
Efektivitas Kecerdasan Buatan Generatif terhadap Hasil Belajar Peserta Didik: Meta-Analisis Studi Eksperimen dan Kuasi-Eksperimen

Roy Marthin Sihombing^{1)*}, Yeremia Christoper Sihotang²⁾, Anggina Pertiwi Nasution³⁾, Balqis Hijriaulia⁴⁾, Siska Anzani⁵⁾, Daldi Altamerando Suramana Sbr⁶⁾, Yosua Nathanael Nainggolan⁷⁾, Elsa Sabrina⁸⁾

^{1,2,3,4,5,6,7,8)} Pendidikan Teknologi Informatika dan Komputer, Fakultas Teknik, Universitas Negeri Medan

Email : roymarthin454@gmail.com
sihotangyeremia123@gmail.com

Abstrak

Penelitian ini bertujuan menganalisis secara kuantitatif pengaruh penggunaan kecerdasan buatan generatif (GenAI) terhadap hasil belajar peserta didik serta mengidentifikasi karakteristik intervensi yang berkaitan dengan variasi pengaruh tersebut. Metode yang digunakan adalah meta-analisis atas studi eksperimen dan kuasi-eksperimen yang terbit pada 2021–2026 dengan pelaporan mengikuti PRISMA 2020. Pencarian dilakukan pada Scopus, ResearchGate, dan Google Scholar, dan sebanyak 20 studi dengan total 3.005 peserta didik disertakan dalam sintesis. Ukuran efek gabungan diestimasi dengan model efek acak, disertai analisis subkelompok dan analisis sensitivitas. Hasil menunjukkan pengaruh positif pada kisaran sedang menuju besar (Hedges' $g = 0,707$; interval kepercayaan 95%: 0,339–1,074) dengan heterogenitas tinggi ($I^2 = 93,1\%$). Pengaruh cenderung lebih besar pada penggunaan GenAI yang terstruktur dengan panduan pedagogis ($g = 0,839$) daripada penggunaan tanpa panduan ($g = 0,034$), serta pada perbandingan dengan pembelajaran tanpa GenAI ($g = 0,861$) daripada dengan umpan balik manusia ($g = 0,200$). Disimpulkan bahwa efektivitas GenAI berkaitan dengan rancangan pedagogis dan konteks penerapannya, sehingga integrasinya perlu dirancang secara terarah.

Kata kunci: kecerdasan buatan generatif, hasil belajar, meta-analisis, panduan pedagogis, ukuran efek

Abstract

This study aimed to quantitatively analyze the effect of generative AI on students' learning outcomes and to identify intervention characteristics associated with variation in that effect. A meta-analysis of experimental and quasi-experimental studies published between 2021 and 2026 was conducted and reported following PRISMA 2020. Searches were carried out in Scopus, ResearchGate, and Google Scholar, and 20 studies involving 3,005 students were included. The pooled effect was estimated using a random-effects model, complemented by subgroup and sensitivity analyses. The results showed a positive effect in the moderate-to-large range (Hedges' $g = 0.707$; 95% CI 0.339–1.074) with high heterogeneity ($I^2 = 93.1\%$). Effects tended to be larger when generative AI was used in a structured way with pedagogical guidance ($g = 0.839$) than without guidance ($g = 0.034$), and larger when compared with instruction without generative AI ($g = 0.861$) than with human feedback ($g = 0.200$). It is concluded that the effectiveness of generative AI is associated with pedagogical design and learning context, so its integration should be planned deliberately.

Keywords: generative AI, learning outcomes, meta-analysis, pedagogical guidance, effect size

PENDAHULUAN

Sejak ChatGPT dirilis pada akhir 2022, kecerdasan buatan generatif (*generative artificial intelligence*, selanjutnya disingkat GenAI) berkembang pesat dalam praktik pembelajaran (Deng dkk., 2025). GenAI merupakan teknologi kecerdasan buatan berbasis model bahasa besar (*large language model*, LLM) yang menghasilkan konten berdasarkan masukan pengguna. Dalam pendidikan, teknologi ini digunakan untuk bantuan belajar, umpan balik, tutor berbasis kecerdasan buatan, dan pendampingan tugas (Deng dkk., 2025; Kestin dkk., 2025). Penelitian awal terutama membahas persepsi, pengalaman, dan penerimaan mahasiswa (Alneyadi & Wardat, 2023), sehingga bukti kausal mengenai pengaruh GenAI terhadap hasil belajar belum konsisten. Fokus penelitian kemudian bergeser dari tingkat penggunaan GenAI ke pertanyaan apakah penggunaannya menghasilkan capaian belajar yang lebih tinggi daripada kondisi pembanding.

Studi eksperimen dan kuasi-eksperimen menyediakan bukti kuantitatif mengenai hal tersebut. Pada pembelajaran fisika di pendidikan tinggi, tutor berbasis GenAI yang dirancang berdasarkan praktik pedagogis yang baik menghasilkan pembelajaran yang lebih banyak dalam waktu yang lebih singkat daripada kelas pembelajaran aktif (Kestin dkk., 2025). Pada pembelajaran pemrograman Java, integrasi GenAI yang terstruktur menghasilkan keterampilan berpikir tingkat tinggi (*higher-order thinking skills*, HOTS) dan logika pemrograman yang lebih baik daripada penggunaan ChatGPT tanpa panduan (Nathaniel dkk., 2026). Pada pembelajaran bahasa Inggris, studi kuasi-eksperimen melaporkan efek perlakuan yang signifikan (Hsu, 2025). Hasil belajar mencakup beberapa dimensi, dan pengaruh GenAI dapat berbeda menurut jenis luaran. Deng dkk. (2025) melaporkan efek positif pada kinerja akademik, aspek afektif-motivasi, dan berpikir tingkat tinggi, tetapi efek yang tidak signifikan pada efikasi diri. Chen dan Cheung (2025) melaporkan pola serupa pada pendidikan tinggi, kecuali pada metakognisi.

Namun, hasil penelitian belum seragam. Eksperimen lain melaporkan dampak negatif ChatGPT terhadap kemampuan menulis kreatif (Niloy dkk., 2024), dan umpan balik tulisan yang dihasilkan kecerdasan buatan tidak berbeda signifikan dari umpan balik tutor manusia (Escalante dkk., 2023). Eksperimen lapangan pada hampir seribu siswa sekolah menengah atas menunjukkan bahwa akses ke tutor GPT-4 meningkatkan kinerja selama latihan. Setelah akses dihentikan, siswa pada versi tanpa pengamanan (*guardrails*) memperoleh nilai 17 persen lebih rendah daripada siswa tanpa akses, sedangkan versi dengan pengamanan pedagogis sebagian besar menghilangkan penurunan tersebut (Bastani dkk., 2025). Keragaman hasil ini, termasuk perbedaan akibat rancangan penggunaan, menunjukkan bahwa arah dan besar pengaruh GenAI tidak dapat disimpulkan dari satu studi.

Sintesis kuantitatif terhadap studi eksperimen telah dilakukan dalam beberapa meta-analisis. Deng dkk. (2025; Sun & Zhou, 2024) menganalisis 62 dari 69 studi eksperimen dan melaporkan efek besar pada kinerja akademik ($g+ = 0,712$; $I^2 = 91,789\%$). Meta-analisis lain melaporkan efek sedang ChatGPT terhadap capaian akademik ($g = 0,577$) (Liu dkk., 2025) dan hasil belajar ($g = 0,670$; $I^2 = 91,444\%$) (Wu dkk., 2026). Min Young dan Park (2026) melaporkan $g = 0,573$ untuk capaian belajar, sedangkan sintesis lain memusatkan kajian pada keterlibatan peserta didik (Heung & Chiu, 2025). Perbedaan estimasi tersebut dapat berkaitan dengan perbedaan cakupan dan karakteristik studi, karena bidang pembelajaran, durasi intervensi, dan mode pembelajaran dilaporkan sebagai moderator (Deng dkk., 2025; Liu dkk., 2025; Wu dkk., 2026). Heterogenitas yang tinggi menunjukkan bahwa efek gabungan belum cukup menjelaskan kondisi yang berkaitan dengan efektivitas GenAI.

Sintesis terdahulu memiliki keterbatasan cakupan. Sintesis tersebut membatasi kajian pada ChatGPT, dengan pencarian yang berakhir pada Agustus 2024 (Deng dkk., 2025) dan Juni 2024 (Wu dkk., 2026). Akibatnya, studi yang lebih baru belum tercakup, termasuk studi yang membandingkan GenAI dengan pembelajaran aktif atau penggunaan GenAI tanpa panduan. Moderator yang dianalisis umumnya berupa karakteristik konteks, sedangkan tingkat panduan pedagogis belum diuji sebagai moderator pada sintesis yang ditinjau dalam penelitian ini, meskipun studi terbaru menunjukkan bahwa faktor tersebut berkaitan dengan arah efek.

Penelitian ini bertujuan menganalisis secara kuantitatif pengaruh penggunaan GenAI terhadap hasil belajar peserta didik melalui meta-analisis atas studi eksperimen dan kuasi-eksperimen yang terbit pada 2021–2026. Penelitian ini berbeda dari sintesis terdahulu pada tiga aspek, yaitu cakupan GenAI yang tidak dibatasi pada ChatGPT dan mencakup studi hingga 2026, pemisahan jenis kondisi pembandingan, serta pengujian tingkat panduan pedagogis (penggunaan terstruktur atau berpengamanan dibandingkan dengan penggunaan tanpa panduan) sebagai sumber variasi ukuran efek (Bastani dkk., 2025; Nathaniel dkk., 2026). Tujuan tersebut dijabarkan dalam empat pertanyaan penelitian (PP) berikut.

- PP1. Seberapa besar pengaruh penggunaan GenAI terhadap hasil belajar peserta didik dibandingkan dengan kondisi pembelajaran pembandingan?
- PP2. Seberapa besar heterogenitas pengaruh GenAI terhadap hasil belajar antarpengelitian?

- PP3. Karakteristik penelitian atau intervensi apa yang berkaitan dengan variasi besar pengaruh GenAI terhadap hasil belajar, termasuk tingkat panduan pedagogis penggunaan GenAI dan jenis kondisi pembandingan?
- PP4. Apakah besar pengaruh GenAI berbeda menurut jenis luaran, jenjang pendidikan, bidang pembelajaran, atau karakteristik intervensi?

METODE PENELITIAN

Desain Penelitian dan Pedoman Pelaporan

Penelitian ini merupakan meta-analisis atas studi eksperimen, termasuk uji coba terkendali acak (*randomized controlled trial*, RCT), dan kuasi-eksperimen yang menguji pengaruh GenAI terhadap hasil belajar peserta didik. Pelaporan mengikuti pedoman *Preferred Reporting Items for Systematic Reviews and Meta-Analyses* (PRISMA) 2020 (Page dkk., 2021). Prosedur pencarian, pengodean, perhitungan ukuran efek (*effect size*), dan penilaian risiko bias mengacu pada *Cochrane Handbook* dan panduan meta-analisis (Harrer dkk., 2021).

Strategi Pencarian Literatur

Pencarian dilakukan pada Scopus, ResearchGate, dan Google Scholar pada 17 September 2026 dengan batas tahun terbit 2021–2026. Kata kunci yang digunakan adalah sebagai berikut.

- "Generative AI" "learning outcomes" students
- "Generative Artificial Intelligence" "academic achievement" students
- ChatGPT "learning outcomes" students
- ChatGPT "academic performance" university students
- "Generative AI" "student performance" education

Hasil kelima pencarian digabungkan, kemudian duplikat dihapus sebelum tahap penapisan (*screening*).

Kriteria Inklusi dan Eksklusi

GenAI didefinisikan secara operasional sebagai sistem berbasis model generatif, misalnya LLM, yang menghasilkan teks, kode, atau konten baru sebagai respons terhadap masukan peserta didik. Definisi ini mencakup ChatGPT, tutor atau *chatbot* berbasis LLM, dan aplikasi yang dibangun di atas LLM. Sistem kecerdasan buatan nongeneratif, seperti aplikasi berbasis aturan, pengenalan suara, atau sistem rekomendasi, dikeluarkan. Kriteria inklusi dan eksklusi disusun menggunakan komponen PICOS, yaitu populasi, intervensi, pembandingan, luaran, dan desain penelitian, serta data dan publikasi, dan disajikan pada Tabel 1.

Tabel 1. Kriteria inklusi dan eksklusi (PICOS)

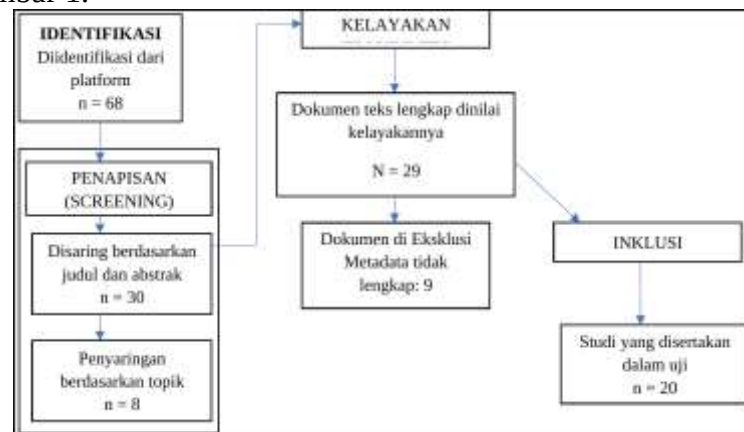
Komponen	Inklusi	Eksklusi
Populasi	Siswa, mahasiswa, atau peserta didik pada pendidikan formal	Guru dalam jabatan, profesional, populasi umum
Intervensi	GenAI (sesuai definisi operasional) yang digunakan dalam aktivitas pembelajaran	Kecerdasan buatan nongeneratif; studi yang hanya menilai kinerja model kecerdasan buatan
Pembandingan	Kelompok tanpa GenAI (pembelajaran konvensional atau aktif) untuk analisis utama; pembandingan lain (GenAI tanpa panduan, umpan balik atau tutor manusia) dikodekan dan dianalisis terpisah	Tanpa kelompok pembandingan
Luaran	Hasil belajar yang terukur secara objektif (tes, kinerja, keterampilan) dengan statistik yang memadai; domain luaran dikodekan terpisah	Persepsi atau kepuasan saja; luaran afektif tidak digabungkan dengan hasil belajar

Desain	Eksperimen dan kuasi-eksperimen; metode campuran (<i>mixed-methods</i>) apabila komponen kuantitatifnya memenuhi kriteria	Noneksperimen; kualitatif, survei atau korelasional, ulasan, konseptual, opini
Data dan publikasi	Artikel jurnal yang telah ditelaah sejawat (<i>peer-reviewed</i>), berbahasa Inggris, terbit pada 2021–2026, dengan teks lengkap tersedia dan statistik memadai (M , SD , n ; atau t , F , η^2 , dan seterusnya)	Duplikasi; apabila satu sampel dilaporkan di beberapa publikasi, hanya satu yang digunakan; statistik tidak memadai atau tidak konsisten secara internal sehingga ukuran efek tidak dapat dihitung

Tabel 1 menjadi acuan dalam menentukan studi yang disertakan dan dikeluarkan. Dua klarifikasi kriteria eksklusi, yaitu desain noneksperimen dan statistik yang tidak konsisten, dirumuskan pada tahap penilaian teks lengkap dan dilaporkan sebagai deviasi dari protokol awal. Kriteria tersebut dimaksudkan agar studi yang dipilih menggunakan GenAI dalam aktivitas pembelajaran, memiliki kelompok pembanding, mengukur hasil belajar secara objektif, dan menyediakan data statistik yang memadai.

Proses Seleksi Studi

Seleksi dilakukan oleh dua penelaah (A dan B), dan keputusan pada setiap tahap diambil berdasarkan kesepakatan. Pencarian pada Scopus, ResearchGate, dan Google Scholar menghasilkan 68 rekaman, dan 67 rekaman tersisa setelah satu duplikat dihapus. Pada penapisan judul dan abstrak, 30 rekaman dikeluarkan dan 37 rekaman dinilai layak untuk penilaian teks lengkap. Delapan artikel dikeluarkan pada tahap ini karena kecerdasan buatan yang digunakan tidak memenuhi definisi GenAI (Tabel 1). Dari 29 artikel yang tersisa, 9 artikel dikeluarkan karena metadata tidak lengkap, sehingga 20 studi disertakan dalam sintesis. Alur seleksi disajikan pada Gambar 1.



Gambar 1. Diagram Alir PRISMA

Gambar 1 menunjukkan alur seleksi penelitian yang dimulai dari identifikasi sebanyak 68 rekaman. Pada tahap penapisan, rekaman disaring berdasarkan judul dan abstrak, kemudian dilanjutkan dengan penilaian kelayakan dokumen teks lengkap. Setelah dokumen yang tidak memenuhi kriteria dikeluarkan, studi yang memenuhi persyaratan akhir dimasukkan ke dalam sintesis meta-analisis.

Ekstraksi dan Pengodean Data

Dua penelaah mengekstraksi dan mengodekan data, dan perbedaan diselesaikan melalui diskusi. Variabel yang dikodekan disajikan pada Tabel 2. Dua variabel menjadi fokus analisis moderator. Tingkat panduan pedagogis dikodekan sebagai terstruktur/berpengaman apabila studi melaporkan paling sedikit satu dari tiga hal berikut: pelatihan atau panduan penggunaan yang eksplisit, pembatasan sistem (misalnya memberikan petunjuk, bukan jawaban langsung), atau kerangka pedagogis yang mengatur peran GenAI. Penggunaan bebas tanpa struktur dikodekan sebagai standar/tanpa panduan. Jenis pembanding dikodekan sesuai dengan Tabel 2.

Tabel 2. Skema pengodean

Variabel	Kategori
Identitas dan desain	Penulis, tahun, negara; RCT/kuasi-eksperimen/silang (<i>crossover</i>); kesetaraan pretes
Sampel	Jumlah per kelompok; jenjang (sekolah / pendidikan tinggi)
Bidang pembelajaran	Bahasa dan menulis; pemrograman dan komputasi; sains dan matematika; kesehatan; lainnya
Bentuk GenAI	ChatGPT digunakan langsung; tutor atau <i>chatbot</i> berbasis LLM yang disesuaikan; lainnya
Tingkat panduan pedagogis	Terstruktur/berpengaman; standar/tanpa panduan; tidak dilaporkan
Jenis pembanding	Konvensional/tanpa GenAI; pembelajaran aktif; GenAI tanpa panduan; umpan balik atau tutor manusia
Durasi intervensi	≤ 4 minggu; > 4 minggu
Domain luaran	Prestasi akademik/pengetahuan; keterampilan bahasa/menulis; berpikir tingkat tinggi/kritis; keterampilan pemrograman/komputasional; lainnya
Data statistik	M , SD , n ; atau t , F , η^2 ; koefisien terkontrol kovariat

Tabel 2 menunjukkan bahwa data yang dikodekan mencakup identitas dan desain penelitian, sampel, bidang pembelajaran, bentuk GenAI, tingkat panduan pedagogis, jenis pembanding, durasi intervensi, domain luaran, serta data statistik. Pengodean ini digunakan untuk menjaga konsistensi pencatatan data dan mendukung analisis moderator sesuai dengan karakteristik masing-masing studi.

Penilaian Risiko Bias

Risiko bias studi acak dinilai dengan *Risk of Bias 2* (RoB 2) dan studi nonacak dengan *Risk of Bias in Non-randomised Studies of Interventions* (ROBINS-I) (Higgins dkk., 2024). Hasil penilaian digunakan pada analisis sensitivitas dan tidak digunakan sebagai kriteria eksklusi. Penilaian dengan RoB 2 mencakup domain bias dari proses randomisasi, deviasi dari intervensi yang direncanakan, data luaran yang hilang, pengukuran luaran, dan pelaporan hasil yang selektif. ROBINS-I menilai bias akibat perancu (*confounding*), seleksi partisipan, klasifikasi intervensi, deviasi dari intervensi yang dimaksudkan, data yang hilang, pengukuran luaran, serta pelaporan hasil yang selektif. Penilaian dilakukan secara independen oleh dua penelaah, dan perbedaan hasil diselesaikan melalui diskusi hingga tercapai kesepakatan.

Perhitungan Ukuran Efek

Ukuran efek utama adalah Hedges' g , bertanda positif apabila hasil berpihak pada kelompok GenAI. Nilai tersebut dihitung dari selisih rerata pascates yang distandardisasi dengan simpangan baku gabungan, kemudian dikoreksi terhadap bias sampel kecil:

$$g = J \times \frac{(M_E - M_C)}{SD_{gabungan}} \quad (1)$$

dengan J sebagai faktor koreksi sampel kecil yang didefinisikan sebagai $J = 1 - 3/(4(n_E + n_C - 2) - 1)$. Galat baku (*standard error*, SE) untuk Hedges' g dihitung menggunakan persamaan (2) berikut:

$$SE(g) = \sqrt{J} \cdot \sqrt{\frac{n_E + n_C}{n_E \cdot n_C} + \frac{g^2}{2(n_E + n_C)}} \quad (2)$$

Apabila rerata dan simpangan baku tidak tersedia, ukuran efek diturunkan dari t , F , atau η^2 menggunakan rumus konversi standar (Harrer dkk., 2021). Interpretasi besar efek mengikuti konvensi Cohen (0,2 kecil, 0,5 sedang, dan 0,8 besar). Nilai g yang telah disesuaikan dengan kovariat digunakan hanya apabila nilai yang tidak disesuaikan tidak tersedia, dan nilai tersebut dikodekan agar dapat diuji pada analisis sensitivitas.

Tiga aturan berikut digunakan untuk menjaga independensi data:

- Satu ukuran efek per studi per domain luaran; apabila terdapat beberapa ukuran dalam satu domain, nilainya dirata-ratakan.
- Studi dengan dua lengan perlakuan dan satu kontrol membagi sampel kontrol secara proporsional.

- Desain silang (*crossover*) dianalisis pada analisis sensitivitas atau dihitung dengan memperhitungkan korelasi antarkondisi apabila statistiknya tersedia.

Analisis Data

Analisis utama (PP1) menggunakan model efek acak (*random-effects*) dengan estimator *restricted maximum likelihood* (REML) dan penyesuaian Hartung-Knapp karena jumlah studi terbatas. Heterogenitas (PP2) dilaporkan dengan statistik Q , τ^2 , I^2 , dan interval prediksi. Studi yang membandingkan GenAI terstruktur dengan GenAI tanpa panduan menjawab pertanyaan yang berbeda dari studi yang membandingkan GenAI dengan pembelajaran tanpa GenAI. Oleh karena itu, studi tersebut tidak digabungkan ke dalam analisis utama dan disintesis secara terpisah, secara naratif apabila jumlahnya kurang dari tiga studi.

Untuk PP3 dan PP4, moderator kategorikal (tingkat panduan pedagogis, jenis pembandingan, domain luaran, jenjang, bidang, durasi, desain, dan bentuk GenAI) diuji dengan analisis subkelompok model efek campuran (*mixed-effects*). Subkelompok dianalisis hanya apabila memuat paling sedikit tiga studi, dan hasilnya diposisikan sebagai eksploratif. Meta-regresi dilakukan hanya apabila tersedia paling sedikit 10 studi. Apabila terdapat ukuran efek ganda yang tidak dapat dihindari, digunakan estimasi varians kokoh, *robust variance estimation* (Pustejovsky & Tipton, 2022).

Analisis sensitivitas mencakup pengeluaran satu studi secara bergantian (*leave-one-out*); pengecualian studi berisiko bias tinggi; pengecualian studi kuasi-eksperimen, studi dengan g terkontrol kovariat, dan studi silang; serta identifikasi pencilan. Bias publikasi diperiksa dengan plot corong (*funnel plot*), uji Egger (apabila terdapat paling sedikit 10 studi per analisis), dan metode *trim-and-fill*, dengan catatan bahwa daya uji rendah pada heterogenitas tinggi. Analisis dijalankan dengan perangkat lunak R (paket *metafor* dan *meta*).

HASIL DAN PEMBAHASAN

Karakteristik Studi yang Disertakan

Sebanyak 20 studi primer dengan total 3.005 peserta didik memenuhi kriteria sintesis akhir. Studi tersebut terbit pada 2024-2026 dan menggunakan desain eksperimen, kuasi-eksperimen, atau *crossover* untuk membandingkan GenAI dengan pembelajaran tanpa GenAI, pembelajaran aktif, atau umpan balik manusia (Tabel 3). Bidang bahasa dan menulis paling banyak diwakili, sedangkan studi bidang kesehatan umumnya menggunakan luaran objektif berupa tes atau kinerja praktik. Variasi desain dan kondisi pembandingan tersebut perlu dipertimbangkan dalam menafsirkan efek gabungan.

Tabel 3. Karakteristik studi yang disertakan dalam meta-analisis

No.	Studi	Tahun	N dianalisis	Jenjang	Bidang	Desain	Pembandingan
1	Gan dkk.	2024	110	Pendidikan tinggi	Kesehatan	RCT	Konvensional/tanpa GenAI
2	Yang, Hsu, dan Wu	2025	153	Sekolah	Pemrograman	Kuasi-eksperimen	Konvensional/tanpa GenAI
3	Nguyen	2025	76	Pendidikan tinggi	Bahasa dan menulis	RCT	Konvensional/tanpa GenAI
4	Bastani dkk.	2025	839	Sekolah	Sains dan matematika	RCT klaster	Konvensional/tanpa GenAI
5	Huang dkk.	2025	187	Pendidikan tinggi	Kesehatan	RCT	Video/konvensional
6	de la Puente dkk.	2024	95	Pendidikan tinggi	Ilmu sosial	RCT	Pembelajaran aktif
7	Kestin dkk.	2025	194	Pendidikan tinggi	Sains dan matematika	RCT silang	Pembelajaran aktif
8	Alnemrat dkk.	2025	120	Pendidikan tinggi	Bahasa dan menulis	Kuasi-eksperimen	Umpan balik guru

9	Ilyas dkk.	2026	100	Pendidikan tinggi	Kesehatan	RCT silang	Konvensional/tanpa GenAI
10	Roza dkk.	2026	150	Pendidikan tinggi	Bahasa dan menulis	Kuasi-eksperimen	Umpan balik manusia
11	Yang, Li, dan Huang	2026	83	Pendidikan tinggi	Pemrograman	Kuasi-eksperimen	Tanpa umpan balik individual
12	Zhang	2025	259	Pendidikan tinggi	Bahasa dan menulis	RCT	Umpan balik manusia
13	Ng, Tan, dan Leung	2024	74	Pendidikan tinggi	Prestasi akademik	Kuasi-eksperimen	Chatbot berbasis aturan
14	Barcaui	2025	85	Pendidikan tinggi	Retensi pengetahuan	RCT	Pembelajaran tradisional
15	El Fathi dkk.	2025	120	Pendidikan tinggi	Sains dan matematika	Kuasi-eksperimen	Pembelajaran tradisional
16	Shahsavar dkk.	2024	60	Pendidikan tinggi	Bahasa dan menulis	Eksperimen longitudinal	Pembelajaran menulis konvensional
17	Lu dan Zeng	2025	71	Sekolah	Bahasa dan menulis	Kuasi-eksperimen	Tanpa umpan balik
18	Cen dan Shakibaei	2026	92	Pendidikan formal	Bahasa dan menulis	RCT tiga lengan	Tanpa kecerdasan buatan
19	Dai, Gong, dan Ma	2026	65	Pendidikan tinggi	Kesehatan	RCT silang	VR tanpa ChatGPT
20	Hu, Gong, dan Hashim	2026	72	Pendidikan tinggi	Bahasa dan menulis	Intervensi metode campuran	Umpan balik guru

Ukuran Efek Tiap Studi

Ukuran efek setiap studi dikonversi ke Hedges' g dengan arah positif apabila hasil lebih tinggi pada kelompok GenAI. Hanya satu kontras utama per studi yang digunakan agar satu sampel tidak memberikan bobot berlebih (Tabel 4).

Tabel 4. Ukuran efek utama setiap studi

No.	Studi	Hedges' g	IK 95%
1	(Gan dkk., 2024)	0,395	0,021-0,770
2	(T.-C. Yang dkk., 2025)	-0,287	-0,604-0,030
3	(Huu Hoang, 2025)	1,238	0,752-1,724
4	(Bastani dkk., 2025)	-0,014	-0,106-0,078
5	(Huang dkk., 2025)	0,666	0,373-0,960
6	(de la Puente dkk., 2024)	0,583	0,176-0,990
7	(Kestin dkk., 2025)	0,541	0,298-0,784
8	(Alnemrat dkk., 2025)	0,148	-0,208-0,504
9	(Maria Ilyas dkk., 2026)	0,571	0,174-0,967
10	(Roza dkk., 2026)	0,094	-0,296-0,483
11	(C. C. Y. Yang dkk., 2026)	1,260	0,787-1,733
12	(Zhang, 2025)	0,191	0,102-0,280
13	(Ng dkk., 2024)	0,739	0,268-1,211
14	(Barcaui, 2025)	-0,680	-1,118-0,243
15	(El Fathi dkk., 2025)	1,544	1,136-1,953
16	(Shahsavar dkk., 2024)	1,691	1,099-2,283
17	(Lu & Zeng, 2025)	2,600	1,963-3,236
18	(Cen & Shakibaei, 2026)	2,060	1,435-2,684
19	(Dai dkk., 2026)	0,623	0,357-0,889
20	(Hu dkk., 2026)	0,716	0,239-1,193

Arah dan besar efek tidak seragam. Sebagian studi memiliki efek mendekati nol atau negatif, sedangkan studi lain memiliki efek sedang hingga besar. Nilai efek yang tinggi perlu ditafsirkan bersama desain penelitian, jenis luaran, dan kondisi pembandingan, dan tidak dapat dianggap berlaku umum.

Pengaruh Keseluruhan GenAI terhadap Hasil Belajar

Estimasi utama menggunakan model efek acak (*random-effects*) dengan REML dan penyesuaian Hartung-Knapp karena studi berasal dari konteks, populasi, dan instrumen yang berbeda (Tabel 5).

Tabel 5. Hasil meta-analisis efek acak dan analisis sensitivitas

Analisis	k	Hedges' g	IK 95%	Keterangan
Efek acak REML + Hartung–Knapp	20	0,707	0,339-1,074	Analisis utama
Tanpa Alnemrat dkk.	19	0,738	0,354-1,121	Sensitivitas risiko bias tinggi
Tanpa studi kuasi-eksperimen	13	0,629	0,217-1,041	Sensitivitas desain
Tanpa studi silang	17	0,735	0,294-1,177	Sensitivitas desain
Leave-one-out, estimasi minimum	19	0,611	0,285-0,936	Mengeluarkan Lu dan Zeng
Leave-one-out, estimasi maksimum	19	0,775	0,422-1,128	Mengeluarkan Barcaui
Uji Egger	20	-	$p = 0,387$	Tidak menunjukkan bukti statistik asimetri

Efek gabungan berada pada kisaran sedang menuju besar dengan interval kepercayaan yang tidak mencakup nol. Estimasi ini relatif dekat dengan hasil meta-analisis terdahulu (Deng dkk. 0,712; Deng dkk. 0,577; Wu dkk. 0,670), meskipun cakupan studinya tidak identik. Heterogenitas tergolong tinggi ($Q = 275,98$; $\tau^2 = 0,550$; $I^2 = 93,1\%$), dan interval prediksi ($-0,785$ hingga $2,199$) mencakup nol. Dengan demikian, studi baru pada konteks lain masih dapat menghasilkan efek kecil atau negatif, sehingga rerata gabungan tidak dapat digunakan untuk menggambarkan efektivitas GenAI secara universal. Arah efek tetap positif pada seluruh analisis sensitivitas, dan tidak ada satu studi pun yang menjelaskan keseluruhan efek gabungan.

Variasi Ukuran Efek Berdasarkan Bidang, Pembandingan, dan Panduan Pedagogis

Analisis subkelompok bersifat eksploratif dan ditafsirkan secara formal hanya apabila tersedia paling sedikit tiga studi (Tabel 6).

Tabel 6. Analisis subkelompok eksploratif

Karakteristik	Subkelompok	k	Hedges' g	IK 95%	I ²
Bidang	Bahasa dan menulis	8	1,061	0,263-1,858	94,5%
	Kesehatan	4	0,586	0,411-0,760	0%
	Sains dan matematika	3	0,672	-1,279-2,624	97,0%
Pembandingan	Non-GenAI	14	0,861	0,341-1,382	95,0%
	Umpan balik	4	0,200	0,028-0,372	38,8%
Panduan pedagogis	Terstruktur/berpengaman	12	0,839	0,420-1,259	93,7%
	Standar/tanpa panduan	4	0,034	-0,945-1,013	91,3%
	Tidak dilaporkan	3	1,127	-2,092-4,345	95,4%

Bidang bahasa dan menulis menghasilkan estimasi tinggi dengan heterogenitas besar, bidang kesehatan menghasilkan estimasi yang lebih rendah dengan heterogenitas rendah, dan bidang sains dan matematika memiliki ketidakpastian terbesar karena jumlah studinya sedikit. Efek terhadap pembandingan non-GenAI lebih tinggi daripada efek terhadap umpan balik manusia, sehingga besar efek berbeda menurut kondisi pembandingan. Studi dengan penggunaan GenAI terstruktur atau berpengaman juga menghasilkan estimasi yang lebih tinggi daripada penggunaan tanpa panduan. Hubungan ini belum dapat dinyatakan sebagai hubungan kausal karena studi terstruktur juga berbeda dalam bidang, luaran, durasi, dan pembandingan. Kategori dengan satu atau dua studi, seperti pemrograman dan komputasi, tidak digabungkan secara formal.

Pembahasan Temuan Utama

Penggunaan GenAI berkaitan dengan peningkatan hasil belajar secara keseluruhan, tetapi besar efeknya bergantung pada konteks penerapan. Tugas menulis memungkinkan umpan balik kebahasaan yang langsung dan berulang sehingga skor dapat meningkat dalam waktu singkat, seperti pada studi Shamsavar dkk. (2024) serta Gan dkk. (2024). Namun, luaran tersebut sering berupa kualitas tulisan yang dikerjakan selama intervensi, sehingga transfer ke tugas baru dan keberlanjutan keterampilan belum tentu terukur.

Retensi pengetahuan menunjukkan perbedaan antara kinerja sesaat dan pembelajaran yang bertahan. Barcaui (2025) melaporkan efek negatif pada tes retensi tertunda, dengan kelompok pembelajaran tradisional memperoleh hasil lebih tinggi daripada kelompok ChatGPT. Heterogenitas yang rendah pada bidang kesehatan dapat berkaitan dengan tes yang lebih terstandar dan intervensi yang lebih spesifik, meskipun jumlah studinya terbatas sehingga temuan ini bersifat eksploratif. Efek yang lebih kecil terhadap umpan balik manusia dapat berkaitan dengan keunggulan GenAI dalam menyediakan umpan balik yang cepat, individual, dan berulang, sedangkan manfaat tambahan di atas umpan balik manusia yang berkualitas lebih terbatas.

Panduan pedagogis merupakan salah satu faktor yang berkaitan dengan variasi efek secara konsisten pada analisis ini. Penggunaan terstruktur mengarahkan peserta didik untuk menggunakan kecerdasan buatan sebagai pemberi petunjuk, umpan balik, atau pendamping berpikir, bukan sebagai sumber jawaban akhir, sehingga keterlibatan kognitif dapat dipertahankan. Bastani dkk. (2025) melaporkan konsekuensi belajar yang berbeda antara tutor GPT-4 berpengaman dan tutor tanpa pengaman setelah akses dihentikan. Dengan demikian, hubungan antara GenAI dan hasil belajar tidak dapat disederhanakan menjadi "positif" atau "negatif" karena kondisi implementasi berperan, sejalan dengan peran bidang, durasi, mode pembelajaran, dan rancangan intervensi sebagai moderator pada penelitian terdahulu.

Risiko Bias dan Kekokohan Temuan

Sebagian besar studi memiliki sejumlah kekhawatiran metodologis terkait randomisasi, penggunaan kelas utuh, pengukuran luaran, atrisi (*attrition*), atau pelaporan statistik, dan studi Alnemrat dkk. (2025) dinilai berisiko bias serius. Penilaian ini digunakan untuk menguji kestabilan sintesis dan tidak digunakan untuk mengeluarkan studi. Estimasi tetap positif setelah beberapa kelompok studi dikeluarkan, sehingga kekokohan (*robustness*) temuan dipahami sebagai kestabilan arah efek, bukan kestabilan besar efek. Uji Egger tidak menunjukkan asimetri yang kuat, tetapi hasil ini tidak membuktikan ketiadaan bias publikasi karena heterogenitas yang tinggi menurunkan daya diagnostik uji tersebut.

Implikasi Temuan

Efektivitas GenAI lebih tepat dipahami sebagai hasil interaksi antara teknologi, rancangan pedagogis, karakteristik tugas, dan kondisi pembelajaran. Penggunaan tanpa struktur tidak dapat diasumsikan memberikan manfaat yang sama dengan intervensi yang menyertakan panduan penggunaan, batasan jawaban, atau proses reflektif. Oleh karena itu, perhatian dapat diarahkan dari adopsi teknologi ke kualitas integrasinya. Implementasi juga perlu menyesuaikan tujuan belajar dan bentuk evaluasi, dengan memprioritaskan luaran yang menilai penerapan pengetahuan secara mandiri atau retensi dalam jangka waktu tertentu, dan tidak hanya kinerja sesaat setelah penggunaan kecerdasan buatan. Heterogenitas yang tinggi merupakan temuan yang perlu dilaporkan bersama efek gabungan.

KESIMPULAN

Hasil meta-analisis menunjukkan bahwa penggunaan GenAI berkaitan dengan hasil belajar peserta didik yang lebih tinggi dibandingkan dengan kondisi pembelajaran pembandingan, dengan besar pengaruh pada kategori sedang menuju besar dan arah yang tetap positif pada seluruh analisis sensitivitas. Pengaruh tersebut bervariasi antarpencarian, sehingga hasil gabungan tidak berlaku sama pada setiap konteks, dan penelitian baru masih dapat menunjukkan pengaruh yang kecil atau negatif. Variasi tersebut berkaitan dengan karakteristik penelitian dan intervensi. Pengaruh GenAI cenderung lebih tinggi pada penggunaan terstruktur dengan panduan atau pengaman pedagogis daripada penggunaan tanpa panduan, lebih tinggi pada perbandingan dengan pembelajaran tanpa GenAI daripada dengan umpan balik manusia, serta berbeda menurut

bidang pembelajaran dan jenis hasil belajar. Pengaruh tampak besar pada kemampuan menulis, tetapi tidak selalu bertahan pada retensi pengetahuan jangka panjang. Temuan ini bersifat eksploratif karena jumlah studi pada sebagian kelompok terbatas dan kualitas metodologis studi primer beragam. Dengan demikian, efektivitas GenAI berkaitan dengan cara teknologi tersebut diintegrasikan ke dalam rancangan pembelajaran. Hasil ini dapat menjadi dasar pertimbangan bagi pendidik dan pengambil kebijakan dalam merancang penggunaan GenAI yang terarah, dan penelitian selanjutnya disarankan menguji hasil belajar jangka panjang pada lebih banyak bidang dan jenjang pendidikan.

REFERENSI

- Alnemrat, A., Aldamen, H., Almashour, M., Al-Deaibes, M., & AlSharefeen, R. (2025). AI vs. teacher feedback on EFL argumentative writing: a quantitative study. *Frontiers in Education*, 10. <https://doi.org/10.3389/educ.2025.1614673>
- Alneyadi, S., & Wardat, Y. (2023). ChatGPT: Revolutionizing student achievement in the electronic magnetism unit for eleventh-grade students in Emirates schools. *Contemporary Educational Technology*, 15(4), ep448–ep448. <https://doi.org/10.30935/cedtech/13417>
- Barcaui, A. (2025). ChatGPT as a cognitive crutch: Evidence from a randomized controlled trial on knowledge retention. *Social Sciences & Humanities Open*, 12, 102287. <https://doi.org/10.1016/j.ssaho.2025.102287>
- Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakçı, Ö., & Mariman, R. (2025). Generative AI without guardrails can harm learning: Evidence from high school mathematics. *Proceedings of the National Academy of Sciences*, 122(26). <https://doi.org/10.1073/pnas.2422633122>
- Cen, X., & Shakibaei, G. (2026). Enhancing creative writing through AI-powered co-creation with cognitive and emotional outcomes. *Scientific Reports*, 16(1), 4407. <https://doi.org/10.1038/s41598-025-34416-2>
- Chen, S., & Cheung, A. C. K. (2025). Effect of generative artificial intelligence on university students learning outcomes: A systematic review and meta-analysis. *Educational Research Review*, 49, 100737. <https://doi.org/10.1016/j.edurev.2025.100737>
- Dai, X., Gong, Y., & Ma, L.-Y. (2026). Effects of ChatGPT-generated immediate feedback integrated into VR-based OSCEs on nursing students' performance: a randomized crossover study. *BMC Nursing*, 25(1), 396. <https://doi.org/10.1186/s12912-026-04633-9>
- de la Puente, M., Torres, J., Troncoso, A. L. B., Meza, Y. Y. H., & Carrascal, J. X. M. (2024). Investigating the use of chatGPT as a tool for enhancing critical thinking and argumentation skills in international relations debates among undergraduate students. *Smart Learning Environments*, 11(1), 55. <https://doi.org/10.1186/s40561-024-00347-0>
- 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>
- El Fathi, T., Saad, A., Larhzil, H., Lamri, D., & Al Ibrahim, E. M. (2025). Integrating generative AI into STEM education: enhancing conceptual understanding, addressing misconceptions, and assessing student acceptance. *Disciplinary and Interdisciplinary Science Education Research*, 7(1), 6. <https://doi.org/10.1186/s43031-025-00125-z>
- Escalante, J., Pack, A., & Barrett, A. (2023). AI-generated feedback on writing: insights into efficacy and ENL student preference. *International Journal of Educational Technology in Higher Education*, 20(1), 57. <https://doi.org/10.1186/s41239-023-00425-2>
- Gan, W., Ouyang, J., Li, H., Xue, Z., Zhang, Y., Dong, Q., Huang, J., Zheng, X., & Zhang, Y. (2024). Integrating ChatGPT in Orthopedic Education for Medical Undergraduates: Randomized Controlled Trial. *Journal of Medical Internet Research*, 26, e57037. <https://doi.org/10.2196/57037>
- Harrer, M., Cuijpers, P., Furukawa, T. A., & Ebert, D. D. (2021). *Doing Meta-Analysis with R*. Chapman and Hall/CRC. <https://doi.org/10.1201/9781003107347>

- Heung, Y. M. E., & Chiu, T. K. F. (2025). How ChatGPT impacts student engagement from a systematic review and meta-analysis study. *Computers and Education: Artificial Intelligence*, 8, 100361. <https://doi.org/10.1016/j.caeai.2025.100361>
- Higgins, J. P. T., Thomas, J., Chandler, J., Cumpston, M., Li, T., & Page, M. J. (2024). *Cochrane Handbook for Systematic Reviews of Interventions version 6.5*. Cochrane. <https://www.cochrane.org/handbook>
- Hsu, L. (2025). Entangled cognition in EFL education: The role of generative AI. *Contemporary Educational Technology*, 17(4), ep619–ep619. <https://doi.org/10.30935/cedtech/17621>
- Hu, H., Gong, Q., & Hashim, H. (2026). The Impact of ChatGPT-Generated Feedback on English Academic Writing Proficiency and Engagement: A Mixed-Methods Study. *Sage Open*, 16(3). <https://doi.org/10.1177/21582440261480859>
- Huang, S., Wen, C., Bai, X., Li, S., Wang, S., Wang, X., & Yang, D. (2025). Exploring the Application Capability of ChatGPT as an Instructor in Skills Education for Dental Medical Students: Randomized Controlled Trial. *Journal of Medical Internet Research*, 27, e68538–e68538. <https://doi.org/10.2196/68538>
- Huu Hoang, N. (2025). ChatGPT-assisted language learning: Effects on Vietnamese English majors' writing skills and motivation. *The JALT CALL Journal*, 21(2), 102429. <https://doi.org/10.29140/jaltcall.v21n2.102429>
- Kestin, G., Miller, K., Klales, A., Milbourne, T., & Ponti, G. (2025). AI tutoring outperforms in-class active learning: an RCT introducing a novel research-based design in an authentic educational setting. *Scientific Reports*, 15(1), 17458. <https://doi.org/10.1038/s41598-025-97652-6>
- Liu, Z., Zuo, H., & Lu, Y. (2025). The Impact of <scp>ChatGPT</scp> on Students' Academic Achievement: A Meta-Analysis. *Journal of Computer Assisted Learning*, 41(4). <https://doi.org/10.1111/jcal.70096>
- Lu, D. (Jason), & Zeng, Y. (2025). Exploring the use of ChatGPT-generated model texts as a feedback instrument: EFL students' text quality and perceptions. *Innovation in Language Learning and Teaching*, 1–21. <https://doi.org/10.1080/17501229.2025.2525341>
- Maria Ilyas, Rehan Ahmed Khan, Rahila Yasmeen, Zainab Kamal, & Ul Ain, N. (2026). Impact of ChatGPT-assisted personalized learning on teaching acute abdomen to undergraduate medical students: A randomized crossover study. *Pakistan Journal of Medical Sciences*, 42(8), 1943–1948. <https://doi.org/10.12669/pjms.42.8.14827>
- Min Young, D., & Park, Y. (2026). Meta-Analysis of ChatGPT's Influence on Learning Achievement. *The International Review of Research in Open and Distributed Learning*, 27(1), 265–289. <https://doi.org/10.19173/irrodl.v27i1.8775>
- Nathaniel, J., Oyelere, S. S., Suhonen, J., & Tedre, M. (2026). An experimental study of structured generative AI integration to mitigate pedagogical, cognitive, and ethical barriers in programming education. *Frontiers in Computer Science*, Volume 8-2026. <https://doi.org/10.3389/fcomp.2026.1789829>
- Ng, D. T. K., Tan, C. W., & Leung, J. K. L. (2024). Empowering student self-regulated learning and science education through <scp>ChatGPT</scp>: A pioneering pilot study. *British Journal of Educational Technology*, 55(4), 1328–1353. <https://doi.org/10.1111/bjet.13454>
- Niloy, A. C., Akter, S., Sultana, N., Sultana, J., & Rahman, S. I. U. (2024). Is Chatgpt a menace for creative writing ability? An experiment. *Journal of Computer Assisted Learning*, 40(2), 919–930. <https://doi.org/10.1111/jcal.12929>
- Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., ... Moher, D. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, n71–n71. <https://doi.org/10.1136/bmj.n71>
- Pustejovsky, J. E., & Tipton, E. (2022). Meta-analysis with Robust Variance Estimation: Expanding the Range of Working Models. *Prevention Science*, 23(3), 425–438. <https://doi.org/10.1007/s11121-021-01246-3>

- Roza, V., Melani, M., reflinda, R., Rianto, R., Ahida, R., Arsyad, S., Sakti, G., & Mellian, Y. (2026). AI vs. Human Feedback in EFL Writing: Examining the Interplay Between Feedback Modalities and Digital Proficiency. *Teaching English as a Second or Foreign Language--TESL-EJ*, 30(1). <https://doi.org/10.55593/ej.30117int2>
- Shahsavar, Z., Kafipour, R., Khojasteh, L., & Pakdel, F. (2024). Is artificial intelligence for everyone? Analyzing the role of ChatGPT as a writing assistant for medical students. *Frontiers in Education*, 9. <https://doi.org/10.3389/educ.2024.1457744>
- Sun, L., & Zhou, L. (2024). Does Generative Artificial Intelligence Improve the Academic Achievement of College Students? A Meta-Analysis. *Journal of Educational Computing Research*, 62(7), 1676–1713. <https://doi.org/10.1177/07356331241277937>
- Wu, X., Zhu, P., Zhang, J., Yin, M., & Wang, Y. (2026). ChatGPT's impact on student learning outcomes: a meta-analysis of 35 experimental studies. *Humanities and Social Sciences Communications*, 13(1), 684. <https://doi.org/10.1057/s41599-026-07019-z>
- Yang, C. C. Y., Li, M., & Huang, A. Y. Q. (2026). Leveraging Large Language Models to Enhance Self-Regulated Learning in Programming Education With Explainable <scp>AI</scp>. *Journal of Computer Assisted Learning*, 42(2). <https://doi.org/10.1002/jcal.70206>
- Yang, T.-C., Hsu, Y.-C., & Wu, J.-Y. (2025). The effectiveness of ChatGPT in assisting high school students in programming learning: evidence from a quasi-experimental research. *Interactive Learning Environments*, 33(6), 3726–3743. <https://doi.org/10.1080/10494820.2025.2450659>
- Zhang, K. (2025). Enhancing Critical Writing Through AI Feedback: A Randomized Control Study. *Behavioral Sciences*, 15(5), 600. <https://doi.org/10.3390/bs15050600>