dengan data siswa. Akibatnya, representasi yang penting dalam kurikulum fisika sekolah, seperti diagram siklus termodinamika, diagram sinar optika, diagram tingkat energi, serta representasi dinamis berupa animasi dan video, hampir tidak dikaji. Kesulitan AI dalam koordinasi spasial yang ditemukan pada diagram listrik-magnet (S4, S8) memberi alasan untuk menduga bahwa kesulitan serupa juga muncul pada representasi lain yang menuntut penalaran spasial, misalnya diagram sinar dan muka gelombang. Dugaan ini perlu diuji secara empiris.

Dari sudut teori multirepresentasi, korpus juga timpang dalam hal fungsi representasi yang dikaji. Dalam kerangka DeFT (Ainsworth, 2006), representasi berfungsi melengkapi informasi, membatasi interpretasi, dan membangun pemahaman yang lebih dalam. Sebagian besar studi hanya menguji apakah AI dapat membaca atau menghasilkan representasi yang benar. Hanya S2 dan S7 yang menyentuh fungsi membangun pemahaman, yaitu bagaimana representasi yang dimediasi AI membantu siswa mengoordinasikan grafik dengan konsep atau mengoordinasikan visualisasi virtual dengan perangkat nyata. Dengan kata lain, riset saat ini lebih banyak menjawab apakah representasi buatan AI benar, dan belum menjawab apakah representasi tersebut membantu siswa belajar. Pergeseran fokus dari ketepatan representasi ke fungsi representasi dalam pembelajaran menjadi kebutuhan utama bidang ini, sejalan dengan kedudukan kompetensi representasional sebagai tujuan pembelajaran fisika (Kozma & Russell, 1997; Kohl & Finkelstein, 2005).

Hubungan setiap peran dengan kerangka DeFT dapat dirinci lebih jauh. Sebagai pembuat, AI berpotensi mendukung fungsi melengkapi karena dapat dengan cepat menyediakan representasi tambahan, misalnya grafik dari data atau simulasi dari soal verbal. Namun, ketika representasi yang dihasilkan keliru, fungsi membatasi interpretasi justru berbalik menjadi sumber miskonsepsi (S1, S12). Sebagai penafsir, AI belum dapat diandalkan untuk menjalankan fungsi penerjemahan antarrepresentasi yang menurut Ainsworth (1999, 2006) menjadi syarat terbangunnya pemahaman mendalam; penggunaan AI sebagai penafsir tanpa verifikasi bahkan berisiko mengambil alih proses penerjemahan yang seharusnya dilakukan siswa. Sebagai mediator, AI paling dekat dengan fungsi membangun

pemahaman karena umpan balik dan rutinitas pengecekan mendorong siswa menghubungkan grafik dengan konsep atau visualisasi virtual dengan perangkat nyata (S2, S7). Sebagai penilai, AI berpotensi memberi umpan balik cepat atas representasi siswa, tetapi galat dalam membaca persamaan dan grafik (S13) dapat memberi sinyal yang keliru tentang kualitas representasi siswa. Dengan demikian, kontribusi AI terhadap kompetensi representasional tidak ditentukan oleh perannya semata, tetapi oleh sejauh mana peran tersebut tetap menuntut siswa melakukan penerjemahan dan koordinasi representasi secara aktif.

Kinerja AI dan Dampaknya terhadap Pembelajaran (RQ3)

Temuan RQ3 dapat dimaknai melalui tiga pola utama, yaitu kesenjangan antara penalaran dan penglihatan, peran kendali dalam menentukan keandalan, serta potensi pedagogis dari kesalahan AI.

Paradoks "bernalar tanpa melihat"

Temuan paling konsisten dalam kajian ini adalah bahwa AI generatif mampu menyusun strategi penalaran fisika yang benar, tetapi sering gagal menerapkannya pada representasi visual yang konkret. Pola ini muncul pada TUG-K (S15), BEMA (S8), soal AP (S4), dan tugas diagram (S14). Dalam kerangka DeFT (Ainsworth, 2006), kesulitan ini terletak pada fungsi **penerjemahan** dan **koordinasi** antarrepresentasi. AI menguasai representasi verbal-simbolik, tetapi gagal memetakan representasi tersebut ke fitur spasial seperti kemiringan, titik potong, arah vektor, dan orientasi tiga dimensi. Hal ini menjelaskan mengapa kaidah tangan kanan muncul sebagai titik lemah di tiga studi berbeda (S4, S8, S9).

Pola kesalahan AI ternyata mirip dengan pola kesalahan siswa pemula. Klassen dan Friege (2026) menunjukkan bahwa pemula mengandalkan strategi pencocokan terbalik dari pilihan jawaban, sedangkan ahli memindai diagram secara terstruktur berdasarkan skema. Pembacaan grafik oleh AI yang melompat ke kesimpulan tanpa memeriksa sumbu dan interval (S2, S3) menyerupai perilaku pemula tersebut. Bedanya, AI menyampaikan kesimpulan keliru dengan bahasa yang fasih dan yakin. Kefasihan inilah yang membuat kesalahan representasional AI lebih berbahaya daripada kesalahan siswa, karena diagram yang rapi dan penjelasan yang lancar memberi kesan otoritas sehingga siswa yang belum memiliki kompetensi representasional cenderung menerimanya tanpa kritik (Fulminar, 2026a).

Kendali sebagai faktor yang diduga memengaruhi keandalan

Berdasarkan sintesis lintas peran, penulis menginterpretasikan bahwa keandalan representasi berbasis AI tampak lebih berkaitan dengan tingkat kendali atas keluaran AI daripada dengan merek atau versi model. Pada satu ujung, AI dibiarkan menggambar diagram secara bebas dan menghasilkan kesalahan konseptual yang sistematis (S1, S12). Pada titik tengah, AI menulis kode atas arahan pengguna yang memahami fisikanya, sehingga grafik yang dihasilkan dapat dikendalikan dan diverifikasi (S3, S11). Pada ujung lain, AI hanya mengekstraksi parameter, sementara hukum fisika dijalankan oleh mesin simulasi yang terkunci (S6), atau guru memegang kendali konseptual atas setiap keluaran (S2, S10). Hasil yang lebih andal cenderung muncul pada ujung kendali tinggi. Perlu ditegaskan bahwa pola ini merupakan interpretasi penulis atas studi-studi yang berbeda desain, bukan hubungan sebab-akibat yang telah diuji secara empiris. Hanya S6 yang membandingkan dua tingkat kendali secara langsung, dan itu pun dalam studi pilot dengan jumlah soal terbatas. Dugaan ini perlu diuji melalui studi yang memanipulasi tingkat kendali secara sistematis. Temuan ini sejalan dengan argumen bahwa AI di kelas sebaiknya diposisikan sebagai alat yang dikurasi, bukan sebagai sumber pengetahuan yang mandiri (Kasneci et al., 2023).

Implikasinya, pertanyaan tentang model mana yang terbaik kurang produktif. Paul, Kaur, Chakraborty, dan Shrestha (2026) menemukan bahwa peringkat model berubah dari tahun ke tahun, sedangkan Polverini dan Gregorcic (2025) menemukan bahwa model mahal tidak selalu lebih baik. Pertanyaan yang lebih berguna bagi guru dan pengembang adalah bagian representasi mana yang boleh diserahkan kepada AI dan bagian mana yang harus dikunci oleh fisika atau oleh guru.

Kesalahan AI sebagai sumber belajar

Ketidaksempurnaan AI ternyata dapat dimanfaatkan secara pedagogis. Ramkorun (2024) menyebut keterbatasan ChatGPT sebagai berkah karena memaksa pengguna menggunakan kemampuan kognitifnya untuk menilai data. Fulminar (2026a) mengamati bahwa diskusi panel ahli tentang kesalahan diagram AI memunculkan artikulasi eksplisit atas prinsip dan konvensi representasi. Gaurina

dan Štibi (2026) mengubah gagasan ini menjadi rutinitas kelas yang konkret, yaitu setiap klaim AI harus diuji terhadap sumbu, satuan, kemiringan, luas, dan interval grafik sebelum diterima atau ditolak.

Dari sudut teori, praktik ini melatih **kompetensi metarepresentasional**, yaitu kemampuan menilai kecocokan dan kualitas suatu representasi (diSessa, 2004). Siswa tidak hanya belajar membaca grafik, tetapi juga belajar menilai apakah sebuah representasi buatan mesin benar secara fisis. Namun, pendekatan ini bergantung pada guru yang memiliki kompetensi representasional kuat. Tanpa guru yang mampu mendeteksi kesalahan AI, strategi belajar dari kesalahan AI berisiko berubah menjadi belajar kesalahan dari AI.

Tantangan dan Arah Penelitian (RQ4)

Tantangan yang dilaporkan studi primer (Tabel 5) menunjukkan bahwa hambatan pemanfaatan representasi berbasis AI tidak hanya bersifat teknis, tetapi juga pedagogis, metodologis, dan etis. Bagian ini membahas implikasi tantangan tersebut terhadap kesetaraan akses, memetakan seluruh temuan dalam analisis SWOT, lalu merumuskan implikasi praktis dan agenda riset.

Aksesibilitas, biaya, dan kesetaraan

AI menurunkan hambatan teknis dan finansial untuk menghasilkan representasi berkualitas. Guru tanpa perangkat lunak berbayar dapat membuat grafik ilmiah (S3), guru tanpa latar pemrograman dapat membangun sistem akuisisi data (S11), dan guru sekolah menengah dapat membangun laboratorium virtual sesuai konteks lokal (S10). Analisis biaya pada S9 juga menunjukkan bahwa model yang lebih murah dapat memadai untuk tugas tertentu, yang relevan bagi sekolah dengan anggaran terbatas. Di sisi lain, S8 dan S15 memperingatkan bahwa AI belum layak digunakan sebagai alat bantu bagi siswa tunanetra atau berpenglihatan rendah, karena justru pada pembacaan visual itulah AI paling sering keliru. Bagi Indonesia, ketimpangan fasilitas laboratorium antarsekolah menjadikan pendekatan seperti S10 dan S11 berpotensi memperluas akses praktikum, sementara ketiadaan studi dari Indonesia dalam korpus ini menunjukkan peluang riset yang besar.

Analisis SWOT

Dengan mengadaptasi pendekatan Zhang dan Tur (2024), temuan RQ1 sampai RQ4 dirangkum dalam analisis SWOT pada Tabel 7. Pengelompokan setiap unsur menggunakan dua kriteria, yaitu sumber (internal atau eksternal terhadap teknologi AI) dan arah dampak (mendukung atau menghambat pembelajaran representasional), mengikuti logika SWOT pada kajian ChatGPT dalam pendidikan (Farrokhnia, Banihashem, Noroozi, & Wals, 2024). Kekuatan dan kelemahan merupakan karakteristik internal kinerja AI yang teramati langsung dalam studi primer, sedangkan peluang dan ancaman merupakan kondisi eksternal, yaitu konteks penggunaan di kelas, desain pembelajaran, dan ekosistem teknologi, yang dapat memperbesar manfaat atau risiko AI. Dengan cara ini, analisis SWOT tidak sekadar merangkum temuan, tetapi juga mengarahkan strategi: kekuatan dimanfaatkan untuk menangkap peluang, misalnya kemampuan menulis kode untuk arsitektur hibrida, sedangkan kelemahan diantisipasi agar tidak berubah menjadi ancaman, misalnya interpretasi visual yang lemah yang dapat menyebarkan miskonsepsi.

Tabel 7. Analisis SWOT Representasi Berbasis AI dalam Pembelajaran Fisika

Dimensi	Temuan utama	Studi pendukung
Kekuatan	Strategi penalaran sering benar, mampu menulis kode visualisasi dan simulasi, dan umpan balik personal yang mengarahkan atensi visual	S3, S7, S11, S15
Kelemahan	Interpretasi visual lemah, kesalahan arah vektor, titik tangkap gaya, dan kaidah tangan kanan serta jarang memanfaatkan visual secara spontan	S1, S4, S5, S8, S12, S14
Peluang	Arsitektur AI terbatas, pedagogi kritik keluaran AI, laboratorium virtual murah dan model berbiaya rendah yang memadai	S2, S6, S9, S10
Ancaman	Penyebaran miskonsepsi melalui visual yang meyakinkan, ketergantungan siswa, penilaian otomatis yang tidak valid dan kedaluwarsa temuan	S1, S9, S12, S13

Implikasi dan agenda riset

Bagi guru fisika, sintesis ini menghasilkan empat prinsip praktis. Pertama, diagram atau grafik buatan AI tidak digunakan sebagai bahan ajar tanpa verifikasi, terutama untuk diagram benda bebas, arah vektor, dan situasi tiga dimensi. Kedua, AI dimanfaatkan sebagai penulis kode visualisasi atau mitra pengembangan simulasi, dengan guru tetap memverifikasi isi fisikanya. Ketiga, kesalahan AI diubah menjadi tugas belajar melalui rutinitas pengecekan terhadap sumbu, satuan, kemiringan, luas,

titik tangkap gaya, dan konvensi tanda. Keempat, *prompt* siswa diarahkan dari permintaan jawaban ke permintaan penjelasan. Bagi pengembang media, temuan S6 menyarankan arsitektur hibrida, yaitu model bahasa menangani pemahaman soal dan dialog, sedangkan representasi dihasilkan oleh mesin fisika atau pustaka grafik yang deterministik. Bagi lembaga pendidikan tenaga kependidikan, kompetensi representasional dan kompetensi menilai keluaran AI perlu menjadi bagian eksplisit dari kurikulum calon guru fisika.

Dalam konteks Indonesia, implikasi tersebut dapat diterjemahkan ke dalam tiga ranah. Pertama, pada perkuliahan calon guru fisika dan program pendidikan profesi guru, mahasiswa perlu dilatih menilai diagram dan grafik buatan AI sebagai bagian dari pengetahuan teknologi-pedagogi-konten (Mishra & Koehler, 2006), termasuk dimensi etis penggunaan AI yang dirumuskan Celik (2023) sebagai Intelligent-TPACK. Küchemann et al. (2023) menunjukkan bahwa calon guru fisika dapat memanfaatkan ChatGPT untuk menyusun tugas fisika, tetapi kualitas tugas tetap bergantung pada penilaian kritis calon guru terhadap keluaran AI. Tugas mengoreksi diagram benda bebas atau grafik gerak buatan AI dapat dijadikan kegiatan perkuliahan yang sekaligus melatih kompetensi representasional calon guru. Kedua, bagi guru fisika sekolah menengah, rutinitas pengecekan seperti pada S2 dapat diintegrasikan ke dalam pembelajaran berbasis proyek dan asesmen formatif dalam Kurikulum Merdeka, misalnya dengan meminta siswa membandingkan grafik gerak buatan AI dengan data praktikum mereka sendiri. Ketiga, pengembangan laboratorium virtual dan antarmuka data berbantuan AI (S10, S11) berpotensi membantu sekolah dengan fasilitas laboratorium terbatas, dengan syarat guru tetap memverifikasi isi fisika dan interaksi dilakukan dalam bahasa yang dipahami siswa. Ketiga implikasi ini masih perlu diuji secara empiris karena belum ada studi dari Indonesia dalam korpus ini.

Arah penelitian yang diusulkan studi primer pada bagian Hasil umumnya masih bersifat perluasan cakupan, seperti memperbesar sampel dan menambah topik. Berdasarkan kesenjangan yang teridentifikasi dalam kajian ini, penulis mengusulkan lima agenda riset yang lebih spesifik. Agenda pertama adalah bergeser dari kemampuan AI ke pembelajaran siswa melalui studi kuasi-eksperimen atau eksperimen acak dengan kelompok kontrol, instrumen tervalidasi dalam bentuk paralel yang setara, dan ukuran retensi jangka panjang. Agenda kedua adalah mengoperasionalkan kemampuan siswa mendeteksi dan memperbaiki kesalahan representasi buatan AI sebagai variabel hasil tersendiri. Agenda ketiga adalah memperluas topik dan representasi ke termodinamika, optika, fisika modern, serta representasi dinamis seperti animasi dan video buatan AI. Agenda keempat adalah memperluas konteks penelitian ke Indonesia dan negara berkembang lain, termasuk interaksi dengan AI dalam bahasa lokal dan pada sekolah dengan infrastruktur terbatas. Agenda kelima adalah menerapkan standar pelaporan riset AI yang mencakup versi dan tanggal akses model, *prompt* lengkap, jumlah pengulangan, modalitas masukan, serta reliabilitas penilai.

Keterbatasan Kajian

Kajian ini memiliki beberapa keterbatasan yang dilaporkan sesuai anjuran PRISMA 2020. Penelusuran hanya dilakukan pada satu basis data, yaitu Scopus, dengan filter bidang subjek *Social Sciences*, sehingga artikel pendidikan fisika yang terbit di jurnal yang hanya diklasifikasikan dalam bidang Physics and Astronomy atau Computer Science, serta artikel yang terindeks di Web of Science, ERIC, atau SINTA, berpotensi tidak terjaring. Satu artikel juga dikeluarkan karena teks lengkapnya tidak dapat diakses. Jumlah studi yang dianalisis relatif kecil dan sebagian besar terbit pada 2026, sehingga sintesis ini mencerminkan bidang yang masih sangat muda. Pembatasan pada artikel berbahasa Inggris dan jurnal bereputasi berpotensi mengabaikan praktik yang dilaporkan dalam bahasa lain atau di prosiding. Heterogenitas desain dan ukuran hasil tidak memungkinkan meta-analisis, sehingga perbandingan pada Gambar 2 bersifat ilustratif. Selain itu, tipologi empat peran AI dihasilkan secara induktif dari korpus ini dan perlu diuji pada korpus yang lebih luas. Kajian selanjutnya disarankan memperluas penelusuran ke bidang Physics and Astronomy dan Computer Science pada Scopus serta ke basis data Web of Science dan ERIC. Penilaian MMAT dan pengodean dilakukan oleh tim penulis sehingga tetap mengandung unsur subjektivitas, meskipun telah diminimalkan melalui penilaian independen dan konsensus.

KESIMPULAN

Kesimpulan

Kajian sistematis terhadap 15 studi berdasarkan PRISMA 2020 menghasilkan dua temuan utama. Pertama, AI berperan sebagai pembuat, penafsir, mediator, dan penilai representasi fisika, dengan peran pembuat dan penafsir sebagai yang paling banyak dikaji. Kedua, kemampuan AI belum seimbang: AI relatif mampu menyusun penalaran verbal, tetapi lemah dalam menafsirkan dan menghasilkan representasi visual, terutama yang menuntut penalaran spasial. Bukti mengenai pengaruh AI terhadap hasil belajar siswa masih terbatas pada dua studi berskala kecil dengan kekuatan bukti rendah hingga sedang, sehingga efektivitas AI dalam pembelajaran fisika belum dapat disimpulkan.

Secara konseptual, kajian ini menegaskan bahwa ketepatan AI dalam menyelesaikan soal fisika tidak sama dengan kemampuannya mendukung kompetensi representasional siswa. Melalui kerangka DeFT, kontribusi AI tampak bergantung pada sejauh mana penggunaannya tetap menuntut siswa menerjemahkan, mengoordinasikan, dan mengevaluasi representasi. Kajian ini juga mengajukan dugaan bahwa tingkat kendali atas keluaran AI, baik melalui arsitektur sistem maupun mediasi guru, lebih berkaitan dengan keandalan representasi daripada versi model.

Penelitian selanjutnya perlu bergeser dari pengujian akurasi AI menuju studi eksperimental dengan kelompok kontrol yang mengukur dampak representasi berbasis AI terhadap kompetensi representasional dan kemampuan siswa mengkritik keluaran AI, memperluas topik di luar mekanika, serta melibatkan konteks Indonesia dan negara berkembang lain.

REFERENSI

Abreu, M., Suárez, A., Stari, C., & Marti, A. C. (2026). Exploring the potential of ChatGPT for feedback and evaluation in experimental physics. *Physics Education, 61*, 045048. https://doi.org/10.1088/1361-6552/ae8110

Ainsworth, S. (1999). The functions of multiple representations. *Computers & Education, 33*(2–3), 131–152. https://doi.org/10.1016/S0360-1315(99)00029-9

Ainsworth, S. (2006). DeFT: A conceptual framework for considering learning with multiple representations. *Learning and Instruction, 16*(3), 183–198.

Beichner, R. J. (1994). Testing student interpretation of kinematics graphs. *American Journal of Physics, 62*(8), 750–762.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77–101.

Celik, I. (2023). Towards Intelligent-TPACK: An empirical study on teachers' professional knowledge to ethically integrate artificial intelligence (AI)-based tools into education. *Computers in Human Behavior, 138*, 107468. https://doi.org/10.1016/j.chb.2022.107468

Coban, A., Dzsotjan, D., Küchemann, S., Durst, J., Kuhn, J., & Hoyer, C. (2025). AI support meets AR visualization for Alice and Bob: Personalized learning based on individual ChatGPT feedback in an AR quantum cryptography experiment for physics lab courses. *EPJ Quantum Technology, 12*, 15. https://doi.org/10.1140/epjqt/s40507-025-00310-z

Deng, R., Jiang, M., Yu, X., Lu, Y., & Liu, S. (2025). Does ChatGPT enhance student learning? A systematic review and meta-analysis of experimental studies. *Computers & Education, 227*, 105224. https://doi.org/10.1016/j.compedu.2024.105224

Ding, L., Chabay, R., Sherwood, B., & Beichner, R. (2006). Evaluating an electricity and magnetism assessment tool: Brief electricity and magnetism assessment. *Physical Review Special Topics—Physics Education Research, 2*(1), 010105.

diSessa, A. A. (2004). Metarepresentation: Native competence and targets for instruction. *Cognition and Instruction, 22*(3), 293–331.

Farrokhnia, M., Banihashem, S. K., Noroozi, O., & Wals, A. (2024). A SWOT analysis of ChatGPT: Implications for educational practice and research. *Innovations in Education and Teaching International, 61*(3), 460–474. https://doi.org/10.1080/14703297.2023.2195846

Fulminar, L. P. (2026a). An aid or not? Examining ChatGPT-generated visualizations in physics problem solving. *Physics Education, 61*, 025034.

Fulminar, L. P. (2026b). When AI draws forces: Evaluation of free-body diagrams generated by ChatGPT, Gemini, and Copilot. *Physics Education, 61*, 035007.

Garg, A. (2026). Teacher-guided AI-assisted development of virtual physics laboratories: Classroom case studies in secondary physics. *Physics Education, 61*, 055024.

Gaurina, M., & Štibi, I. (2026). Using generative AI to support students' interpretation of kinematics graphs: A classroom-based approach. *Physics Education, 61*, 055020.

Gusenbauer, M., & Haddaway, N. R. (2020). Which academic search systems are suitable for systematic reviews or meta-analyses? Evaluating retrieval qualities of Google Scholar, PubMed, and 26 other resources. *Research Synthesis Methods, 11*(2), 181–217. https://doi.org/10.1002/jrsm.1378

Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. *The Physics Teacher, 30*(3), 141–158.

Hong, Q. N., Fàbregues, S., Bartlett, G., Boardman, F., Cargo, M., Dagenais, P., … Pluye, P. (2018). The Mixed Methods Appraisal Tool (MMAT) version 2018 for information professionals and researchers. *Education for Information, 34*(4), 285–291.

Kahaleh, R., & Lopez, V. (2025). Evaluating large language models in high school physics education: Addressing misconceptions and fostering conceptual understanding. *Physics Education, 60*, 025013.

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*, 102274.

Kitchenham, B. (2004). *Procedures for performing systematic reviews* (Technical Report TR/SE-0401). Keele, UK: Keele University.

Klassen, M., & Friege, G. (2026). Difficulties in diagram interpretation in physics: An eye-tracking study of visual attention and diagram interpretation strategies. *Physical Review Physics Education Research, 22*, 010121. https://doi.org/10.1103/58hn-493c

Kohl, P. B., & Finkelstein, N. D. (2005). Student representational competence and self-assessment when solving physics problems. *Physical Review Special Topics—Physics Education Research, 1*(1), 010104.

Kortemeyer, G. (2023). Could an artificial-intelligence agent pass an introductory physics course? *Physical Review Physics Education Research, 19*(1), 010132. https://doi.org/10.1103/PhysRevPhysEducRes.19.010132

Kozma, R. B., & Russell, J. (1997). Multimedia and understanding: Expert and novice responses to different representations of chemical phenomena. *Journal of Research in Science Teaching, 34*(9), 949–968.

Küchemann, S., Steinert, S., Revenga, N., Schweinberger, M., Dinc, Y., Avila, K. E., & Kuhn, J. (2023). Can ChatGPT support prospective teachers in physics task development? *Physical Review Physics Education Research, 19*(2), 020128. https://doi.org/10.1103/PhysRevPhysEducRes.19.020128

Martin, F., Zhuang, M., & Schaefer, D. (2024). Systematic review of research on artificial intelligence in K-12 education (2017–2022). *Computers and Education: Artificial Intelligence, 6*, 100195. https://doi.org/10.1016/j.caeai.2023.100195

Mishra, P., & Koehler, M. J. (2006). Technological pedagogical content knowledge: A framework for teacher knowledge. *Teachers College Record, 108*(6), 1017–1054. https://doi.org/10.1111/j.1467-9620.2006.00684.x

Mortimer, R. (2026). Democratising scientific software in physics education: AI-assisted custom instrumentation for teachers. *Physics Education, 61*, 035043.

O'Connor, C., & Joffe, H. (2020). Intercoder reliability in qualitative research: Debates and practical guidelines. *International Journal of Qualitative Methods, 19*, 1–13. https://doi.org/10.1177/1609406919899220

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ, 372*, n71.

Paul, B., Kaur, J., Chakraborty, S., & Shrestha, S. (2026). How well do AI systems solve AP physics? A comparative evaluation of large language models on algebra-based free response questions. *Physics Education, 61*, 045008.

Polverini, G., & Gregorcic, B. (2024). Performance of ChatGPT on the test of understanding graphs in kinematics. *Physical Review Physics Education Research, 20*, 010109. https://doi.org/10.1103/PhysRevPhysEducRes.20.010109

Polverini, G., & Gregorcic, B. (2025). Multimodal large language models and physics visual tasks: Comparative analysis of performance and costs. *European Journal of Physics, 46*, 055708. https://doi.org/10.1088/1361-6404/ae03f8

Polverini, G., Melin, J., Önerud, E., & Gregorcic, B. (2025). Performance of ChatGPT on tasks involving physics visual representations: The case of the brief electricity and magnetism assessment. *Physical Review Physics Education Research, 21*, 010154. https://doi.org/10.1103/PhysRevPhysEducRes.21.010154

Ramkorun, B. (2024). Graph plotting of 1-D motion in introductory physics education using scripts generated by ChatGPT 3.5. *Physics Education, 59*, 025020.

Rosengrant, D., Van Heuvelen, A., & Etkina, E. (2009). Do students use and understand free-body diagrams? *Physical Review Special Topics—Physics Education Research, 5*(1), 010108.

Sachithanand, K. (2026). Constrained vs. unconstrained free-fall and projectile-motion AI problem-to-simulation generation: A pilot study. *Physics Education, 61*, 055025.

Van Heuvelen, A. (1991). Learning to think like a physicist: A review of research-based instructional strategies. *American Journal of Physics, 59*(10), 891–897.

Yeadon, W., & Hardy, T. (2024). The impact of AI in physics education: A comprehensive review from GCSE to university levels. *Physics Education, 59*(2), 025010. https://doi.org/10.1088/1361-6552/ad1fa2

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education – where are the educators? *International Journal of Educational Technology in Higher Education, 16*, 39.

Zhang, P., & Tur, G. (2024). A systematic review of ChatGPT use in K-12 education. *European Journal of Education, 59*(2), e12599. https://doi.org/10.1111/ejed.12599