Jurnal Pendidikan MIPA
Volume 16. Nomor 5, Tahun 2026 | ISSN: 2088-0294 | e-ISSN: 2621-9166
https://doi.org/10.37630/jpm.v16i5.5237

Generative AI dalam Pembelajaran Pembuktian Matematis: Tinjauan Sistematis terhadap Peluang, Keterbatasan, dan Implikasi Pedagogis

Yan Sando Purdatama Rija[1),*], **Marnilawati**[1)]

[1)]Fakultas Keguruan dan Ilmu Pendidikan, Universitas Tanjungpura

*Corresponding Author: f2181251016@student.untan.ac.id

ABSTRAK

Generative Artificial Intelligence (GenAI) semakin banyak dimanfaatkan dalam pembelajaran matematika, padahal pembuktian matematis menuntut ketelitian logis yang justru menjadi kelemahan model bahasa besar. Penelitian ini bertujuan memetakan karakteristik, peluang, keterbatasan, dan implikasi pedagogis pemanfaatan GenAI dalam pembelajaran pembuktian matematis. Metode yang digunakan adalah *systematic literature review* yang mengikuti pedoman PRISMA 2020. Pencarian pada basis data Scopus menghasilkan 140 rekaman, dan setelah penyaringan serta penilaian kelayakan, 16 studi yang terbit pada 2024–2026 dianalisis secara deskriptif dan tematik. Hasil menunjukkan bahwa studi didominasi penelitian kualitatif berskala kecil pada jenjang perguruan tinggi, dan hanya satu studi yang dilakukan pada jenjang sekolah. Penilaian kualitas dengan MMAT 2018 menunjukkan bahwa sebagian besar studi (sembilan dari 16) hanya memenuhi dua kriteria atau kurang. GenAI berpeluang menjadi sumber gagasan pembuktian, pemberi umpan balik dalam skala besar, *scaffold* penalaran, dan sarana latihan berpikir kritis, terutama untuk fungsi komunikasi dan penemuan bukti. Namun, GenAI kerap menghasilkan bukti yang tidak sahih, memberikan umpan balik yang dangkal, dan berisiko menimbulkan ketergantungan sehingga fungsi verifikasi bukti paling terancam. Kajian ini merumuskan enam implikasi pedagogis dengan prinsip bahwa GenAI perlu diposisikan sebagai mitra belajar yang keluarannya selalu diverifikasi. Penelitian lanjutan direkomendasikan pada jenjang sekolah, konteks negara berkembang, dan desain eksperimental.

Kata Kunci: ChatGPT; Generative AI; Pembuktian Matematis; Penalaran Deduktif; Systematic Literature Review

This is an open access article under the CC - BY license.

PENDAHULUAN

Pembuktian menempati posisi sentral dalam matematika karena menjadi cara utama untuk menetapkan kebenaran suatu pernyataan (Hanna, 2000). Fungsi bukti mencakup verifikasi, eksplanasi, sistematisasi, penemuan, dan komunikasi (de Villiers, 1990), sedangkan bukti yang bernilai pedagogis adalah bukti yang menjelaskan mengapa suatu pernyataan benar (Hanna, 2000). Stylianides (2007) menegaskan bahwa pembuktian perlu diajarkan pada berbagai jenjang, dengan syarat argumen menggunakan pernyataan yang telah diterima, bentuk penalaran yang sahih, serta representasi yang dipahami bersama. Meskipun demikian, kemampuan menyusun bukti masih menjadi kesulitan yang terus dilaporkan. Banyak siswa meyakini kebenaran pernyataan umum berdasarkan otoritas eksternal atau beberapa contoh empiris, bukan berdasarkan penalaran deduktif (Harel & Sowder, 1998). Balacheff (1988) menyebut kecenderungan menyimpulkan kebenaran umum dari beberapa kasus tersebut sebagai empirisme naif.

Pembuktian juga berkaitan erat dengan argumentasi matematis. Argumentasi yang dibangun siswa saat menyusun dugaan sering menjadi landasan konstruksi bukti, meskipun peralihan dari argumentasi ke bukti tidak selalu berlangsung mulus (Pedemonte, 2007). Kedekatan konseptual tersebut menjadi dasar kajian ini untuk mengikutsertakan studi argumentasi matematis yang bertujuan menjustifikasi klaim secara deduktif.

Peluncuran ChatGPT pada November 2022 menandai babak baru pemanfaatan kecerdasan buatan dalam pendidikan. Model bahasa besar (*large language model*/LLM) mampu menghasilkan penjelasan, contoh, dan umpan balik dalam bahasa alami secara instan sehingga dipandang berpotensi mendukung pembelajaran yang dipersonalisasi (Kasneci et al., 2023). Penelitian tentang integrasi kecerdasan buatan dalam pendidikan matematika pun meningkat pesat (Nguyen & Pham, 2025), dan tinjauan terhadap pemanfaatan ChatGPT

ISSN: 2088-0294 | e-ISSN: 2621-9166
https://doi.org/10.37630/jpm.v16i5.5237

melaporkan manfaatnya bagi perencanaan pembelajaran, dukungan belajar, dan motivasi siswa (Almarashdi et al., 2024).

Namun, karakteristik LLM menimbulkan persoalan khusus ketika diterapkan pada pembuktian. LLM menghasilkan teks berdasarkan pola probabilistik bahasa, bukan melalui penalaran deduktif yang terverifikasi (Bender et al., 2021), sehingga rentan mengalami halusinasi, yakni menghasilkan keluaran yang terdengar meyakinkan tetapi tidak sesuai dengan fakta atau logika (Ji et al., 2023). Evaluasi terhadap ChatGPT menunjukkan bahwa kemampuan matematisnya, termasuk dalam menyusun bukti, masih berada di bawah kemampuan mahasiswa pascasarjana matematika pada umumnya (Frieder et al., 2023). Persoalan tersebut krusial karena satu langkah yang tidak sahih sudah cukup untuk menggugurkan keseluruhan bukti. GenAI juga berpotensi menjadi otoritas eksternal baru bagi siswa sekaligus menghasilkan argumen yang bersifat empiris.

Sejumlah studi empiris mulai menelaah pemanfaatan GenAI pada aktivitas pembuktian dengan temuan yang beragam. Yoon et al. (2024) menemukan bahwa penggunaan GenAI oleh mahasiswa saat membuktikan dipengaruhi oleh konsepsi tentang bukti, konsepsi tentang GenAI, dan pertimbangan etis. Davor (2026) melaporkan peningkatan kemampuan konstruksi bukti geometri siswa SMA melalui pembelajaran berbantuan GenAI, sedangkan Agad et al. (2025) mendapati bahwa mahasiswa kurang memercayai umpan balik ChatGPT atas bukti mereka. Pengembangan sistem tutor dan penilai bukti berbasis LLM juga mulai dilaporkan (Patel et al., 2026; Zhao et al., 2025). Keragaman temuan tersebut menunjukkan bahwa peran GenAI dalam pembelajaran pembuktian belum dipahami secara utuh.

Upaya sintesis literatur tentang GenAI dalam pendidikan matematika sebenarnya telah dilakukan, tetapi cakupannya masih umum. Almarashdi et al. (2024) mengulas pemanfaatan ChatGPT dalam pembelajaran matematika secara luas, Nguyen dan Pham (2025) memetakan tren integrasi kecerdasan buatan selama 2015–2025, sedangkan Dela Torre dan Sayco (2026) menyintesis manfaat GenAI bagi pemecahan masalah dan umpan balik. Muhamad Sani et al. (2026) mencatat bahwa penelitian GenAI dalam pendidikan matematika terkonsentrasi pada jenjang perguruan tinggi serta topik aljabar dan kalkulus. Sejauh penelusuran penulis, belum ada tinjauan sistematis yang secara khusus menyintesis temuan tentang GenAI dalam pembelajaran pembuktian matematis, padahal studi empiris pada topik ini terus bertambah sejak 2024. Kajian khusus tersebut diperlukan karena pembuktian menuntut ketelitian logis pada setiap langkah, sedangkan kelemahan utama LLM justru terletak pada kesahihan logis tersebut. Kesimpulan umum tentang manfaat GenAI dalam pendidikan matematika belum tentu berlaku pada pembelajaran pembuktian.

Berdasarkan uraian tersebut, penelitian ini bertujuan memetakan dan menyintesis hasil penelitian tentang pemanfaatan GenAI dalam pembelajaran pembuktian matematis melalui *systematic literature review* (SLR). Penelitian dipandu oleh empat pertanyaan, yaitu (RQ1) bagaimana karakteristik studi ditinjau dari tahun publikasi, negara, jenjang, desain penelitian, dan alat GenAI; (RQ2) peluang apa yang ditawarkan GenAI; (RQ3) keterbatasan dan risiko apa yang dilaporkan; serta (RQ4) implikasi pedagogis apa yang dapat dirumuskan bagi guru dan dosen. Kajian ini diharapkan menyajikan peta penelitian yang khusus memfokuskan pada GenAI dan pembuktian, menjadi dasar rekomendasi pedagogis yang konkret, serta mengidentifikasi agenda penelitian selanjutnya, termasuk pada jenjang sekolah dan konteks Indonesia.

METODE

Penelitian ini menggunakan desain SLR dengan pelaporan yang mengikuti pedoman PRISMA 2020 (Page et al., 2021). Scopus dipilih karena merupakan basis data bibliografis multidisiplin yang mengindeks jurnal dan prosiding bereputasi, termasuk publikasi pendidikan matematika dan ilmu komputer yang relevan dengan topik kajian. Pencarian dilakukan pada 21 Oktober 2026 melalui fitur *Advanced Search* pada judul, abstrak, dan kata kunci. *Search string* disusun dari tiga blok konsep, yaitu istilah GenAI, istilah aktivitas pembuktian, dan istilah konteks pendidikan, dengan rentang tahun 2022–2026 karena ChatGPT dirilis pada November 2022. *Search string* lengkap yang digunakan adalah sebagai berikut.

TITLE-ABS-KEY (("generative AI" OR "generative artificial intelligence" OR "large language model*" OR "LLM" OR "ChatGPT" OR "GPT-4" OR "chatbot*" OR "Gemini" OR "Copilot") AND ("mathematical proof*" OR "proof writing" OR "proof construction" OR "proving" OR "proof-based" OR "mathematical reasoning" OR "mathematical argumentation") AND ("mathematics education" OR "mathematics learning" OR "mathematics teaching" OR

ISSN: 2088-0294 | e-ISSN: 2621-9166
https://doi.org/10.37630/jpm.v16i5.5237

"mathematics course*" OR student* OR teacher* OR undergraduate*)) AND PUBYEAR > 2021 AND PUBYEAR < 2027

Hasil pencarian dibatasi pada artikel jurnal dan makalah prosiding berbahasa Inggris, lalu diekspor dalam format CSV. Studi diikutsertakan apabila (1) membahas GenAI atau LLM dalam konteks belajar atau mengajar pembuktian maupun penalaran deduktif matematis, (2) terbit pada 2022–2026, (3) berupa artikel jurnal atau makalah prosiding, (4) berbahasa Inggris, dan (5) teks lengkapnya dapat diakses. Kriteria pertama diperjelas melalui dua aturan lingkup. Aturan pertama menetapkan bahwa studi argumentasi diikutsertakan apabila peserta menjustifikasi klaim matematis secara deduktif, sedangkan aturan kedua menetapkan bahwa pembuktian formal pada mata kuliah ilmu komputer, seperti bukti induksi, logika proposisional, dan ketakterputusan (*undecidability*), dipandang sebagai pembuktian matematis. Studi dikeluarkan apabila (E1) merupakan studi teknis kecerdasan buatan tanpa konteks pembelajaran, (E2) berada dalam konteks pendidikan di luar matematika, (E3) tidak berkaitan dengan pembuktian atau penalaran deduktif, (E4) tidak memuat data penelitian primer, atau (E5) tidak mengkaji GenAI.

Seleksi studi berlangsung melalui tahap identifikasi, penyaringan judul dan abstrak, pengambilan teks lengkap, dan penilaian kelayakan. Duplikat diperiksa berdasarkan judul dan DOI. Rekaman yang belum dapat diputuskan dari abstrak diberi label "ragu" dan diperiksa melalui teks lengkap. Teks lengkap diupayakan melalui akses institusi, repositori akses terbuka, dan permintaan kepada penulis. Penyaringan dan penilaian kelayakan dilakukan oleh dua peninjau secara independen, dan perbedaan keputusan diselesaikan melalui diskusi. Tingkat kesepakatan antarpeninjau pada tahap penyaringan mencapai 95,3% (122 dari 128 rekaman), dengan koefisien kappa Cohen sebesar 0,81. Penyaringan awal judul dan abstrak dibantu Claude (Anthropic; model claude-haiku-5-5) untuk memberikan usulan label Include, Ragu, atau Exclude beserta kode eksklusi E1–E5. Instruksi yang diberikan memuat kriteria inklusi dan eksklusi tersebut serta aturan bahwa rekaman yang tidak dapat diputuskan dari abstrak diberi label Ragu; prompt lengkap disertakan sebagai materi tambahan. Dari 128 rekaman, usulan awal berjumlah 9 Include, 15 Ragu, dan 104 Exclude. Seluruh usulan diverifikasi dan diputuskan oleh kedua peninjau manusia. Perubahan hanya terjadi pada 15 usulan Ragu, yang diputuskan menjadi 7 Include dan 8 Exclude, sedangkan tidak ada usulan Include atau Exclude yang diubah. Claude juga membantu menyusun lembar ekstraksi dan draf penilaian MMAT; seluruh hasilnya diperiksa ulang oleh penulis.

Data diekstraksi menggunakan lembar yang memuat identitas studi, negara, jenjang atau partisipan, jumlah partisipan, desain penelitian, alat GenAI, jenis bukti, peran GenAI, serta temuan yang dikelompokkan ke dalam peluang (RQ2), keterbatasan (RQ3), dan implikasi pedagogis (RQ4). Karakteristik metodologis, seperti ukuran sampel, jenis publikasi, dan versi model GenAI, juga dicatat untuk mendukung penafsiran temuan. Setiap studi diberi kode S1–S16. RQ1 dijawab melalui analisis deskriptif, sedangkan RQ2–RQ4 dijawab melalui sintesis tematik (Thomas & Harden, 2008). Temuan setiap studi dikodekan secara induktif, dikelompokkan menjadi tema, lalu ditafsirkan menggunakan dua lensa, yaitu fungsi bukti (de Villiers, 1990) dan peran GenAI dalam pembelajaran.

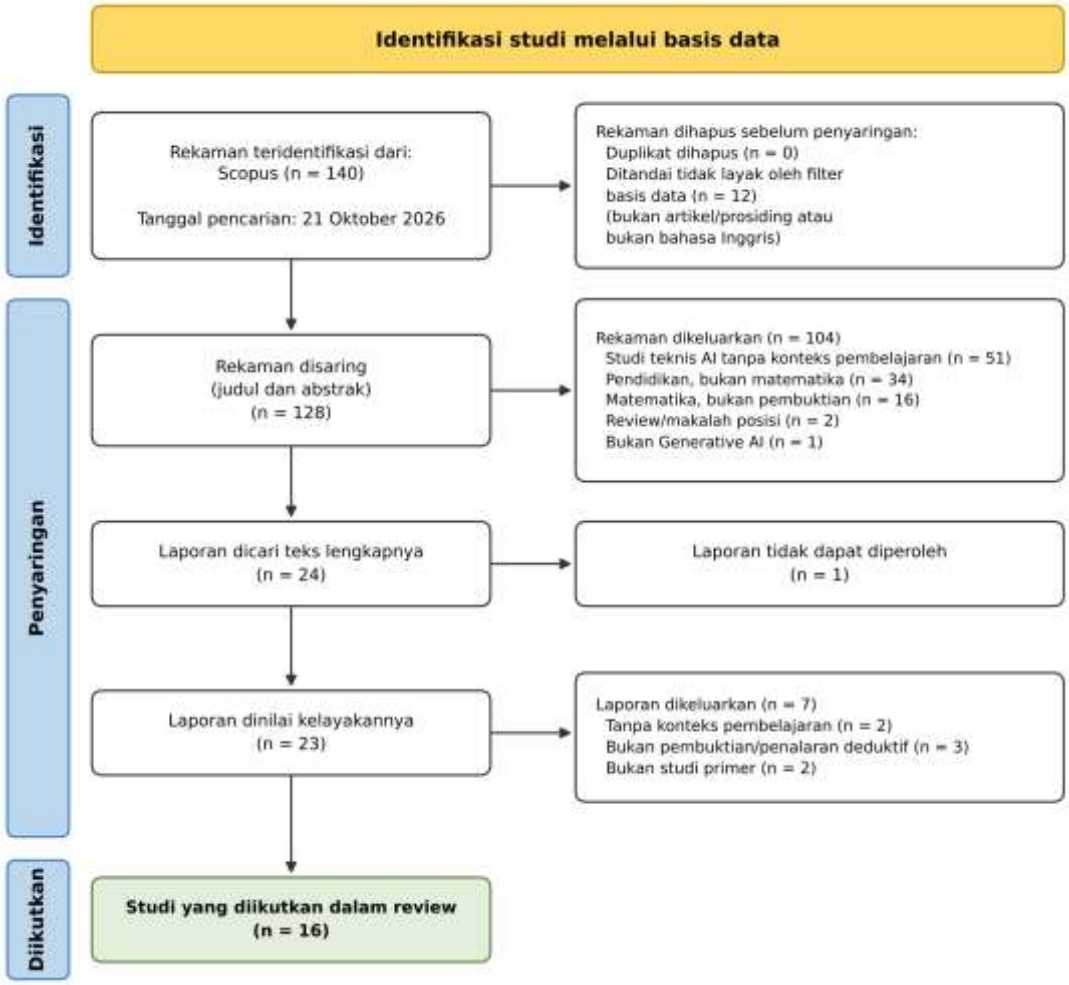
Kualitas metodologis setiap studi dinilai menggunakan *Mixed Methods Appraisal Tool* (MMAT) versi 2018 (Hong et al., 2018). MMAT terdiri atas dua pertanyaan skrining dan lima kriteria untuk setiap kategori desain, sedangkan skor studi adalah jumlah kriteria yang terpenuhi dari lima. Untuk studi yang tidak melibatkan partisipan manusia, MMAT hanya digunakan secara indikatif karena instrumen ini tidak dirancang untuk mengevaluasi kinerja sistem atau model.

HASIL DAN PEMBAHASAN

Pencarian pada Scopus menghasilkan 140 rekaman. Filter basis data mengeluarkan 12 rekaman, dan tidak ditemukan duplikat sehingga 128 rekaman disaring berdasarkan judul dan abstrak. Penyaringan tersebut mengeluarkan 104 rekaman yang sebagian besar merupakan studi teknis kecerdasan buatan tanpa konteks pembelajaran (n = 51). Rekaman lainnya dikeluarkan karena berada dalam konteks pendidikan di luar matematika (n = 34), tidak berkaitan dengan pembuktian (n = 16), berupa tinjauan atau makalah posisi (n = 2), dan tidak mengkaji GenAI (n = 1). Tingginya angka tersebut menunjukkan bahwa istilah "penalaran matematis" banyak digunakan dalam riset kecerdasan buatan untuk menguji kemampuan model, bukan untuk mengkaji pembelajaran. Sebanyak 24 laporan dicari teks lengkapnya, satu laporan tidak dapat diperoleh, dan penilaian kelayakan terhadap 23 laporan mengeluarkan tujuh laporan, yaitu dua laporan tanpa konteks pembelajaran, tiga

ISSN: 2088-0294 | e-ISSN: 2621-9166
https://doi.org/10.37630/jpm.v16i5.5237

laporan yang tidak berkaitan dengan pembuktian, dan dua laporan tanpa data penelitian primer. Akhirnya, 16 studi diikutsertakan dalam kajian. Alur seleksi studi disajikan pada Gambar 1.



Gambar 1. Diagram alir PRISMA 2020 seleksi studi

Karakteristik ke-16 studi dirangkum pada Tabel 1. Seluruh studi terbit pada 2024–2026, dengan lima studi pada 2024, tiga pada 2025, dan delapan pada 2026, sedangkan tidak ada studi tahun 2022–2023 yang memenuhi kriteria. Separuh studi terbit pada 2026 meskipun pencarian hanya mencakup publikasi hingga Oktober 2026 sehingga bidang ini tergolong baru dan berkembang cepat. Sepuluh studi terbit di jurnal dan enam dalam prosiding, dua di antaranya berupa poster dua halaman (S15 dan S16).

Tabel 1. Karakteristik studi yang dikaji

Kode	Penulis (Tahun)	Negara	Partisipan	n	Desain	Alat GenAI	Peran GenAI	Jenis bukti dan konteks
S1	Drushlyak et al. (2024)	Ukraina	Calon guru matematika	17	Eksperimen pedagogis satu kelompok	ChatGPT 3.5	Objek evaluasi	Ketaksamaan
S2	Dilling dan Herrmann (2024)	Jerman	Calon guru SD	129	Kualitatif eksploratif	ChatGPT 3.5/4	Asisten pembuktian	Geometri
S3	Yoon et al. (2024)	Korea Selatan	Mahasiswa S1	3	Kualitatif eksploratif	ChatGPT 3.5	Asisten / mitra bernalar	Teori bilangan
S4	Poulsen et al. (2024)	Amerika Serikat	Mahasiswa S1	–	Evaluasi kinerja LLM	GPT-4, GPT-3.5	Penyelesai soal	Matematika diskret (IK)
S5	Urhan et al. (2024)	Turki	Mahasiswa S1	1	Studi kasus kualitatif	ChatGPT	Asisten / mitra bernalar	Argumentasi kalkulus
S6	Agad et al. (2025)	Filipina	Mahasiswa S1	6	Kualitatif	ChatGPT	Umpan balik / penilai	Geometri proyektif
S7	Zhao et al. (2025)	Amerika Serikat	Mahasiswa S1	169	Pengembangan sistem dan studi pengguna	LLM (Llemma, GPT-3)	Umpan balik / penilai	Induksi matematis (IK)

ISSN: 2088-0294 | e-ISSN: 2621-9166
https://doi.org/10.37630/jpm.v16i5.5237

Kode	Penulis (Tahun)	Negara	Partisipan	n	Desain	Alat GenAI	Peran GenAI	Jenis bukti dan konteks
S8	Tithi et al. (2025)	Amerika Serikat	Mahasiswa S1	–	Evaluasi kinerja LLM	GPT-4o, DeepSeek-V3, dll.	Pembuat petunjuk	Logika proposisional (IK)
S9	Kartika & Clark (2026)	Indonesia	Mahasiswa S1	94	Kualitatif komparatif	ChatGPT 3.5, YouChat	Objek evaluasi	Teori bilangan
S10	Patel et al. (2026)	Amerika Serikat	Mahasiswa S1 (sasaran)	–	Pengembangan sistem	LLM dan Lean	Tutor	Aritmetika Peano (IK)
S11	Davor (2026)	Ghana	Siswa SMA	86	Kuasi-eksperimen dengan kelompok kontrol	Tidak disebutkan	Scaffold	Geometri
S12	Yoon et al. (2026)	Korea Selatan	Mahasiswa S1	30	Perbandingan dua kondisi (desain silang)	ChatGPT	Asisten / mitra bernalar	Pernyataan matematis
S13	Urhan et al. (2026)	Turki	Mahasiswa S1	2	Studi kasus kualitatif	ChatGPT 3.5	Asisten / mitra bernalar	Argumentasi kalkulus
S14	Zhuang et al. (2026)	Amerika Serikat	Calon guru matematika	10	Kualitatif	Chatbot berbasis ChatGPT	Tutor (simulasi)	Argumentasi kolektif
S15	Frable et al. (2026)	Amerika Serikat	Mahasiswa S1	–	Evaluasi kinerja LLM	GPT-4	Tutor	Teori komputasi (IK)
S16	Dykstra et al. (2026)	Amerika Serikat	Mahasiswa S1	–	Evaluasi kinerja LLM	ChatGPT-4	Umpan balik / penilai	Teori komputasi (IK)

Keterangan: negara ditentukan berdasarkan afiliasi penulis pertama; tanda "–" menunjukkan studi tanpa partisipan manusia secara langsung. (IK) = konteks mata kuliah ilmu komputer (aturan lingkup kedua). Peran GenAI dikategorikan penulis berdasarkan deskripsi setiap studi.

Studi berasal dari delapan negara dengan dominasi Amerika Serikat (n = 7), disusul Korea Selatan dan Turki (masing-masing n = 2), serta Jerman, Ukraina, Filipina, Ghana, dan Indonesia (masing-masing n = 1). Sebaran jenjangnya sangat timpang: dua belas studi melibatkan atau menyasar mahasiswa S1, tiga studi melibatkan calon guru, dan hanya satu studi yang dilakukan pada jenjang sekolah (S11), sedangkan jenjang SMP dan SD belum tersentuh. Desain kualitatif paling banyak digunakan (n = 7), umumnya dengan partisipan yang sedikit, bahkan tiga studi hanya melibatkan satu sampai tiga orang (S3, S5, S13). Empat studi mengevaluasi kinerja LLM tanpa partisipan langsung, dua studi mengembangkan sistem, dan hanya S11 yang menggunakan kelompok kontrol. S1 merupakan eksperimen pedagogis satu kelompok yang membandingkan skor siswa sebelum dan sesudah perlakuan, sedangkan S12 membandingkan dua kondisi pembelajaran dengan desain silang tanpa kelompok kontrol. Model GPT dari OpenAI digunakan dalam 14 studi, lima di antaranya secara eksplisit memakai ChatGPT 3.5, sedangkan model lain hanya muncul sesekali, seperti YouChat (S9), DeepSeek-V3 (S8), LLM khusus matematika Llemma (S7), dan kombinasi LLM dengan pembukti teorema Lean (S10). Satu studi bahkan tidak menyebutkan alat yang digunakan (S11). Topik bukti yang paling sering dikaji adalah geometri (S2, S6, S11). Peran GenAI yang paling sering ditemukan adalah sebagai asisten pembuktian atau mitra bernalar (n = 5), disusul pemberi umpan balik dan penilai (n = 3) serta tutor (n = 3). Peran tersebut dikategorikan oleh penulis berdasarkan deskripsi setiap studi dan disajikan pada Tabel 1. Enam studi (S4, S7, S8, S10, S15, dan S16) berasal dari konteks mata kuliah ilmu komputer dan ditandai (IK) pada Tabel 1.

Pola tersebut sejalan dengan temuan Muhamad Sani et al. (2026) bahwa penelitian GenAI dalam pendidikan matematika terkonsentrasi pada jenjang perguruan tinggi dan didominasi oleh ChatGPT. Ketimpangan jenjang menjadi persoalan penting karena pembuktian telah diperkenalkan sejak jenjang sekolah (Stylianides, 2007), sementara siswa sekolah semakin mudah mengakses GenAI. Sedikitnya studi dari negara berkembang, termasuk Indonesia, juga menunjukkan bahwa bukti empiris tentang topik ini masih berasal dari konteks pendidikan yang terbatas.

Sintesis tematik menghasilkan enam tema peluang, lima tema keterbatasan, dan enam implikasi pedagogis yang dirangkum pada Tabel 2. Peluang pertama adalah GenAI sebagai sumber gagasan dan titik berangkat penyusunan bukti. Mahasiswa pada studi Yoon et al. (2024) menggunakan argumen ChatGPT sebagai titik awal, lalu mengembangkan bukti dengan strateginya sendiri, bahkan seorang partisipan mengadopsi pembuktian dengan kontradiksi setelah berinteraksi dengan ChatGPT. Yoon et al. (2026) menunjukkan bahwa manfaat tersebut bergantung pada waktu penggunaan. Penggunaan GenAI di awal membuat proses lebih efisien, sedangkan penggunaan setelah penyusunan draf mandiri membantu mahasiswa meninjau dan merevisi bukti.

ISSN: 2088-0294 | e-ISSN: 2621-9166
https://doi.org/10.37630/jpm.v16i5.5237

Tabel 2. Sintesis peluang, keterbatasan, dan implikasi pedagogis

Aspek	Tema	Studi
Peluang (RQ2)	Sumber gagasan dan titik berangkat penyusunan bukti	S3, S12, S13
	Umpan balik dan penilaian bukti dalam skala besar †	S7, S8, S15, S16
	Scaffold penalaran dan konstruksi bukti	S11
	Kesalahan GenAI sebagai bahan latihan berpikir kritis	S1, S13
	Dukungan terhadap komunikasi dan struktur bukti	S5, S6, S13
	Simulasi pedagogis bagi calon guru	S14
Keterbatasan (RQ3)	Ketidakandalan logika dalam bukti yang dihasilkan	S1, S3, S5, S9, S13
	Kinerja menurun pada bukti yang kompleks dan formal †	S4, S8, S10, S15
	Umpan balik dan penilaian yang dangkal serta tidak konsisten	S6, S7, S16
	Ketergantungan dan penerimaan pasif terhadap keluaran GenAI	S3, S12, S13
	Literasi AI yang rendah dan pertimbangan etis	S2, S3
Implikasi (RQ4)	Mengatur waktu penggunaan GenAI	S12
	Memosisikan GenAI sebagai *scaffold*, bukan penyedia bukti	S11, S15
	Menjadikan validasi bukti GenAI sebagai aktivitas belajar	S1, S9, S13
	Mengembangkan literasi AI dan keterampilan *prompting*	S2, S14
	Mempertahankan peran sentral guru dan dosen	S6, S7, S11, S16
	Mengembangkan GenAI yang keluarannya dapat diverifikasi †	S7, S8, S10

Keterangan: † = tema yang seluruh studinya berasal dari konteks mata kuliah ilmu komputer.

Peluang kedua adalah umpan balik dan penilaian bukti dalam skala besar. Sistem penilai berbasis LLM untuk bukti induksi yang dikembangkan Zhao et al. (2025) lebih akurat daripada sebagian besar penilai manusia, dan studi pengguna terhadap 169 mahasiswa menunjukkan bahwa mahasiswa mampu memperbaiki bukti secara signifikan. Tithi et al. (2025) melaporkan akurasi penyusunan bukti logika proposisional hingga 86,7% dan akurasi petunjuk (*hint*) sebesar 75% untuk 1.050 kondisi pengerjaan siswa. Kritik GPT-4 terhadap bukti siswa pada mata kuliah teori komputasi juga dinilai cukup akurat (Frable et al., 2026), dan LLM dapat membantu dosen menilai tugas pembuktian pada kelas besar (Dykstra et al., 2026).

Peluang ketiga adalah GenAI sebagai *scaffold* penalaran dan konstruksi bukti. Davor (2026) memberikan satu-satunya bukti eksperimental: siswa SMA yang belajar dengan dukungan GenAI memperoleh skor penalaran dan konstruksi bukti geometri yang lebih tinggi secara signifikan dibandingkan kelompok kontrol, $F(1, 83) = 50{,}30$, $p < 0{,}001$, dengan ukuran efek yang besar. Kualitas metodologis studi ini terbatas (skor MMAT 2 dari 5), dengan sampel dari satu sekolah dan kovariat hanya pra-tes, sehingga temuan tersebut perlu diuji ulang. Peran GenAI pada studi tersebut dibatasi sebagai penyaji representasi dinamis dan pengaju *prompt* terarah yang menuntut siswa menjustifikasi hubungan geometris, serta sengaja dirancang agar tidak memberikan bukti lengkap.

Peluang keempat adalah pemanfaatan kesalahan GenAI sebagai bahan latihan berpikir kritis. Kegiatan menganalisis dan memperbaiki bukti ketaksamaan yang keliru dari ChatGPT berpengaruh positif terhadap kemampuan berpikir kritis 17 calon guru matematika (Drushlyak et al., 2024). Mahasiswa berkemampuan tinggi juga mampu menyaring kesalahan respons ChatGPT dan menjadikannya sumber inspirasi (Urhan et al., 2026).

Peluang kelima adalah dukungan terhadap komunikasi dan struktur bukti. Revisi mahasiswa setelah menerima umpan balik ChatGPT lebih banyak memperbaiki cara mengungkapkan dan menata argumen daripada penalarannya, termasuk mengingatkan asumsi yang terlewat (Agad et al., 2025). ChatGPT juga memberikan dukungan rasional terhadap komponen jaminan (*warrant*) argumen yang berkaitan dengan definisi kontinuitas, turunan, dan garis singgung (Urhan et al., 2024). Peluang keenam muncul pada pendidikan calon guru, yaitu chatbot berbasis ChatGPT yang membantu sebagian besar calon guru memahami argumentasi matematis kolektif dan langkah guru yang memfasilitasinya (Zhuang et al., 2026).

Pola peluang tersebut dapat dipahami melalui karakteristik LLM. Model yang dilatih untuk menghasilkan teks yang lazim dan runtut wajar menunjukkan keunggulan pada aspek penyajian argumen, seperti merumuskan kalimat, menata langkah, dan mengingatkan asumsi yang terlewat, sedangkan kesahihan logis setiap langkah tidak dijamin oleh mekanisme prediksi bahasa (Bender et al., 2021; Ji et al., 2023). Ketimpangan tersebut tampak jelas pada temuan Agad et al. (2025) bahwa umpan balik ChatGPT lebih banyak memengaruhi cara penyajian

ISSN: 2088-0294 | e-ISSN: 2621-9166
https://doi.org/10.37630/jpm.v16i5.5237

bukti daripada penalarannya. Manfaat GenAI karenanya paling besar ketika peran GenAI dirancang secara sengaja, seperti pada studi Davor (2026), Zhao et al. (2025), dan Zhuang et al. (2026), bukan ketika GenAI digunakan secara bebas tanpa arahan.

Kapabilitas GenAI dalam pembuktian juga terus meningkat. GPT-4 menyelesaikan rata-rata 73% soal Proof Blocks tanpa pengecoh, jauh melampaui GPT-3.5 yang hanya sekitar 11% (Poulsen et al., 2024), sedangkan kombinasi LLM dengan pembukti teorema Lean memungkinkan umpan balik yang kebenarannya dapat diverifikasi secara formal (Patel et al., 2026). Ditinjau dari fungsi bukti (de Villiers, 1990), peluang yang ditemukan terutama mendukung fungsi komunikasi dan penemuan, sedangkan fungsi eksplanasi terdukung ketika GenAI diposisikan sebagai *scaffold* yang menuntut justifikasi.

Keterbatasan yang paling sering dilaporkan adalah ketidakandalan logika. ChatGPT melakukan kesalahan saat membuktikan ketaksamaan dan tidak mampu memperbaikinya meskipun telah ditunjukkan (Drushlyak et al., 2024), menghasilkan argumen yang saling bertentangan dengan kualitas yang dipengaruhi bahasa *prompt* (Yoon et al., 2024), serta memberikan respons yang tidak stabil terkait hubungan kontinuitas dan turunan (Urhan et al., 2024). Sebagian besar responsnya kepada mahasiswa juga tidak memenuhi syarat rasionalitas epistemik (Urhan et al., 2026). Kartika dan Clark (2026) menemukan bahwa GenAI cenderung bertumpu pada empirisme naif, sedangkan sebagian mahasiswa justru menunjukkan penalaran deduktif yang lebih fleksibel. Temuan tersebut selaras dengan catatan Almarashdi et al. (2024) tentang lemahnya ChatGPT pada soal turunan dan geometri ruang.

Kinerja GenAI juga menurun pada bukti yang kompleks dan formal. Tingkat keberhasilan GPT-4 turun dari 73% menjadi 44% pada soal Proof Blocks yang memuat pengecoh (Poulsen et al., 2024), sebagian besar kegagalan GPT-4 sebagai tutor terjadi pada tugas bukti formal seperti reduksi (Frable et al., 2026), dan petunjuk yang dihasilkan LLM lemah dalam menjelaskan alasan suatu langkah (Tithi et al., 2025). Penerjemahan bukti siswa ke bahasa formal pun masih menjadi tantangan teknis (Patel et al., 2026).

Umpan balik GenAI juga belum memenuhi kebutuhan pembelajaran. ChatGPT menilai bukti baris demi baris, bukan sebagai alur logis yang utuh, memberikan umpan balik yang kontradiktif, dan tidak selaras dengan ekspektasi dosen sehingga kepercayaan mahasiswa rendah (Agad et al., 2025). Penilaian ChatGPT-4 terhadap tugas pembuktian juga sensitif terhadap format jawaban dan menghasilkan nilai yang bervariasi (Dykstra et al., 2026). Temuan tersebut memberikan gambaran yang lebih berhati-hati daripada kesimpulan Dela Torre dan Sayco (2026) tentang umpan balik GenAI yang tepat waktu dalam pendidikan matematika secara umum, sebab umpan balik yang andal baru tampak pada sistem yang dilatih secara khusus, seperti pada studi Zhao et al. (2025).

Risiko berikutnya adalah ketergantungan. Mahasiswa yang menggunakan GenAI sebelum menyusun draf sendiri cenderung menjadikan bukti GenAI sebagai acuan utama (Yoon et al., 2026), sedangkan hanya mahasiswa berkemampuan tinggi yang mampu menyaring kesalahan ChatGPT (Urhan et al., 2026). Temuan berbasis sampel kecil tersebut mengisyaratkan bahwa siswa yang paling membutuhkan bantuan justru paling rentan terhadap kesalahan GenAI. Efektivitas GenAI juga dibatasi oleh rendahnya literasi AI, sebagaimana tampak pada calon guru yang memiliki pengetahuan dangkal dan banyak miskonsepsi tentang LLM (Dilling & Herrmann, 2024), serta oleh pertimbangan etis yang membuat sebagian mahasiswa enggan menggunakan GenAI (Yoon et al., 2024).

Ditinjau dari fungsi bukti, keterbatasan tersebut paling mengancam fungsi verifikasi. Bukti yang runtut secara bahasa tetapi keliru secara logis lebih sulit dikenali sehingga siswa dengan skema bukti eksternal (Harel & Sowder, 1998) berisiko menerima bukti GenAI hanya karena tampilannya meyakinkan. Kecenderungan GenAI pada empirisme naif (Balacheff, 1988) juga berpotensi memperkuat skema bukti empiris yang selama ini menjadi hambatan siswa.

Penafsiran temuan juga perlu mempertimbangkan keterbatasan metodologis kumpulan studi. Penilaian MMAT (Tabel 3) menunjukkan bahwa sembilan dari 16 studi hanya memenuhi dua kriteria atau kurang, sehingga temuan dari studi tersebut perlu dibaca sebagai indikasi awal. Dari enam studi konteks ilmu komputer, tiga tema pada Tabel 2 (ditandai †) bersumber seluruhnya dari studi tersebut, yaitu umpan balik skala besar, kinerja yang menurun pada bukti formal, dan GenAI yang keluarannya dapat diverifikasi. Ketiga tema itu sebaiknya dibaca sebagai temuan dari konteks bukti formal dan belum tentu berlaku untuk pembuktian geometri atau aljabar di kelas matematika. Sebagian besar studi menggunakan desain kualitatif dengan partisipan yang

ISSN: 2088-0294 | e-ISSN: 2621-9166
https://doi.org/10.37630/jpm.v16i5.5237

sedikit, hanya satu studi yang menggunakan kelompok kontrol, dan studi tersebut tidak menyebutkan alat GenAI yang digunakan. Lima studi memakai ChatGPT 3.5 sehingga sebagian keterbatasan yang dilaporkan mungkin tidak lagi sepenuhnya berlaku bagi model generasi terbaru, mengingat besarnya perbedaan kinerja antarversi model pada studi Poulsen et al. (2024). Lima studi juga tidak melibatkan partisipan manusia secara langsung, dan belum ada studi yang mengkaji dampak penggunaan GenAI terhadap kemampuan pembuktian dalam jangka panjang.

Tabel 3. Penilaian kualitas studi dengan MMAT 2018

Kode	Kategori MMAT	Skor	Catatan utama penilaian
S1	Kuantitatif non-randomized	1/5	Satu kelompok pra-pasca tanpa pembanding; rubrik buatan penulis tanpa data validitas; kepatuhan pelaksanaan tidak dilaporkan
S2	Kualitatif	4/5	Klaim hasil tidak didukung analisis kualitas bukti mahasiswa
S3	Kualitatif	4/5	Kerangka SIPE-AI dikembangkan dan diuji pada data yang sama; partisipan sangat sedikit
S4	Kuantitatif deskriptif ‡	2/5	Representativitas kumpulan soal, penanganan keluaran gagal, dan analisis per topik belum dijelaskan
S5	Kualitatif	5/5	Studi kasus satu mahasiswa dengan think-aloud dan analisis Toulmin–Habermas
S6	Kualitatif	3/5	Jumlah peserta tidak konsisten antara abstrak dan metode (6 dan 7); metode analisis tematik tidak dijelaskan
S7	Mixed methods	3/5	Komponen kualitatif tanpa prosedur analisis yang dijelaskan; evaluasi inti tanpa partisipan manusia
S8	Kuantitatif deskriptif ‡	2/5	Representativitas sampel, penanganan keluaran yang gagal, dan uji statistik belum dilaporkan lengkap
S9	Kualitatif	2/5	Rincian pengumpulan data GenAI tidak dilaporkan; contoh pekerjaan mahasiswa memuat kesalahan hitung
S10	Kuantitatif deskriptif ‡	1/5	Sistem tanpa partisipan mahasiswa; 21 dari 44 bukti dipilih; metrik belum divalidasi
S11	Kuantitatif non-randomized	2/5	Sampel kenyamanan dari satu sekolah; kovariat hanya pra-tes; data fidelitas tidak tersedia; ukuran sampel perlu diklarifikasi
S12	Kualitatif	5/5	Semua kriteria terpenuhi; catatan kecil pada label tabel
S13	Kualitatif	5/5	Studi kasus dua mahasiswa; pengkodean tanpa koder kedua dan tanpa uji keandalan
S14	Mixed methods	1/5	Alasan desain campuran tidak dijelaskan; integrasi hanya naratif; n = 10 tanpa uji keandalan
S15	Kuantitatif deskriptif ‡	1/5	Sampel tidak representatif; rubrik tidak konsisten; ambang klasifikasi tidak dibenarkan
S16	Kuantitatif deskriptif ‡	1/5	Pihak penilai tidak konsisten antara abstrak dan metode; koreksi perbandingan ganda tidak dilaporkan

Keterangan: ‡ = studi tanpa partisipan manusia; MMAT hanya indikatif untuk kategori ini. Penilaian disusun dari teks lengkap dengan bantuan AI dan diperiksa ulang oleh penulis.

Peluang dan keterbatasan tersebut menghasilkan enam implikasi pedagogis. Pertama, waktu penggunaan GenAI perlu diatur dengan mendahulukan upaya pembuktian mandiri karena mahasiswa menilai urutan tersebut lebih bermanfaat bagi pemahaman konseptual (Yoon et al., 2026), sejalan dengan gagasan *productive failure* (Kapur, 2008). Kedua, GenAI perlu diposisikan sebagai *scaffold* yang mendorong justifikasi, bukan sebagai penyedia bukti lengkap, sebagaimana rancangan pada studi Davor (2026) dan tutor sokratik pada studi Frable et al. (2026).

Ketiga, validasi bukti GenAI dapat dijadikan aktivitas belajar, misalnya dengan meminta siswa menganalisis bukti yang keliru (Drushlyak et al., 2024) atau membandingkan argumen empiris GenAI dengan argumen deduktif (Kartika & Clark, 2026). Aktivitas tersebut perlu disertai bimbingan, misalnya daftar periksa kesahihan bukti, karena siswa berkemampuan rendah kesulitan menyaring kesalahan (Urhan et al., 2026). Keempat, literasi AI dan keterampilan *prompting* perlu dikembangkan, terutama bagi calon guru (Dilling & Herrmann, 2024; Zhuang et al., 2026), disertai kebijakan yang jelas tentang batas penggunaan GenAI yang dapat diterima. Temuan bahwa kualitas argumen ChatGPT dipengaruhi bahasa *prompt* (Yoon et al., 2024) juga patut diperhatikan oleh pendidik di Indonesia karena siswa kemungkinan besar berinteraksi dengan GenAI menggunakan bahasa Indonesia.

Kelima, guru dan dosen tetap memegang peran sentral sebagai mediator yang menyelaraskan keluaran GenAI dengan standar kesahihan bukti (Agad et al., 2025; Davor, 2026), sedangkan penilaian otomatis lebih tepat difungsikan sebagai umpan balik formatif daripada penentu nilai akhir (Dykstra et al., 2026; Zhao et al., 2025). Keenam, pengembangan GenAI yang keluarannya dapat diverifikasi, misalnya melalui kombinasi LLM dan pembukti teorema, perlu didorong agar GenAI dapat mendukung fungsi verifikasi (Patel et al., 2026; Tithi

ISSN: 2088-0294 | e-ISSN: 2621-9166
https://doi.org/10.37630/jpm.v16i5.5237

et al., 2025). Keenam implikasi tersebut bertumpu pada satu prinsip: GenAI sebaiknya diposisikan sebagai mitra belajar yang keluarannya selalu perlu diverifikasi, bukan sebagai otoritas yang menentukan kebenaran bukti.

Seiring meningkatnya kapabilitas GenAI, pertanyaan yang lebih penting bergeser dari kemampuan GenAI dalam membuktikan menjadi cara siswa terlibat dengan bukti yang dihasilkannya. Pembelajaran pembuktian karenanya perlu memberi porsi yang lebih besar pada kemampuan memahami dan mengevaluasi bukti, yang telah lama dipandang sebagai komponen penting pembelajaran pembuktian (Mejía-Ramos et al., 2012). Agenda penelitian selanjutnya meliputi studi pada jenjang SMP dan SMA, studi di negara berkembang termasuk Indonesia dengan bahasa *prompt* selain bahasa Inggris, desain eksperimental dan longitudinal, pengujian model GenAI terbaru, serta studi tentang praktik guru dalam jabatan.

Perbandingan dengan studi di luar sampel. Beberapa publikasi yang tidak memenuhi kriteria inklusi melaporkan pola yang sejalan dengan temuan ini. ChatGPT 3.5 melakukan penalaran melingkar ketika membuktikan kesamaan besar sudut bertolak belakang (McGalliard & Otten, 2025), ChatGPT gagal membuktikan salah satu teorema geometri yang berhasil dibuktikan GeoGebra Discovery (Botana et al., 2024), dan GPT-4 kesulitan pada bukti yang menuntut lema antara serta perencanaan yang panjang (Collins et al., 2024). Kawski (2025) menilai bahwa kekeliruan logis LLM dapat meningkatkan keterlibatan mahasiswa di kelas berbasis inkuiri. Karena publikasi tersebut berada di luar sampel, temuannya tidak dimasukkan dalam sintesis dan hanya digunakan sebagai pembanding.

Kajian ini memiliki sejumlah keterbatasan. Pencarian hanya dilakukan pada Scopus dan publikasi berbahasa Inggris, satu laporan tidak dapat diperoleh, dan *search string* tidak memuat istilah seperti *justification* dan *deductive reasoning*. Perluasan lingkup melalui dua aturan menambah keragaman konteks studi, tetapi membuat studi yang dihimpun heterogen dari sisi konteks dan tingkat formalitas bukti sehingga temuan perlu ditafsirkan dengan hati-hati. Penilaian MMAT bersifat indikatif karena dilakukan dengan bantuan AI dan perlu diverifikasi lebih lanjut, serta sebagian temuan tentang keterbatasan model berpotensi cepat usang seiring perkembangan GenAI.

SIMPULAN

Kajian terhadap 16 studi yang terbit pada 2024–2026 menunjukkan bahwa penelitian tentang GenAI dalam pembelajaran pembuktian matematis masih didominasi studi kualitatif berskala kecil di perguruan tinggi, terutama di Amerika Serikat, dengan ChatGPT sebagai alat utama dan kualitas metodologis yang umumnya terbatas, sedangkan jenjang sekolah dan konteks Indonesia hampir belum tersentuh. GenAI berpeluang menjadi sumber gagasan pembuktian, pemberi umpan balik dalam skala besar, *scaffold* penalaran, sarana latihan berpikir kritis, pendukung komunikasi bukti, dan simulasi pedagogis bagi calon guru, terutama untuk fungsi komunikasi dan penemuan bukti. Peluang tersebut dibatasi oleh ketidakandalan logika, menurunnya kinerja pada bukti formal, umpan balik yang dangkal, risiko ketergantungan, dan rendahnya literasi AI, yang secara bersama paling mengancam fungsi verifikasi bukti. Guru dan dosen disarankan mendahulukan kerja mandiri sebelum penggunaan GenAI, memosisikan GenAI sebagai *scaffold*, menjadikan validasi bukti GenAI sebagai aktivitas belajar, mengembangkan literasi AI, mempertahankan peran sebagai mediator, dan mendorong pemanfaatan GenAI yang keluarannya dapat diverifikasi. Penelitian selanjutnya perlu diarahkan pada jenjang SMP dan SMA, konteks negara berkembang termasuk Indonesia, serta desain eksperimental dan longitudinal dengan model GenAI terbaru.

Daftar Pustaka

Agad, A., Rocena, L. A., Cabalo, J. M. S., & Victorio, C. N. (2025). Revising with AI: Undergraduate students' perceptions and trust in AI-generated feedback on geometric proofs. *Proceedings of the 30th Asian Technology Conference in Mathematics (ATCM 2025)*.

Almarashdi, H. S., Jarrah, A. M., Abu Khurma, O., & Gningue, S. M. (2024). Unveiling the potential: A systematic review of ChatGPT in transforming mathematics teaching and learning. *Eurasia Journal of Mathematics, Science and Technology Education*, *20*(12), em2555. https://doi.org/10.29333/ejmste/15739

Balacheff, N. (1988). Aspects of proof in pupils' practice of school mathematics. In D. Pimm (Ed.), *Mathematics, teachers and children* (pp. 216–235). Hodder & Stoughton.

ISSN: 2088-0294 | e-ISSN: 2621-9166
https://doi.org/10.37630/jpm.v16i5.5237

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Botana, F., Recio, T., & Vélez, M. P. (2024). On using GeoGebra and ChatGPT for geometric discovery. *Computers*, *13*(8), 187. https://doi.org/10.3390/computers13080187

Collins, K. M., Jiang, A. Q., Frieder, S., Wong, L., Zilka, M., Bhatt, U., Lukasiewicz, T., Wu, Y., Tenenbaum, J. B., Hart, W., Gowers, T., Li, W., Weller, A., & Jamnik, M. (2024). Evaluating language models for mathematics through interactions. *Proceedings of the National Academy of Sciences*, *121*(24), e2318124121. https://doi.org/10.1073/pnas.2318124121

Davor, I. (2026). Generative AI-supported instruction as a cognitive scaffold: Effects on senior high school students' geometric reasoning and proof construction. *Frontiers in Education*, *11*, 1869961. https://doi.org/10.3389/feduc.2026.1869961

de Villiers, M. (1990). The role and function of proof in mathematics. *Pythagoras*, *24*, 17–24.

Dela Torre, J., & Sayco, J. (2026). Generative artificial intelligence in mathematics education: A systematic literature review using the PRISMA framework. *The Asian Conference on Education & International Development 2026 Official Conference Proceedings* (pp. 591–606). IAFOR. https://doi.org/10.22492/issn.2189-101X.2026.48

Dilling, F., & Herrmann, M. (2024). Using large language models to support pre-service teachers mathematical reasoning—An exploratory study on ChatGPT as an instrument for creating mathematical proofs in geometry. *Frontiers in Artificial Intelligence*, *7*, 1460337. https://doi.org/10.3389/frai.2024.1460337

Drushlyak, M., Lukashova, T., Sabadosh, Y., Melnikov, I., & Semenikhina, O. (2024). Using ChatGPT for the development of critical thinking in youth: Example of inequality proof. *2024 47th ICT and Electronics Convention (MIPRO 2024) - Proceedings*, 334–339. https://doi.org/10.1109/MIPRO60963.2024.10569759

Dykstra, N., Wesley, R., & Dougherty, R. E. (2026). On the efficacy of using large language models for automatic grading of CS theory problems. *Proceedings of the 57th ACM Technical Symposium on Computer Science Education V.2 (SIGCSE TS 2026)*, 1311–1312. https://doi.org/10.1145/3770761.3777261

Frable, W., Trevino, A., & Dougherty, R. E. (2026). Viability of large language models as CS theory tutors. *Proceedings of the 57th ACM Technical Symposium on Computer Science Education V.2 (SIGCSE TS 2026)*, 1335–1336. https://doi.org/10.1145/3770761.3777358

Frieder, S., Pinchetti, L., Chevalier, A., Griffiths, R.-R., Salvatori, T., Lukasiewicz, T., Petersen, P. C., & Berner, J. (2023). Mathematical capabilities of ChatGPT. *Advances in Neural Information Processing Systems*, *36*. https://papers.neurips.cc/paper_files/paper/2023/hash/58168e8a92994655d6da3939e7cc0918-Abstract-Datasets_and_Benchmarks.html

Hanna, G. (2000). Proof, explanation and exploration: An overview. *Educational Studies in Mathematics*, *44*(1–2), 5–23. https://doi.org/10.1023/A:1012737223465

Harel, G., & Sowder, L. (1998). Students' proof schemes: Results from exploratory studies. In A. H. Schoenfeld, J. Kaput, & E. Dubinsky (Eds.), *Research in collegiate mathematics education III* (pp. 234–283). American Mathematical Society.

Hong, Q. N., Fàbregues, S., Bartlett, G., Boardman, F., Cargo, M., Dagenais, P., Gagnon, M.-P., Griffiths, F., Nicolau, B., O'Cathain, A., Rousseau, M.-C., Vedel, I., & Pluye, P. (2018). The Mixed Methods Appraisal Tool (MMAT) version 2018 for information professionals and researchers. *Education for Information*, *34*(4), 285–291. https://doi.org/10.3233/EFI-180221

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, *55*(12), 1–38. https://doi.org/10.1145/3571730

Kapur, M. (2008). Productive failure. *Cognition and Instruction*, *26*(3), 379–424.

ISSN: 2088-0294 | e-ISSN: 2621-9166
https://doi.org/10.37630/jpm.v16i5.5237

https://doi.org/10.1080/07370000802212669

Kartika, H., & Clark, L. J. (2026). Comparative analysis of mathematical arguments in algebraic proofs: Generative AI vs. university students approaches. *Mathematics Teaching-Research Journal*, *18*(1), 281–299.

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., … Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, *103*, 102274. https://doi.org/10.1016/j.lindif.2023.102274

Kawski, M. (2025). ChatGPT and AI enhancing undergrad math. *Proceedings of the 30th Asian Technology Conference in Mathematics (ATCM 2025)*.

McGalliard, W., & Otten, S. (2025). AI responses to challenging problems and educator responses to AI availability. *Digital Experiences in Mathematics Education*, *11*(2), 319–332. https://doi.org/10.1007/s40751-024-00167-4

Mejía-Ramos, J. P., Fuller, E., Weber, K., Rhoads, K., & Samkoff, A. (2012). An assessment model for proof comprehension in undergraduate mathematics. *Educational Studies in Mathematics*, *79*(1), 3–18. https://doi.org/10.1007/s10649-011-9349-7

Muhamad Sani, N. S. A., Omar, M. A. A., & Abdullah, M. R. (2026). Exploring the role of generative artificial intelligence in transforming mathematics teaching and learning: A systematic mapping review. *Innovative Teaching and Learning Journal*, *10*(2), 155–174. https://doi.org/10.11113/itlj.v10.256

Nguyen, D. T., & Pham, Q. V. (2025). The evolving landscape of AI integration in mathematics education: A systematic review of trends (2015–2025). *Eurasia Journal of Mathematics, Science and Technology Education*, *21*(10), em2714. https://doi.org/10.29333/ejmste/17078

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*, *372*, n71. https://doi.org/10.1136/bmj.n71

Patel, M., Bhattacharyya, R., Lu, T., Mehta, A., Voss, N., Norouzi, N., & Ranade, G. (2026). LeanTutor: Towards a verified AI mathematical proof tutor. *Proceedings of the AAAI Conference on Artificial Intelligence*, *40*(47), 40670–40678. https://doi.org/10.1609/aaai.v40i47.41514

Pedemonte, B. (2007). How can the relationship between argumentation and proof be analysed? *Educational Studies in Mathematics*, *66*(1), 23–41. https://doi.org/10.1007/s10649-006-9057-x

Poulsen, S., Sarsa, S., Prather, J., Leinonen, J., Becker, B. A., Hellas, A., Denny, P., & Reeves, B. N. (2024). Solving proof block problems using large language models. *Proceedings of the 55th ACM Technical Symposium on Computer Science Education (SIGCSE 2024)*, *1*, 1063–1069. https://doi.org/10.1145/3626252.3630928

Stylianides, A. J. (2007). Proof and proving in school mathematics. *Journal for Research in Mathematics Education*, *38*(3), 289–321. https://doi.org/10.2307/30034869

Thomas, J., & Harden, A. (2008). Methods for the thematic synthesis of qualitative research in systematic reviews. *BMC Medical Research Methodology*, *8*, 45. https://doi.org/10.1186/1471-2288-8-45

Tithi, S. D., Ramesh, A. K., DiMarco, C., Tian, X., Alam, N., Fazeli, K., & Barnes, T. (2025). The promise and limits of LLMs in constructing proofs and hints for logic problems in intelligent tutoring systems. *Computers and Education: Artificial Intelligence*, *9*, 100490. https://doi.org/10.1016/j.caeai.2025.100490

Urhan, S., Gençaslan, O., & Dost, Ş. (2024). An argumentation experience regarding concepts of calculus with ChatGPT. *Interactive Learning Environments*, *32*(10), 7186–7211. https://doi.org/10.1080/10494820.2024.2308093

Urhan, S., Gençaslan, O., & Dost, Ş. (2026). Is ChatGPT a rational assistant for university students during mathematical reasoning? *Science & Education*. Advance online publication.

ISSN: 2088-0294 | e-ISSN: 2621-9166
https://doi.org/10.37630/jpm.v16i5.5237

https://doi.org/10.1007/s11191-026-00725-9

Yoon, H., Hwang, J., Lee, K., Roh, K. H., & Kwon, O. N. (2024). Students' use of generative artificial intelligence for proving mathematical statements. *ZDM – Mathematics Education*, *56*(7), 1531–1551. https://doi.org/10.1007/s11858-024-01629-0

Yoon, H., Lee, Y., Lee, K., & Kwon, O. N. (2026). Before or after? Beyond timing of students' interaction with generative artificial intelligence in proving mathematical statements. *ZDM – Mathematics Education*, *58*(5), 945–962. https://doi.org/10.1007/s11858-026-01800-9

Zhao, C., Silva, M., & Poulsen, S. (2025). Autograding mathematical induction proofs with natural language processing. *International Journal of Artificial Intelligence in Education*, *35*(5), 3202–3232. https://doi.org/10.1007/s40593-025-00498-2

Zhuang, Y., Yao, X., Davis, T. J., & Nudze, W. Y. (2026). Utilizing customized ChatGPT-based chatbots to support preservice teachers' understanding of collective mathematical argumentation. *European Journal of STEM Education*, *11*(1), Article 39. https://doi.org/10.20897/ejsteme/18918