**TECHNOVATAR Jurnal Teknologi, Industri, dan Informasi**
Vol. 4, No. 4, Oktober 2026, pp. 1 – 19
**Open Access**:
https://journal.awatarapublisher.com/index.php/technovatar
**ISSN: 3031-8874, DOI: 10.61434/technovatar.v4i4.401**

# PEMANFAATAN LARGE LANGUAGE MODEL DALAM PENDIDIKAN PEMROGRAMAN: SYSTEMATIC LITERATURE REVIEW BERBASIS PRISMA 2020

Bambang Irawan[1*]
1 Program Studi Teknik Informatika, Fakultas Teknik dan Ilmu Komputer, Universitas Muhadi Setiabudi, Brebes, Jawa Tengah, Indonesia
Email: bambangumus@gmail.com

**INFO ARTIKEL**
Sejarah Artikel:
Diterima, 21 September 2026
Direvisi, 28 September 2026
Disetujui, 1 Oktober 2026

**KEYWORDS**
Large language model
Programming education
Systematic literature review
PRISMA
Academic integrity

***ABSTRACT***

*The use of large language models (LLMs) in programming education is growing rapidly, yet evidence on its benefits, risks, and best practices has not been mapped in an integrated way. This study aimed to map the characteristics of empirical studies, synthesize benefits and risks including plagiarism and overreliance, and identify evidence-supported instructional practices. This systematic literature review followed PRISMA 2020. Searches were run through Consensus on 9 October 2026 using 14 queries, yielding 140 records. Screening, extraction, and six-item quality assessment were performed by a single AI-assisted reviewer based on abstracts. A total of 68 studies published between 2023 and 2026 were included and synthesized thematically. The most frequently reported benefit was on-demand feedback and debugging help (30 studies), followed by motivation and self-efficacy (9). Of 21 studies reporting learning outcomes, 10 were positive, 5 mixed, 5 null, and 1 negative. Main risks were overreliance (22 studies), unreliable or solution-leaking output (20), and academic integrity concerns (9). Promising practices include pedagogical scaffolding, prompt literacy training, self-regulation support, process-based assessment, human oversight, and explicit AI-use policies. Findings are preliminary because the search was limited to one aggregator and screening relied on abstracts.*

*This work is licensed under a Creative Commons Attribution 4.0 International License.*

**KATA KUNCI**
Large language model
Pendidikan pemrograman
Systematic literature review
PRISMA
Integritas akademik

**ABSTRAK**

Pemanfaatan large language model (LLM) dalam pembelajaran pemrograman berkembang cepat, tetapi bukti tentang manfaat, risiko, dan praktik terbaiknya belum terpetakan secara terpadu. Penelitian ini bertujuan memetakan karakteristik studi empiris, merangkum manfaat dan risiko, termasuk plagiarisme dan ketergantungan, serta mengidentifikasi praktik pembelajaran yang didukung bukti. Tinjauan literatur sistematis ini dilaporkan menurut PRISMA 2020. Penelusuran dilakukan melalui Consensus pada 9 Oktober 2026 dengan 14 kueri, menghasilkan 140 rekaman. Seleksi, ekstraksi, dan penilaian kualitas enam butir dilakukan oleh satu penelaah dengan bantuan asisten AI berdasarkan abstrak. Sebanyak 68 studi terbit 2023 sampai 2026 disertakan dan disintesis secara naratif tematik. Manfaat yang paling sering dilaporkan adalah umpan balik dan bantuan debugging sesuai permintaan (30 studi), disusul motivasi dan efikasi diri (9). Dari 21 studi dengan ukuran hasil belajar, 10 positif, 5 campuran, 5 nol, dan 1 negatif. Risiko utama adalah ketergantungan berlebih (22 studi), keluaran tidak andal atau membocorkan solusi (20), dan integritas akademik (9). Praktik yang menjanjikan meliputi perancah pedagogis, pelatihan literasi prompt, dukungan regulasi diri, asesmen berbasis proses, pengawasan manusia, dan kebijakan penggunaan AI yang eksplisit. Temuan bersifat awal karena penelusuran terbatas pada satu agregator dan penyaringan berbasis abstrak.

**CORRESPONDING AUTHOR**
Bambang Irawan
Program Studi Teknik Informatika,
Universitas Muhadi Setiabudi
Jl. Pangeran Diponegoro KM 2,
Pesantunan, Wanasari, Brebes 52252, Jawa Tengah, Indonesia
Email: bambangumus@gmail.com

## PENDAHULUAN

Pemrograman merupakan kompetensi inti bagi mahasiswa informatika dan semakin penting bagi bidang lain, tetapi bagi pemula kegiatan ini menuntut penguasaan sintaks, penalaran logis, dan strategi pemecahan masalah sekaligus. Sejak 2022, *large language model* (LLM) seperti ChatGPT dan GitHub Copilot dapat menghasilkan kode, menjelaskan pesan galat, dan memberi petunjuk langkah berikutnya dari deskripsi bahasa alami. Kemampuan tersebut telah diuji dalam berbagai bentuk, mulai dari penjelasan galat kompilasi yang dipasang pada kompilator untuk ribuan mahasiswa (Taylor et al., 2023), umpan balik atas pekerjaan mahasiswa (Azaiz et al., 2024), hingga asisten pengajar digital dengan pagar pengaman (Liffiton et al., 2023; Kazemitabaar et al., 2024).

Bukti mengenai dampak LLM pada pembelajaran pemrograman belum seragam. Sebagian eksperimen melaporkan peningkatan berpikir komputasional, efikasi diri, dan motivasi (Yılmaz et al., 2023b), sedangkan eksperimen lain tidak menemukan perbedaan kinerja yang signifikan (Kosar et al., 2024; Xue et al., 2024) atau bahkan melaporkan capaian belajar yang lebih rendah dibanding metode konvensional (Yang et al., 2025). Di sisi lain, kemudahan menghasilkan solusi lengkap memunculkan kekhawatiran integritas akademik, sehingga muncul penelitian tentang deteksi kode buatan LLM (Hoq et al., 2024; Ramachandra et al., 2026) dan survei tentang kecurangan mahasiswa (de Souza Santos et al., 2026). Kekhawatiran lain adalah ketergantungan berlebih, yaitu mahasiswa yang langsung meminta solusi lengkap daripada memahami kesalahannya (Rahe et al., 2025).

Beberapa tinjauan sistematis telah merangkum literatur ini. Tinjauan yang teridentifikasi dalam penelusuran ini mencakup tinjauan LLM dalam pendidikan ilmu komputer secara umum (Raihan et al., 2024), tinjauan khusus pada proses penilaian pembelajaran pemrograman (Pereira et al., 2025), tinjauan dengan kerangka enam dimensi atas 75 makalah (Fu et al., 2025), tinjauan atas 42 studi yang terbit sampai Maret 2024 (Zhu et al., 2025), pemetaan literatur tentang deteksi plagiarisme kode buatan AI (Simmons et al., 2024), tinjauan naratif atas 29 studi (Farooq et al., 2026), dan tinjauan lingkup atas 146 studi pendidikan teknik dan komputasi (Álvarez Ariza et al., 2025). Tinjauan-tinjauan tersebut umumnya menyoroti satu aspek tertentu, mencakup periode yang lebih awal, atau tidak menilai kualitas studi primer secara terstruktur. Pertumbuhan publikasi yang cepat sepanjang 2025 sampai 2026 juga menuntut pembaruan sintesis, terutama untuk menimbang manfaat, risiko, dan praktik pembelajaran dalam satu kerangka yang dapat dipakai pendidik.

Berdasarkan celah tersebut, kebaruan penelitian ini terletak pada sintesis integratif yang memetakan karakteristik bukti empiris, menyeimbangkan manfaat umpan balik dengan risiko kognitif (ketergantungan dan alih daya pemikiran), serta merumuskan rekomendasi pedagogis berbasis bukti melalui PRISMA 2020 disertai penilaian kualitas studi primer. Tujuan penelitian adalah memetakan karakteristik studi, merangkum manfaat, merangkum risiko termasuk plagiarisme dan ketergantungan, dan mengidentifikasi praktik terbaik penggunaan LLM, melalui tinjauan literatur sistematis yang dilaporkan menurut PRISMA 2020 (Page et al., 2021) dengan penilaian kualitas studi. Kontribusi penelitian ini adalah peta bukti terkini beserta penilaian kualitasnya, serta implikasi praktis bagi pendidik pemrograman di perguruan tinggi, termasuk di Indonesia.

## METODE PENELITIAN

### Protokol dan registrasi

Penelitian ini merupakan tinjauan literatur sistematis yang dilaporkan mengikuti pernyataan PRISMA 2020 (Page et al., 2021) dan mengacu pada panduan tinjauan sistematis bidang rekayasa perangkat lunak (Kitchenham dan Charters, 2007). Protokol tidak didaftarkan pada PROSPERO (karena registri ini berfokus pada luaran kesehatan manusia) maupun repositori lain, namun seluruh tahapan dirancang secara sistematis merujuk panduan Kitchenham dan Charters (2007) serta pernyataan PRISMA 2020. Pertanyaan penelitian, kerangka PICo, kriteria seleksi, dan butir

penilaian kualitas dirumuskan sebelum penyaringan. Dua aturan operasional, yaitu penegasan kriteria data pembelajar atau data otentik pada tahap kelayakan dan ambang skor kualitas, difinalkan selama proses dan sebelum sintesis dilakukan.

**Pertanyaan penelitian dan kerangka PICo**

Pertanyaan penelitian dirumuskan dengan kerangka PICo. *Population* adalah pembelajar dan pengajar pada pembelajaran pemrograman atau ilmu komputer. *Interest* adalah pemanfaatan LLM, yaitu ChatGPT, GitHub Copilot, model GPT, Gemini, serta alat yang dibangun di atas LLM. *Context* adalah pembelajaran formal maupun daring pada jenjang sekolah dan perguruan tinggi. Empat pertanyaan penelitian disajikan pada Tabel 1.

**Tabel 1 Rumusan pertanyaan penelitian**

| Kode | Pertanyaan penelitian |
|---|---|
| RQ1 | Bagaimana karakteristik studi empiris (tahun, desain, jenjang, jenis LLM, dan kualitas) tentang pemanfaatan LLM dalam pembelajaran pemrograman? |
| RQ2 | Manfaat apa yang dilaporkan dari pemanfaatan LLM dalam pembelajaran pemrograman? |
| RQ3 | Risiko apa yang dilaporkan, termasuk plagiarisme, ketergantungan berlebih, dan keandalan keluaran? |
| RQ4 | Praktik pembelajaran apa yang direkomendasikan atau diuji untuk memaksimalkan manfaat dan menekan risiko? |

**Kriteria inklusi dan eksklusi**

Kriteria seleksi disajikan pada Tabel 2. Studi disertakan bila memuat populasi pembelajar atau pengajar pemrograman, melibatkan LLM sebagai fokus intervensi atau objek evaluasi, berupa artikel jurnal atau makalah prosiding berbahasa Inggris atau Indonesia yang terbit 2021 sampai 2026, dan melaporkan data empiris. Studi dieksklusi bila tidak berfokus pada pemrograman, merupakan studi sekunder, berstatus praterbit atau berasal dari venue yang tidak dapat diverifikasi, bukan studi empiris, hanya menguji kapabilitas model tanpa data pembelajar atau data otentik, atau berskor kualitas di bawah ambang.

**Tabel 2 Kriteria inklusi dan eksklusi**

| Kode | Uraian kriteria |
|---|---|
| IC1 | Populasi pembelajar (siswa atau mahasiswa) atau pengajar pada pembelajaran pemrograman atau ilmu komputer |
| IC2 | LLM atau alat berbasis LLM menjadi fokus intervensi atau evaluasi untuk belajar, umpan balik, penilaian, atau pembuatan materi pemrograman |
| IC3 | Artikel jurnal atau makalah prosiding, terbit 2021 sampai 2026, bahasa Inggris atau Indonesia |
| IC4 | Studi empiris (eksperimen, kuasi-eksperimen, survei, wawancara, studi kasus, analisis log, evaluasi alat) dengan data pembelajar atau data otentik |
| IC5 | Melaporkan hasil terkait manfaat, risiko, atau praktik pembelajaran |
| EC1 | Fokus bukan pembelajaran pemrograman (pendidikan umum, bahasa, atau bidang lain) |
| EC2 | Studi sekunder (tinjauan, pemetaan, atau survei literatur) |
| EC3 | Praterbit atau venue tidak dapat diverifikasi |
| EC4 | Artikel opini, komentar, atau konseptual tanpa data empiris |
| EC5 | Hanya uji kapabilitas model tanpa data pembelajar atau data otentik, atau sumber data tidak dinyatakan |
| EC6 | Skor penilaian kualitas kurang dari 3,5 dari 6 |

**Sumber informasi dan strategi pencarian**

Penelusuran dilakukan melalui Consensus, mesin pencari semantik atas lebih dari 400 juta makalah yang bersumber dari Semantic Scholar, OpenAlex, dan PubMed, pada 9 Oktober 2026. Basis data Scopus, IEEE Xplore, dan ACM Digital Library tidak ditelusuri langsung; Consensus dipakai sebagai pintu masuk tunggal dan hasilnya disaring pada tingkat venue. Empat belas kueri dengan sudut pandang berbeda dijalankan untuk mencakup ragam terminologi pendidikan dan

teknologi, dengan batas sepuluh rekaman teratas per kueri pada Consensus untuk menyaring studi paling relevan secara semantik, sembari menyadari batasan potensi bias peringkat algoritma mesin pencari. Untuk keperluan replikasi pada basis data terindeks, string Boolean Scopus yang setara disiapkan sebagai berikut: TITLE-ABS-KEY ( ( "large language model*″ OR llm* OR chatgpt OR "generative ai" OR genai OR copilot OR "gpt-4" ) AND ( programming OR coding OR "computer science education" OR cs1 ) AND ( education OR teaching OR learning OR student *OR course* ) ) AND PUBYEAR > 2020 AND ( LIMIT-TO ( DOCTYPE , "ar" ) OR LIMIT-TO ( DOCTYPE , "cp" ) ). String ini belum dijalankan dalam penelitian ini.

**Tabel 3 Kueri pencarian Consensus dan jumlah rekaman**

| No. | Kueri | Rekaman | Rekaman baru |
|---|---|---|---|
| 1 | large language models programming education systematic review | 10 | 10 |
| 2 | ChatGPT introductory programming course students learning outcomes | 10 | 10 |
| 3 | academic integrity plagiarism LLM generated code student programming assignments | 10 | 10 |
| 4 | GitHub Copilot novice programmers classroom study | 10 | 10 |
| 5 | overreliance on AI code generation students metacognitive difficulties programming | 10 | 8 |
| 6 | LLM-based programming tutor guardrails hints deployed in course | 10 | 10 |
| 7 | students' perceptions of ChatGPT in programming learning survey university | 10 | 7 |
| 8 | LLM-generated feedback on student programming assignments evaluation | 10 | 10 |
| 9 | prompt engineering skills teaching students to code with generative AI | 10 | 9 |
| 10 | ChatGPT pembelajaran pemrograman mahasiswa Indonesia | 10 | 4 |
| 11 | generative AI policy guidelines computing educators best practices integrating LLM into programming courses | 10 | 10 |
| 12 | LLM explanations of code and error messages for novice programmers | 10 | 10 |
| 13 | LLM-generated programming exercises and worked examples evaluated with students | 10 | 7 |
| 14 | equity gender prior experience differences novice use of AI code assistants programming course | 10 | 6 |

**Proses seleksi**

Seleksi dilakukan oleh penelaah tunggal dengan bantuan asisten AI terarah, di mana setiap keluaran penyaringan AI diverifikasi manual secara ketat terhadap kriteria kelayakan guna mencegah halusinasi data dan menjamin kepatuhan kriteria eksklusi. Setelah duplikat dihapus, rekaman disaring berdasarkan judul dan abstrak terhadap kriteria IC1 sampai IC5 dan EC1 sampai EC4. Rekaman yang lolos dinilai kelayakannya pada tahap berikutnya terhadap EC5 dan EC6. Teks lengkap tidak diambil; penilaian kelayakan memakai abstrak lengkap yang dikembalikan Consensus.

**Proses pengumpulan dan butir data**

Data diekstraksi oleh penelaah yang sama ke dalam lembar ekstraksi, yaitu penulis, tahun, venue, desain, ukuran sampel, jenjang, jenis LLM, temuan utama, dan kode tema. Tema manfaat meliputi capaian belajar (B1), motivasi, efikasi diri, dan keterlibatan (B2), efisiensi dan beban kerja (B3), umpan balik, penjelasan, dan bantuan debugging sesuai permintaan (B4), serta otomasi tugas pengajar (B5). Tema risiko meliputi integritas akademik (R1), ketergantungan berlebih dan alih daya kognitif (R2), keluaran yang tidak andal atau membocorkan solusi (R3), serta kesenjangan dan

homogenisasi (R4). Tema praktik meliputi perancah dan pagar pengaman pedagogis (P1), pelatihan literasi prompt (P2), dukungan regulasi diri dan metakognisi (P3), asesmen berbasis proses dan deteksi sebagai alat bantu (P4), pengawasan manusia (P5), serta kebijakan dan tujuan belajar yang eksplisit (P6). Untuk studi berdesain eksperimen atau pra-pasca, arah hasil belajar dikodekan sebagai positif, campuran, nol, atau negatif. Tidak ada penulis studi primer yang dihubungi.

**Penilaian kualitas studi**

Kualitas studi dinilai dengan instrumen enam butir yang diadaptasi dari panduan Kitchenham dan Charters (2007) (Tabel 4). Tiap butir diberi skor 1 (ya), 0,5 (sebagian), atau 0 (tidak), sehingga skor maksimum adalah 6. Studi dengan skor kurang dari 3,5 dieksklusi, skor 5 atau lebih dikategorikan tinggi, 4 sampai 4,5 sedang, dan kurang dari 4 rendah. Penilaian kualitas dilakukan secara terstandarisasi berdasarkan informasi abstrak; disadari bahwa studi primer yang kuat secara metodologi di teks lengkap dapat memperoleh penalti skor apabila rincian sampel tidak tercantum di abstrak.

**Tabel 4 Instrumen penilaian kualitas**

| Kode | Butir penilaian |
|---|---|
| QA1 | Tujuan atau pertanyaan penelitian dinyatakan jelas |
| QA2 | Desain, konteks, dan peserta atau data dideskripsikan memadai |
| QA3 | Pengumpulan dan analisis data sesuai dan transparan (kelompok pembanding, instrumen, triangulasi, atau metrik terdefinisi) |
| QA4 | Hasil dilaporkan jelas dengan statistik, metrik, atau tema yang konkret |
| QA5 | Keterbatasan atau hasil nol dan negatif dilaporkan |
| QA6 | Ukuran sampel atau data memadai (kuantitatif minimal 30, kualitatif minimal 10, atau kumpulan data minimal 100 sampel) |

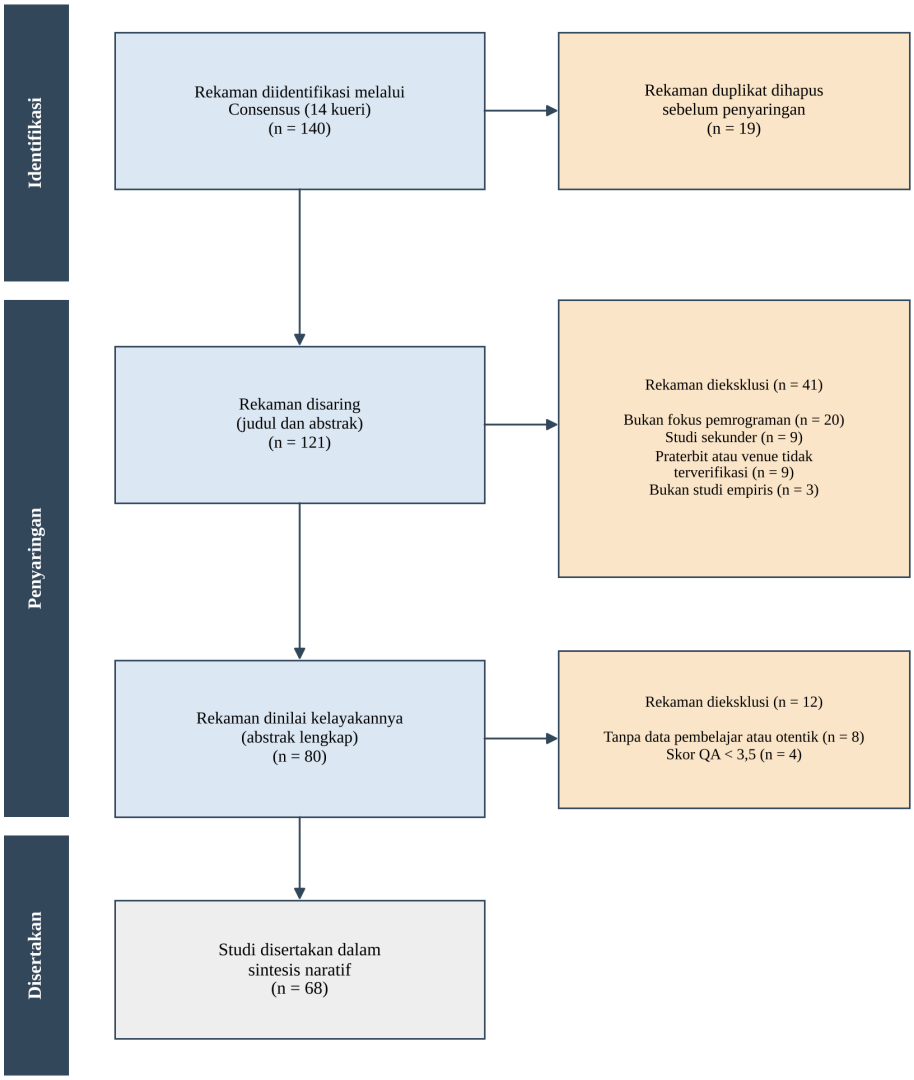
**Metode sintesis**

Sintesis dilakukan secara naratif tematik. Frekuensi tema dihitung per studi, dan satu studi dapat berkontribusi pada lebih dari satu tema. Arah hasil belajar pada studi eksperimen dirangkum dengan penghitungan suara menurut arah efek tanpa pembobotan ukuran sampel. Meta-analisis tidak dilakukan karena heterogenitas desain, ukuran hasil, dan versi LLM, sehingga ukuran efek tidak dihitung. Bias pelaporan dinilai secara informal dengan mencatat bahwa naskah praterbit dikeluarkan dan studi berhasil cenderung lebih mudah terbit. Kepastian bukti tidak dinilai dengan GRADE, melainkan dideskripsikan secara kualitatif menurut jumlah dan desain studi.

## HASIL DAN PEMBAHASAN

**Seleksi studi**

Penelusuran menghasilkan 140 rekaman. Setelah 19 duplikat dihapus, 121 rekaman disaring. Sebanyak 41 rekaman dieksklusi pada tahap penyaringan karena tidak berfokus pada pemrograman (n = 20), merupakan studi sekunder (n = 9), berstatus praterbit atau venue tidak terverifikasi (n = 9), atau bukan studi empiris (n = 3). Dari 80 rekaman yang dinilai kelayakannya, 12 dieksklusi, yaitu delapan karena hanya menguji kapabilitas model tanpa data pembelajar atau data otentik atau tidak menyatakan sumber data, dan empat karena skor kualitas di bawah 3,5. Sebanyak 68 studi disertakan dalam sintesis (Gambar 1). Satu rekaman dari Indonesia yang muncul pada kueri berbahasa Indonesia membahas kesiapan mahasiswa memakai ChatGPT secara umum sehingga dieksklusi karena tidak berfokus pada pemrograman. Daftar lengkap rekaman yang dieksklusi beserta alasannya tersedia dari penulis atas permintaan.




**Gambar 1 Diagram alir PRISMA 2020 proses seleksi studi**

## Karakteristik studi (RQ1)

Studi yang disertakan terbit pada 2023 (n = 12), 2024 (n = 20), 2025 (n = 20), dan 2026 sampai Oktober (n = 16); tidak ada studi terbit 2021 sampai 2022 yang memenuhi kriteria. Sebanyak 36 studi terbit di prosiding konferensi dan 32 di jurnal. Desain yang paling banyak adalah eksperimen atau kuasi-eksperimen (n = 21), diikuti evaluasi keluaran LLM pada data otentik (n = 17), studi observasional atau campuran (n = 11), survei atau wawancara (n = 11), dan pengembangan serta penerapan alat (n = 8). Mayoritas studi dilakukan di perguruan tinggi (n = 61), dengan lima studi di jenjang sekolah (Bhaskar et al. (2025), Chrysopoulos et al. (2025), Liu et al. (2025), Xu et al. (2026), Yang et al. (2025)) dan dua di lingkungan daring (Gabbay et al. (2024), Hellas et al. (2023)). Jenis LLM yang dipakai adalah ChatGPT (n = 24), alat kustom berbasis LLM (n = 15), beberapa model sekaligus (n = 13), GitHub Copilot (n = 5), dan GenAI tanpa spesifikasi model (n = 11). Ukuran sampel tidak dilaporkan pada abstrak 18 studi. Ringkasan setiap studi disajikan pada Tabel 5.

**Tabel 5 Karakteristik studi yang disertakan**

| No. | Studi | Desain | n | LLM | Jenjang | Skor QA |
|---|---|---|---|---|---|---|
| 1 | Balse et al. (2023) | Evaluasi | 30 program dan 5 asisten | ChatGPT | PT | 5,5 |
| 2 | Hellas et al. (2023) | Evaluasi | tidak dilaporkan | Beberapa model | Daring | 4,5 |
| 3 | Lau et al. (2023) | Survei | 20 instruktur | Beberapa model | PT | 6 |
| 4 | Liffiton et al. (2023) | Alat | 52 | Alat LLM | PT | 6 |
| 5 | Prather et al. (2023a) | Observasi | tidak dilaporkan | Copilot | PT | 4,5 |
| 6 | Prather et al. (2023b) | Observasi | 22 wawancara dan survei 20 negara | Beberapa model | PT | 6 |
| 7 | Roest et al. (2023) | Alat | tidak dilaporkan | Alat LLM | PT | 4,5 |
| 8 | Santos et al. (2023) | Evaluasi | tidak dilaporkan | ChatGPT | PT | 4,0 |
| 9 | Sheese et al. (2023) | Observasi | 52 (2.500 kueri) | Alat LLM | PT | 5,5 |
| 10 | Taylor et al. (2023) | Alat | 2.565 mahasiswa | Alat LLM | PT | 6 |
| 11 | Yılmaz et al. (2023a) | Survei | 41 | ChatGPT | PT | 6 |
| 12 | Yılmaz et al. (2023b) | Eksperimen | 45 | ChatGPT | PT | 5 |
| 13 | Azaiz et al. (2024) | Evaluasi | 55 submisi | ChatGPT | PT | 5,5 |
| 14 | Denny et al. (2024) | Survei | 813 | Alat LLM | PT | 5,0 |
| 15 | Estévez-Ayres et al. (2024) | Evaluasi | tidak dilaporkan | Beberapa model | PT | 4,5 |
| 16 | Gabbay et al. (2024) | Evaluasi | tidak dilaporkan | ChatGPT | Daring | 4,5 |
| 17 | Gardella et al. (2024) | Eksperimen | 17 | Copilot | PT | 5 |
| 18 | Groothuijsen et al. (2024) | Observasi | 29 | ChatGPT | PT | 5,5 |
| 19 | Haindl et al. (2024) | Survei | tidak dilaporkan | ChatGPT | PT | 4,0 |
| 20 | Hoq et al. (2024) | Evaluasi | dataset terbuka | ChatGPT | PT | 5,5 |
| 21 | Jury et al. (2024) | Alat | sekitar 400 | Alat LLM | PT | 5,5 |
| 22 | Kazemitabaar et al. (2024) | Alat | 700 mahasiswa | Alat LLM | PT | 6 |
| 23 | Kharrufa et al. (2024) | Survei | 39 survei dan 8 wawancara | Beberapa model | PT | 5,5 |
| 24 | Kosar et al. (2024) | Eksperimen | 182 | ChatGPT | PT | 5,5 |
| 25 | Koutcheme et al. (2024) | Evaluasi | tidak dilaporkan | Beberapa model | PT | 4,0 |
| 26 | Lee et al. (2024) | Eksperimen | tidak dilaporkan | ChatGPT | PT | 3,5 |
| 27 | Prather et al. (2024) | Observasi | 21 | GenAI umum | PT | 6 |
| 28 | Santos et al. (2024) | Eksperimen | 106 | ChatGPT | PT | 6 |
| 29 | Scholl et al. (2024) | Survei | 298 | ChatGPT | PT | 5,5 |
| 30 | Song et al. (2024) | Alat | 60 | Alat LLM | PT | 5,5 |
| 31 | Sun et al. (2024) | Eksperimen | 82 | ChatGPT | PT | 5,5 |
| 32 | Xue et al. (2024) | Eksperimen | 56 | ChatGPT | PT | 6 |
| 33 | Andleeb et al. (2025) | Eksperimen | tidak dilaporkan | ChatGPT | PT | 5 |
| 34 | Bernik et al. (2025) | Evaluasi | 100 dari 315 tugas | Beberapa model | PT | 6 |

| No. | Studi | Desain | n | LLM | Jenjang | Skor QA |
|---|---|---|---|---|---|---|
| 35 | Bhaskar et al. (2025) | Alat | 105 | Alat LLM | K-12 | 3,5 |
| 36 | Chrysopoulos et al. (2025) | Alat | 18 siswa dan 3 guru | Alat LLM | K-12 | 4,0 |
| 37 | Garg et al. (2025) | Eksperimen | 157 | GenAI umum | PT | 5,5 |
| 38 | Güner et al. (2025) | Eksperimen | tidak dilaporkan | ChatGPT | PT | 4,0 |
| 39 | Huang et al. (2025) | Eksperimen | tidak dilaporkan | Alat LLM | PT | 4,0 |
| 40 | Karsa et al. (2025) | Evaluasi | 85 hingga 160 mahasiswa | Beberapa model | PT | 6 |
| 41 | Kerslake et al. (2025) | Observasi | 21 | Alat LLM | PT | 5,5 |
| 42 | Khan et al. (2025) | Observasi | 19 | Alat LLM | PT | 5 |
| 43 | Li et al. (2025) | Eksperimen | tidak dilaporkan | GenAI umum | PT | 4,5 |
| 44 | Liu et al. (2025) | Evaluasi | 14.138 sampel | ChatGPT | K-12 | 5,5 |
| 45 | Lohr et al. (2025) | Evaluasi | tidak dilaporkan | GenAI umum | PT | 4,0 |
| 46 | Nathaniel et al. (2025) | Eksperimen | 25 | GenAI umum | PT | 5,5 |
| 47 | Rahe et al. (2025) | Eksperimen | 37 | ChatGPT | PT | 6 |
| 48 | Santos et al. (2025) | Eksperimen | 80 | ChatGPT | PT | 6 |
| 49 | Shah et al. (2025) | Observasi | tidak dilaporkan | Copilot | PT | 4,0 |
| 50 | Shihab et al. (2025) | Eksperimen | 10 | Copilot | PT | 5 |
| 51 | Wang et al. (2025) | Eksperimen | 83 | ChatGPT | PT | 6 |
| 52 | Yang et al. (2025) | Eksperimen | 153 | ChatGPT | K-12 | 6 |
| 53 | Alquran et al. (2026) | Survei | 248 | Beberapa model | PT | 5,5 |
| 54 | Aziz (2026) | Survei | 114 | Copilot | PT | 5,5 |
| 55 | Bai et al. (2026) | Eksperimen | tidak dilaporkan | Alat LLM | PT | 4,0 |
| 56 | Chiu et al. (2026) | Survei | 260 | ChatGPT | PT | 5,5 |
| 57 | de Oliveira et al. (2026) | Observasi | tidak dilaporkan | Beberapa model | PT | 4,0 |
| 58 | de Souza Santos et al. (2026) | Survei | 116 | GenAI umum | PT | 5,5 |
| 59 | Denny et al. (2026) | Evaluasi | 950 | GenAI umum | PT | 6 |
| 60 | Haseeb et al. (2026) | Evaluasi | 57 tugas | Beberapa model | PT | 5,0 |
| 61 | Jukiewicz (2026) | Evaluasi | lebih dari 6.000 tugas | Beberapa model | PT | 5,5 |
| 62 | Liu et al. (2026) | Observasi | 41 | GenAI umum | PT | 6 |
| 63 | Maisch et al. (2026) | Observasi | tidak dilaporkan | GenAI umum | PT | 4,0 |
| 64 | Marmolejo-Ramos et al. (2026) | Survei | 20 | GenAI umum | PT | 5,0 |
| 65 | Ramachandra et al. (2026) | Evaluasi | sekitar 350 mahasiswa | GenAI umum | PT | 5,0 |
| 66 | Vysocan et al. (2026) | Evaluasi | 16.271 sampel | Beberapa model | PT | 6 |
| 67 | Xu et al. (2026) | Eksperimen | 71 | ChatGPT | K-12 | 5,5 |
| 68 | Çoban et al. (2026) | Eksperimen | 30 | Alat LLM | PT | 6 |

Keterangan: PT = perguruan tinggi; n = ukuran sampel atau data pada abstrak; skor QA dari maksimum 6.

**Kualitas studi**

Skor kualitas rata-rata adalah 5,18 dari 6 (SD = 0,76; rentang 3,5 sampai 6). Sebanyak 49 studi berkualitas tinggi, 17 sedang, dan 2 rendah. Butir yang paling lemah adalah ukuran sampel atau data (QA6, rata-rata 0,62) dan pelaporan keterbatasan (QA5, rata-rata 0,75), sedangkan tujuan penelitian (QA1) dan pelaporan hasil (QA4) hampir selalu terpenuhi. Pola ini sejalan dengan temuan tinjauan naratif yang melaporkan sampel kecil, intervensi singkat, dan ketergantungan pada data swalaporan (Farooq et al., 2026).

**Manfaat pemanfaatan LLM (RQ2)**

Sebanyak 54 dari 68 studi melaporkan setidaknya satu manfaat (Tabel 6). Manfaat yang paling sering dilaporkan adalah umpan balik, penjelasan, dan bantuan debugging sesuai permintaan (30 studi). Asisten dengan pagar pengaman yang dipakai selama satu semester dinilai bermanfaat karena selalu tersedia dan membantu menyelesaikan galat (Liffiton et al., 2023), dan dalam penerapan skala besar penjelasan galat kompilasi akurat secara konseptual pada 90 persen kasus galat saat kompilasi (Taylor et al., 2023). Pada survei terhadap 813 mahasiswa, dukungan instan menjadi karakteristik yang paling dihargai (Denny et al., 2024). Mahasiswa juga melaporkan manfaat berupa jawaban cepat, bantuan debugging, dan peningkatan kepercayaan diri (Yılmaz et al., 2023a).

Manfaat afektif dan efisiensi juga konsisten dilaporkan. Motivasi, efikasi diri, dan keterlibatan meningkat pada 9 studi, misalnya pada kelompok yang memakai ChatGPT dengan strategi pembelajaran berregulasi diri (Wang et al., 2025) dan pada pembelajar sekolah dasar (Xu et al., 2026). Efisiensi dan beban kerja membaik pada 7 studi: mahasiswa menyelesaikan tugas 34,9 persen lebih cepat dengan Copilot pada tugas pengembangan di atas kode yang sudah ada (Shihab et al., 2025), dan beban mental turun pada novis yang bekerja di bawah tekanan waktu (Gardella et al., 2024). Bagi pengajar, otomasi penilaian dan pembuatan materi dilaporkan pada 8 studi. Skor LLM berkorelasi tinggi dengan skor dosen (r hingga 0,91; Bernik et al., 2025), sedangkan korelasi pada ujian pemrograman berkisar 0,6 sampai 0,9 dan dipengaruhi inkonsistensi penilai manusia (Karsa et al., 2025). Proyek latihan buatan LLM dinilai sebanding dengan buatan manusia (Song et al., 2024).

Bukti pada capaian belajar lebih berimbang. Di antara 21 studi eksperimen atau pra-pasca yang melaporkan arah hasil belajar, 10 melaporkan hasil positif, 5 campuran, 5 nol, dan 1 negatif. Hasil positif dilaporkan, antara lain, oleh Garg et al. (2025), Xu et al. (2026), Yılmaz et al. (2023b). Hasil nol terjadi pada eksperimen berukuran besar, yaitu 182 mahasiswa (Kosar et al., 2024), 106 mahasiswa (Santos et al., 2024), dan 82 mahasiswa (Sun et al., 2024), sedangkan hasil negatif ditemukan pada 153 siswa sekolah menengah dengan aliran pengalaman, efikasi diri, dan capaian lebih rendah (Yang et al., 2025). Hasil campuran muncul ketika efek pada motivasi tidak diikuti peningkatan pengetahuan (Wang et al., 2025) atau ketika capaian naik tetapi transfer pengetahuan jangka panjang menurun (Li et al., 2025). Dukungan terstruktur, yaitu perancah, petunjuk bertahap, atau pelatihan prompt, muncul pada beberapa studi positif (Bai et al. (2026), Garg et al. (2025), Huang et al. (2025), Nathaniel et al. (2025)), tetapi juga pada dua studi dengan hasil nol atau negatif (Yang et al. (2025), Çoban et al. (2026)), sehingga tidak ada pola konsisten yang dapat disimpulkan karena desain, jenjang, dan ukuran hasil berbeda antarstudi.

**Tabel 6 Ringkasan tema manfaat dan risiko**

| Jenis | Tema | Studi (n) | Contoh studi |
|---|---|---|---|
| Manfaat | Umpan balik, penjelasan, dan bantuan debugging sesuai permintaan | 30 | Liffiton et al. (2023); Taylor et al. (2023); Denny et al. (2024) |

| Jenis | Tema | Studi (n) | Contoh studi |
|---|---|---|---|
| Manfaat | Capaian belajar | 10 | Garg et al. (2025); Xu et al. (2026); Yılmaz et al. (2023b) |
| Manfaat | Motivasi, efikasi diri, dan keterlibatan | 9 | Wang et al. (2025); Yılmaz et al. (2023b); Huang et al. (2025) |
| Manfaat | Efisiensi dan beban kerja | 7 | Shihab et al. (2025); Gardella et al. (2024); Andleeb et al. (2025) |
| Manfaat | Otomasi tugas pengajar (penilaian dan pembuatan materi) | 8 | Bernik et al. (2025); Karsa et al. (2025); Jury et al. (2024) |
| Risiko | Ketergantungan berlebih dan alih daya kognitif | 22 | Rahe et al. (2025); Prather et al. (2024); Li et al. (2025) |
| Risiko | Keluaran tidak andal atau membocorkan solusi | 20 | Hellas et al. (2023); Estévez-Ayres et al. (2024); Taylor et al. (2023) |
| Risiko | Integritas akademik dan plagiarisme | 9 | Ramachandra et al. (2026); de Souza Santos et al. (2026); Lau et al. (2023) |
| Risiko | Kesenjangan dan homogenisasi | 2 | Prather et al. (2024); Denny et al. (2026) |

**Risiko pemanfaatan LLM (RQ3)**

Sebanyak 44 dari 68 studi melaporkan risiko. Risiko yang paling sering adalah ketergantungan berlebih dan alih daya kognitif (22 studi). Pada eksperimen terpantau terhadap 37 mahasiswa, sebagian besar pengguna chatbot langsung meminta solusi lengkap dan terjebak dalam siklus mengirim kode yang salah lalu meminta perbaikan (Rahe et al., 2025). Pengamatan laboratorium atas 21 sesi menunjukkan bahwa mahasiswa yang kesulitan justru mengalami kesulitan metakognitif yang makin besar dan ilusi kompetensi, sehingga AI generatif berpotensi memperlebar kesenjangan antara mahasiswa yang berhasil dan yang tidak (Prather et al., 2024). Alih daya kognitif dan transfer pengetahuan yang melemah ditemukan pula oleh Li et al. (2025), Liu et al. (2026), Marmolejo-Ramos et al. (2026), dan dosen pada studi kasus komputasi ilmiah mengkhawatirkan penurunan kualitas kode serta melemahnya kolaborasi pada pemrograman berpasangan (Groothuijsen et al., 2024).

Risiko keandalan keluaran dilaporkan pada 20 studi. Umpan balik ChatGPT dan Bard hanya mencapai akurasi sekitar 50 persen pada kesalahan konkurensi (Estévez-Ayres et al., 2024). GPT-3.5 menemukan setidaknya satu masalah pada 90 persen permintaan bantuan pemula, tetapi 40 persen keluarannya berupa positif palsu dan solusi lengkap sering diberikan meskipun diminta tidak (Hellas et al., 2023). Larangan memberi solusi juga sering dilanggar pada alat yang diterapkan kepada ribuan mahasiswa (Taylor et al., 2023). Separuh penjelasan galat logika memuat setidaknya satu pernyataan keliru (Balse et al., 2023), dan dalam eksperimen terkontrol penjelasan GPT-4 mengungguli pesan galat standar hanya pada satu dari enam tugas (Santos et al., 2024). Antarmodel, konsistensi penilaian umpan balik sangat rendah (ICC = 0,061; Jukiewicz, 2026).

Risiko integritas akademik dilaporkan pada 9 studi, sebagian besar berupa pengembangan dan evaluasi alat deteksi. Pada satu mata kuliah pengantar dengan sekitar 350 mahasiswa, 10 sampai 15 persen mahasiswa dikenai sanksi karena plagiarisme dari LLM, dan model deteksi mencapai presisi serta recall di atas 98 persen (Ramachandra et al., 2026). Model serupa mencapai akurasi di atas 90 persen pada kode CS1 (Hoq et al., 2024) dan 98,97 persen pada pseudokode siswa sekolah menengah (Liu et al., 2025), tetapi performanya menurun pada jenis tugas yang belum pernah dilihat sehingga detektor disarankan hanya sebagai alat peringkat risiko dengan tinjauan manusia (Vysocan et al., 2026). Survei terhadap 116 mahasiswa rekayasa perangkat lunak menunjukkan bahwa penggunaan LLM yang dipersepsikan tidak pantas terutama terjadi pada tugas pemrograman akibat tekanan waktu dan panduan yang tidak jelas (de Souza Santos et al., 2026), sedangkan wawancara dengan 20 instruktur dari sembilan negara memperlihatkan perpecahan antara melarang dan

mengintegrasikan LLM (Lau et al., 2023). Dua studi juga mengangkat risiko kesenjangan dan homogenisasi, yaitu kesenjangan antara pemrogram yang berhasil dan yang kesulitan (Prather et al., 2024) serta contoh buatan AI yang lebih seragam daripada buatan mahasiswa (Denny et al., 2026).

**Praktik terbaik (RQ4)**

Sebanyak 58 dari 68 studi menguji atau merekomendasikan praktik pembelajaran, yang dikelompokkan menjadi enam kategori (Tabel 7). Praktik yang paling sering adalah perancah dan pagar pengaman pedagogis (24 studi): alat seperti CodeHelp dan CodeAid memberi bantuan tanpa langsung membocorkan solusi (Liffiton et al., 2023; Kazemitabaar et al., 2024), dan mahasiswa lebih menyukai perancah yang menjaga otonomi daripada solusi langsung (Denny et al., 2024). Perancah bertingkat terbukti meningkatkan skor pemrograman dan menurunkan beban kognitif (Huang et al., 2025). Namun, pagar pengaman berbasis prompt belum sepenuhnya dapat diandalkan karena kebocoran solusi masih dilaporkan (Taylor et al., 2023; Hellas et al., 2023).

Pelatihan literasi prompt (7 studi) menunjukkan hasil menjanjikan: kelompok dengan pelatihan prompt terstruktur memperoleh skor pascates tertinggi pada 157 mahasiswa teknik (Garg et al., 2025), kerangka GenAI-Ped dengan pelatihan prompt meningkatkan pemecahan masalah dan logika pemrograman sambil menekan ketergantungan (Nathaniel et al., 2025), dan pelatihan prompt mengubah profil interaksi mahasiswa dari pembangkit kode yang bergantung pada AI menjadi pengguna yang kolaboratif (Güner et al., 2025). Dukungan regulasi diri dan metakognisi (9 studi) mencakup strategi pembelajaran berregulasi diri yang meningkatkan efikasi diri (Wang et al., 2025) serta fase mengkritik keluaran AI dan pemudaran bertahap yang diusulkan dari teori berbasis data (Liu et al., 2026).

Praktik terkait asesmen dan integritas (9 studi) menekankan perancangan ulang tugas, pelaporan penggunaan AI, dan deteksi sebagai alat bantu, bukan sebagai penegak aturan tunggal (Vysocan et al., 2026; Prather et al., 2023b). Pengawasan manusia (17 studi) menjadi tema penting pada penilaian otomatis, yaitu evaluasi hibrida dengan pengawasan pengajar (Bernik et al., 2025), umpan balik LLM sebagai pelengkap umpan balik berbasis uji (Gabbay et al., 2024), dan peninjauan oleh asisten pengajar yang terlatih (Balse et al., 2023). Terakhir, kebijakan dan tujuan belajar yang eksplisit (10 studi) disarankan agar penggunaan AI selaras dengan keterampilan yang hendak dinilai (Groothuijsen et al., 2024; Scholl et al., 2024). Sebagian besar praktik didukung oleh studi observasional atau persepsi, sehingga tingkat dukungan buktinya hanya sedang atau terbatas.

**Tabel 7 Praktik pembelajaran dan tingkat dukungan bukti**

| Praktik | Studi (n) | Eksperimen (n) | Dukungan bukti | Contoh studi |
|---|---|---|---|---|
| P1 Perancah dan pagar pengaman pedagogis | 24 | 6 | Sedang | Liffiton et al. (2023); Kazemitabaar et al. (2024); Huang et al. (2025) |
| P2 Pelatihan literasi prompt | 7 | 4 | Sedang | Garg et al. (2025); Nathaniel et al. (2025); Güner et al. (2025) |
| P3 Dukungan regulasi diri dan metakognisi | 9 | 7 | Sedang | Wang et al. (2025); Liu et al. (2026); Bai et al. (2026) |
| P4 Asesmen berbasis proses dan deteksi sebagai alat bantu | 9 | 0 | Terbatas | Vysocan et al. (2026); Hoq et al. (2024); Prather et al. (2023b) |
| P5 Pengawasan manusia pada umpan balik dan penilaian | 17 | 2 | Terbatas | Bernik et al. (2025); Gabbay et al. (2024); Balse et al. (2023) |
| P6 Kebijakan dan tujuan belajar yang mencakup penggunaan AI | 10 | 3 | Terbatas | Lau et al. (2023); Groothuijsen et al. (2024); Scholl et al. (2024) |

Keterangan: dukungan sedang berarti ada beberapa studi eksperimen dengan hasil positif; terbatas berarti terutama studi evaluasi, observasi, atau persepsi. Penentuan ini kualitatif dan bukan penilaian GRADE.

## PEMBAHASAN

Sintesis ini menunjukkan bahwa LLM bukan intervensi dengan efek seragam. Khusus bagi konteks pendidikan tinggi ilmu komputer di Indonesia, pendidik perlu menyeimbangkan adopsi alat bantu AI dengan perancangan asesmen otentik berbasis proses (process-based assessment) serta pembekalan literasi prompt etis guna mencegah ketergantungan kognitif dan kecurangan kode. Manfaat yang paling konsisten bersifat dukungan, yaitu umpan balik instan, penjelasan, dan bantuan debugging, serta dampak afektif berupa motivasi dan kepercayaan diri. Capaian belajar tidak otomatis meningkat: eksperimen berukuran besar cenderung tidak menemukan perbedaan, dan beberapa studi menunjukkan bahwa manfaat jangka pendek dapat disertai penurunan transfer pengetahuan. Temuan ini sejalan dengan kekhawatiran ketergantungan berlebih dan integritas akademik pada tinjauan sebelumnya (Fu et al., 2025) serta dengan kesimpulan bahwa pengawasan manusia perlu dipadukan dengan umpan balik otomatis (Pereira et al., 2025). Dengan memasukkan studi 2025 sampai 2026 dan menilai kualitas studi, tinjauan ini menambahkan bahwa hasil positif sering dilaporkan pada studi yang menguji dukungan pedagogis terstruktur, meskipun pola ini belum konsisten.

Implikasi praktis bagi pendidik pemrograman di perguruan tinggi dapat diturunkan dari bukti tersebut, dengan catatan bahwa tidak ada studi terpilih yang secara eksplisit berlatar Indonesia pada abstraknya. Pertama, kebijakan penggunaan AI per tugas sebaiknya dirumuskan eksplisit dan diselaraskan dengan capaian pembelajaran, karena kecurangan banyak dikaitkan dengan tekanan waktu dan panduan yang tidak jelas. Kedua, bila LLM dipakai, pilih atau konfigurasikan alat yang memberi petunjuk bertahap dan tidak langsung memberi solusi, sambil menyadari bahwa pagar pengaman berbasis prompt masih bisa bocor. Ketiga, ajarkan literasi prompt dan verifikasi keluaran sebagai bagian kurikulum. Keempat, nilai proses dan pemahaman, misalnya lewat riwayat penyuntingan atau penjelasan lisan, dan perlakukan detektor hanya sebagai alat peringkat risiko. Kelima, pada penilaian otomatis, pertahankan peninjauan manusia karena konsistensi antarmodel dan akurasi pada kesalahan kompleks masih bervariasi.

Penelitian ini memiliki keterbatasan. Pertama, penelusuran hanya memakai satu agregator dengan sepuluh hasil teratas per kueri dan tidak menelusuri Scopus, IEEE Xplore, dan ACM Digital Library secara langsung, sehingga cakupan tidak dapat dianggap menyeluruh. Kedua, seleksi, ekstraksi, dan penilaian kualitas dilakukan oleh satu penelaah dengan bantuan asisten AI tanpa penelaah kedua, sehingga risiko kesalahan dan bias penelaah tidak dapat dikuantifikasi. Ketiga, ekstraksi dan penilaian kualitas bersumber dari abstrak dan bukan teks lengkap, sehingga ukuran sampel, keterbatasan, dan detail metode dapat kurang terwakili. Keempat, tidak ada protokol terdaftar, dan dua aturan operasional difinalkan selama proses. Kelima, penghitungan suara arah efek tidak memperhitungkan ukuran sampel, dan versi LLM yang berubah cepat dari GPT-3.5 hingga model generasi terbaru menyulitkan perbandingan antarstudi. Keenam, hanya publikasi berbahasa Inggris yang akhirnya disertakan. Seluruh angka dalam naskah ini perlu diverifikasi terhadap teks lengkap sebelum dipakai sebagai bukti.

Agenda penelitian selanjutnya meliputi studi longitudinal dengan uji transfer tanpa bantuan AI, replikasi di konteks Indonesia dan kawasan berkembang, kajian kesetaraan menurut bahasa, gender, dan pengetahuan awal, serta pelaporan standar atas versi model, prompt, dan konfigurasi pagar pengaman agar studi dapat dibandingkan.

## KESIMPULAN

Tinjauan sistematis atas 68 studi (2023 sampai 2026) menunjukkan bahwa pemanfaatan LLM dalam pembelajaran pemrograman memberi manfaat nyata pada dukungan umpan balik, motivasi, dan efisiensi, tetapi tidak menjamin peningkatan capaian belajar: dari 21 studi dengan ukuran hasil belajar, 10 positif, 5 campuran, 5 nol, dan 1 negatif. Risiko utama adalah ketergantungan berlebih dan alih daya kognitif (22 studi), keluaran yang tidak andal atau membocorkan solusi (20 studi), dan persoalan integritas akademik (9 studi). Praktik yang paling menjanjikan adalah perancah pedagogis, pelatihan literasi prompt, dukungan regulasi diri, asesmen berbasis proses, pengawasan manusia, serta kebijakan penggunaan AI yang eksplisit. Karena penelusuran terbatas pada satu agregator dan seleksi dilakukan satu penelaah berbasis abstrak, temuan ini perlu dikonfirmasi dengan penelusuran Scopus, IEEE Xplore, dan ACM Digital Library serta penelaahan teks lengkap.

## PERNYATAAN PEMANFAATAN TEKNOLOGI ASISTIF

Para penulis bertanggung jawab penuh atas isi naskah ini. Asisten AI generatif (Claude, Anthropic) dimanfaatkan secara substansial dalam penelusuran literatur melalui Consensus, penyaringan rekaman, ekstraksi data, penilaian kualitas, penghitungan statistik deskriptif, dan penyusunan draf naskah. Para penulis wajib memverifikasi seluruh rekaman, angka, dan sitasi terhadap teks lengkap sumbernya, melengkapi data referensi, dan menyetujui isi sebelum naskah diajukan.


## UCAPAN TERIMA KASIH

Penulis menyampaikan terima kasih yang sebesar-besarnya kepada Program Studi Teknik Informatika, Fakultas Teknik dan Ilmu Komputer, Universitas Muhadi Setiabudi atas dukungan sarana riset, akses basis data literatur, dan fasilitasi akademik selama pelaksanaan systematic literature review ini.


## KONTRIBUSI PENULIS

Bambang Irawan: Conceptualization, Methodology, Investigation, Formal Analysis, Data Curation, Writing - Original Draft, Writing - Review and Editing. Penulis telah membaca dan menyetujui naskah versi akhir.

## PERNYATAAN KETERSEDIAAN DATA

Lembar ekstraksi data 68 studi, matriks penilaian kualitas, dan rekaman kueri Consensus tersedia dari penulis atas permintaan yang wajar untuk tujuan akademik dan replikasi ilmiah.

## PERNYATAAN BEBAS KONFLIK KEPENTINGAN

Para penulis menyatakan bahwa tidak ada konflik kepentingan finansial maupun hubungan pribadi yang dapat mempengaruhi hasil yang dilaporkan dalam naskah ini.

## PERNYATAAN ETIKA PENELITIAN DAN PUBLIKASI

Para penulis menyatakan bahwa penelitian dan penulisan naskah ini telah mematuhi standar etika penelitian dan publikasi, sesuai dengan prinsip ilmiah, serta bebas dari plagiasi.

## REFERENSI


Alquran, H. et al. (2026). Human-AI collaboration in programming education: Student perspectives on LLM-based coding assistants. *Computers*.

Álvarez Ariza, J. et al. (2025). Generative AI in engineering and computing education: A scoping review of empirical studies and educational practices. *IEEE Access*.

Andleeb, S. et al. (2025). ChatGPT in introductory programming: Counterbalanced evaluation of code quality, conceptual learning, and student perceptions. In *Proceedings of the 26th ACM Annual Conference on Cybersecurity and Information Technology Education*.

Azaiz, I. et al. (2024). Feedback-generation for programming exercises with GPT-4. In *Proceedings of the 2024 on Innovation and Technology in Computer Science Education V. 1*.

Aziz, N. (2026). Technology acceptance of GitHub Copilot: A study among computing students. *Journal of Computational Innovation and Analytics (JCIA)*.

Bai, S.-R. et al. (2026). Enhancing the effect of AI-assisted learning: The use of scaffolding strategies to develop students' prompt engineering skills. *Interactive Learning Environments*.

Balse, R. et al. (2023). Evaluating the quality of LLM-generated explanations for logical errors in CS1 student programs. In *Proceedings of the 16th Annual ACM India Compute Conference*.

Bernik, A. et al. (2025). A comparative study of large language models in programming education: Accuracy, efficiency, and feedback in student assignment grading. *Applied Sciences*.

Bhaskar, S. et al. (2025). Designing answer-aware LLM hints to scaffold deeper learning in K-12 programming education. In *Proceedings of the 2025 ACM Conference on International Computing Education Research V.2*.

Chiu, K.-Y. et al. (2026). Perceived effectiveness and usefulness of ChatGPT as an aid for learning programming: A quantitative study based on a novel 8E learning technology model. *Interactive Learning Environments*.

Chrysopoulos, M. et al. (2025). Enhancing high school programming education through LLM-based hint generation. In *2025 International Workshop on Artificial Intelligence and Education (WAIE)*.

Çoban, E. et al. (2026). AI-Code Feedback, an AI-based companion for enhancing introductory programming education. In *2026 IEEE Global Engineering Education Conference (EDUCON)*.

de Oliveira, R. M. et al. (2026). Generative AI in introductory programming for non-CS STEM students: A two-semester study. *IEEE Access*.

de Souza Santos, R. E. et al. (2026). LLM use, cheating, and academic integrity in software engineering education. In *Proceedings of the 34th ACM International Conference on the Foundations of Software Engineering*.

Denny, P. et al. (2024). Desirable characteristics for AI teaching assistants in programming education. In *Proceedings of the 2024 on Innovation and Technology in Computer Science Education V. 1*.

Denny, P. et al. (2026). Assessing the quality and diversity of student- and AI-generated programming exercises from limited examples: Empirical insights from a large-scale blind evaluation. In *Proceedings of the 28th Western Canadian Conference on Computing Education*.

Estévez-Ayres, I. et al. (2024). Evaluation of LLM tools for feedback generation in a course on concurrent programming. *International Journal of Artificial Intelligence in Education*.

Farooq, U. et al. (2026). AI code assistants in programming education: A narrative literature review. *Education Sciences*.

Fu, Q. et al. (2025). Large language models (LLMs) in programming learning: The current research state and agenda. *IEEE Transactions on Learning Technologies*.

Gabbay, H. et al. (2024). Combining LLM-generated and test-based feedback in a MOOC for programming. In *Proceedings of the Eleventh ACM Conference on Learning @ Scale*.

Gardella, N. et al. (2024). Performance, workload, emotion, and self-efficacy of novice programmers using AI code generation. In *Proceedings of the 2024 on Innovation and Technology in Computer Science Education V. 1*.

Garg, A. et al. (2025). Enhancing data analysis and programming skills through structured prompt training: The impact of generative AI in engineering education. *Computers and Education: Artificial Intelligence*.

Groothuijsen, S. et al. (2024). AI chatbots in programming education: Students' use in a scientific computing course and consequences for learning. *Computers and Education: Artificial Intelligence*.

Güner, H. et al. (2025). AI in the classroom: Exploring students' interaction with ChatGPT in programming learning. *Education and Information Technologies*.

Haindl, P. et al. (2024). Students' experiences of using ChatGPT in an undergraduate programming course. *IEEE Access*.

Haseeb, M. et al. (2026). From code to rubrics: A multimodal evaluation of LLM systems for automated feedback generation. In *Companion Proceedings of the ACM Web Conference 2026*.

Hellas, A. et al. (2023). Exploring the responses of large language models to beginner programmers' help requests. In *Proceedings of the 2023 ACM Conference on International Computing Education Research - Volume 1*.

Hoq, M. et al. (2024). Detecting ChatGPT-generated code submissions in a CS1 course using machine learning models. In *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1*.

Huang, A. Y. Q. et al. (2025). The impact of GenAI-enabled coding hints on students' programming performance and cognitive load in an SRL-based Python course. *British Journal of Educational Technology*.

Jukiewicz, M. (2026). Multimodel sentiment analysis of feedback generated by large language models in programming assessment. *Discover Artificial Intelligence*.

Jury, B. et al. (2024). Evaluating LLM-generated worked examples in an introductory programming course. In *Proceedings of the 26th Australasian Computing Education Conference*.

Karsa, Z. I. et al. (2025). Automatic evaluation of programming tasks supported by language models. *IEEE Access*.

Kazemitabaar, M. et al. (2024). CodeAid: Evaluating a classroom deployment of an LLM-based programming assistant that balances student and educator needs. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*.

Kerslake, C. et al. (2025). Exploring student reactions to LLM-generated feedback on explain in plain English problems. In *Proceedings of the 56th ACM Technical Symposium on Computer Science Education V. 1*.

Khan, M. F. A. et al. (2025). Student-perceived cognitive load of LLM-generated programming exercises. In *2025 IEEE 12th International Conference on Data Science and Advanced Analytics (DSAA)*.

Kharrufa, A. et al. (2024). LLMs integration in software engineering team projects: Roles, impact, and a pedagogical design space for AI tools in computing education. *ACM Transactions on Computing Education*.

Kitchenham, B., and Charters, S. (2007). *Guidelines for performing systematic literature reviews in software engineering* (EBSE Technical Report EBSE-2007-01). Keele University and University of Durham.

Kosar, T. et al. (2024). Computer science education in ChatGPT era: Experiences from an experiment in a programming course for novice programmers. *Mathematics*.

Koutcheme, C. et al. (2024). Evaluating language models for generating and judging programming feedback. In *Proceedings of the 56th ACM Technical Symposium on Computer Science Education V. 1*.

Lau, S. et al. (2023). From "ban it till we understand it" to "resistance is futile": How university programming instructors plan to adapt as more students use AI code generation and explanation tools such as ChatGPT and GitHub Copilot. In *Proceedings of the 2023 ACM Conference on International Computing Education Research - Volume 1*.

Lee, J. S. Y. et al. (2024). Code debugging with LLM-generated explanations of programming error messages. In *2024 IEEE 13th International Conference on Engineering Education (ICEED)*.

Li, S. et al. (2025). Generative artificial intelligence-supported programming education: Effects on learning performance, self-efficacy and processes. *Australasian Journal of Educational Technology*.

Liffiton, M. H. et al. (2023). CodeHelp: Using large language models with guardrails for scalable support in programming classes. In *Proceedings of the 23rd Koli Calling International Conference on Computing Education Research*.

Liu, D. et al. (2026). Tool, tutor, or crutch?: A grounded theory of cognitive scaffolding and offloading in AI-assisted programming education. *International Journal of STEM Education*.

Liu, Z.-F. et al. (2025). What are the differences between student and ChatGPT-generated pseudocode? Detecting AI-generated pseudocode in high school programming using explainable machine learning. *Education and Information Technologies*.

Lohr, D. et al. (2025). You're (not) my type - Can LLMs generate feedback of specific types for introductory programming tasks?. *Journal of Computer Assisted Learning*.

Maisch, R. et al. (2026). Criminal minds: How first-year CS students plagiarize code. In *Proceedings of the 34th ACM International Conference on the Foundations of Software Engineering*.

Marmolejo-Ramos, F. et al. (2026). Between tool and trouble. *Learning Letters*.

Nathaniel, J. et al. (2025). Investigating the impact of generative AI integration on the sustenance of higher-order thinking skills and understanding of programming logic. *Computers and Education: Artificial Intelligence*.

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

Pereira, A. et al. (2025). A systematic literature review on large language models applications in computer programming teaching evaluation process. *IEEE Access*.

Prather, J. et al. (2023a). "It's weird that it knows what I want": Usability and interactions with Copilot for novice programmers. *ACM Transactions on Computer-Human Interaction*.

Prather, J. et al. (2023b). The robots are here: Navigating the generative AI revolution in computing education. In *Proceedings of the 2023 Working Group Reports on Innovation and Technology in Computer Science Education*.

Prather, J. et al. (2024). The widening gap: The benefits and harms of generative AI for novice programmers. In *Proceedings of the 2024 ACM Conference on International Computing Education Research - Volume 1*.

Rahe, C. et al. (2025). How do programming students use generative AI?. *Proceedings of the ACM on Software Engineering*.

Raihan, N. et al. (2024). Large language models in computer science education: A systematic literature review. In *Proceedings of the 56th ACM Technical Symposium on Computer Science Education V. 1*.

Ramachandra, A. et al. (2026). Detecting AI-generated code in introductory programming courses. In *Proceedings of the 57th ACM Technical Symposium on Computer Science Education V.1*.

Roest, L. et al. (2023). Next-step hint generation for introductory programming using large language models. In *Proceedings of the 26th Australasian Computing Education Conference*.

Santos, E. A. et al. (2023). Always provide context: The effects of code context on programming error message enhancement. In *Proceedings of the ACM Conference on Global Computing Education Vol 1*.

Santos, E. A. et al. (2024). Not the silver bullet: LLM-enhanced programming error messages are ineffective in practice. In *Proceedings of the 2024 Conference on United Kingdom and Ireland Computing Education Research*.

Santos, E. A. et al. (2025). It's dangerous to prompt alone! Exploring how fine-tuning GPT-4o affects novices' programming error resolution. *IEEE Access*.

Scholl, A. et al. (2024). How novice programmers use and experience ChatGPT when solving programming exercises in an introductory course. In *2024 IEEE Frontiers in Education Conference (FIE)*.

Shah, A. et al. (2025). Students' use of GitHub Copilot for working with large code bases. In *Proceedings of the 56th ACM Technical Symposium on Computer Science Education V. 1*.

Sheese, B. et al. (2023). Patterns of student help-seeking when using a large language model-powered programming assistant. In *Proceedings of the 26th Australasian Computing Education Conference*.

Shihab, M. I. H. et al. (2025). The effects of GitHub Copilot on computing students' programming effectiveness, efficiency, and processes in brownfield coding tasks. In *Proceedings of the 2025 ACM Conference on International Computing Education Research V.1*.

Simmons, A. et al. (2024). AI generated code plagiarism detection in computer science courses: A literature mapping. In *2024 IEEE Frontiers in Education Conference (FIE)*.

Song, T. et al. (2024). A high-quality generation approach for educational programming projects using LLM. *IEEE Transactions on Learning Technologies*.

Sun, D. et al. (2024). Would ChatGPT-facilitated programming mode impact college students' programming behaviors, performances, and perceptions? An empirical study. *International Journal of Educational Technology in Higher Education*.

Taylor, A. et al. (2023). dcc --help: Transforming the role of the compiler by generating context-aware error explanations with large language models. In *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1*.

Vysocan, J. et al. (2026). Detection of AI-generated source code in programming assignments. In *2026 IEEE 30th International Conference on Intelligent Engineering Systems (INES)*.

Wang, Z. et al. (2025). ChatGPT-enhanced self-regulated learning in programming education: Impacts on motivation, self-efficacy, and learning outcomes. *Interactive Learning Environments*.

Xu, J. et al. (2026). Empirical insights into the influence of ChatGPT-facilitated programming on primary learners' performances: A mixed methods approach. *International Journal of STEM Education*.

Xue, Y. et al. (2024). Does ChatGPT help with introductory programming? An experiment of students using ChatGPT in CS1. In *2024 IEEE/ACM 46th International Conference on Software Engineering: Software Engineering Education and Training (ICSE-SEET)*.

Yang, T.-C. et al. (2025). The effectiveness of ChatGPT in assisting high school students in programming learning: Evidence from a quasi-experimental research. *Interactive Learning Environments*.

Yılmaz, R. et al. (2023a). Augmented intelligence in programming learning: Examining student views on the use of ChatGPT for programming learning. *Computers in Human Behavior: Artificial Humans*.

Yılmaz, R. et al. (2023b). The effect of generative artificial intelligence (AI)-based tool use on students' computational thinking skills, programming self-efficacy and motivation. *Computers and Education: Artificial Intelligence*.

Zhu, M. et al. (2025). A systematic review of research on large language models for computer programming education. *arXiv*.