



**GENERATIVE AI-ASSISTED LEARNING IN SECONDARY EDUCATION: A
SYSTEMATIC REVIEW OF EFFECTS ON LEARNING PERFORMANCE,
KNOWLEDGE RETENTION, AND HIGHER-ORDER THINKING IN ADOLESCENTS
AGED 12–18**

Mokhigul Isanova

*MSc Business Intelligence and Analytics,
Westminster International University in Tashkent, Uzbekistan*

Sokhibjamol Boeva

Computer Science, Lecturer, Central Asia University, Tashkent, Uzbekistan

*Corresponding authors: Mokhigul Isanova, misanova@students.wiut.uz,
Sokhibjamol Boeva, s.boeva@centralasian.uz*

ABSTRACT:

ARTICLE INFO:

Received: 06.10.2026

Revised: 07.10.2026

Accepted: 08.10.2026

KEY WORDS:

*generative artificial
intelligence; ChatGPT;
secondary education;
adolescents; learning
performance;
knowledge retention;
higher-order thinking;
systematic review.*

*Generative artificial intelligence (GenAI) tools
such as ChatGPT entered secondary classrooms
within months of their public release in November
2022. Evidence on how these tools affect
adolescents' learning remains scattered, and existing
syntheses are dominated by university samples or
pool K-12 and higher education together.*

Purpose: This systematic review synthesises empirical evidence on the effects of GenAI-assisted learning on the learning performance, knowledge retention and higher-order thinking of secondary school students aged 12–18.

Sample: Thirty-seven empirical studies published between January 2023 and September 2026 were included. Records were identified in ERIC, IEEE Xplore and Google Scholar.

Design and methods: The review followed the PRISMA 2020 statement. Of 616 identified records, 594 database records and 15 Google Scholar records were screened, 91 were assessed for eligibility, and 37 met the inclusion criteria. Studies were coded for context, design, GenAI



tool and outcomes. Findings were synthesised narratively, supported by direction-of-effect vote counting.

Results: Research is recent (25 of 37 studies appeared in 2025) and concentrated in China and Taiwan (17 studies) and in programming classes (12 studies). Of 31 studies measuring learning performance, 23 reported positive effects, 4 reported no difference, 2 reported mixed results and 2 reported negative effects. Only 2 studies measured knowledge retention with a delayed test. Of 21 studies measuring higher-order thinking, 14 reported gains, mostly in computational thinking and critical thinking. Positive findings came mainly from single-group and non-randomised designs; none of the three randomised trials found a GenAI advantage on domain knowledge. Studies that compared guided with unguided GenAI use found benefits only under teacher supervision or structured scaffolding.

Introduction. Models that generate new text, graphics, code, or other content in response to natural language prompts are referred to as generative artificial intelligence (GenAI). This feature became publicly accessible in November 2022 because of large language models (LLMs) like ChatGPT, Gemini, and Copilot. Teenagers swiftly embraced these tools, frequently prior to the establishment of school policies regarding their use (Kasneci et al., 2023; Miao & Holmes, 2023). Nowadays, secondary educators must decide if using GenAI in the classroom benefits or hinders pupils between the ages of 12 and 18. This is a practical topic with little supporting data.

The question matters because adolescence is a period in which foundational knowledge is consolidated and higher-order thinking is still developing. A tool that completes tasks for students may raise the quality of their work while lowering the cognitive effort that produces durable learning. Cognitive psychology describes this risk as cognitive offloading, in which external aids reduce the internal processing that supports memory (Risko & Gilbert, 2016). It also suggests that effortful, "desirable" difficulties during practice improve long-term retention (Bjork & Bjork, 2011). A field experiment in high school mathematics found that unrestricted access to GPT-4 improved practice performance but harmed later test performance, whereas a tutor-style version with guardrails largely avoided this harm (Bastani et al., 2025). These findings make it essential to separate immediate performance from retention and from higher-order thinking.

Learning outcomes examined in this review

This review distinguishes three outcome domains. **Learning performance** refers to achievement on tests, assignments, projects or skill assessments measured during or immediately after an intervention. **Knowledge retention** refers to the persistence of learning, measured with a delayed post-test administered some time after the intervention ends. **Higher-order thinking (HOT)** refers to the analysing, evaluating and creating levels of the revised Bloom's taxonomy (Anderson & Krathwohl, 2001). In this review HOT includes critical thinking, problem solving, creativity, argumentation, scientific reasoning and computational thinking.



Prior reviews and the gap addressed

Several recent syntheses report positive average effects of GenAI on learning. A meta-analysis of 49 studies across K-12 and higher education found large effects on achievement ($g = 0.857$) and motivation (Liu et al., 2025). A meta-analysis of 19 studies found a large overall effect on academic performance ($g = 0.683$). It also found a striking moderator: students with teacher support gained substantially ($g = 1.426$), whereas students without it barely gained ($g = 0.077$) (Gu & Yan, 2025). Other reviews report moderate effects of GenAI conversational agents on cognitive and non-cognitive skills (Suo et al., 2026) and on computational thinking (Xu et al., 2026).

Three limitations of this literature motivate the present review. First, most syntheses pool school and university learners, even though adolescents differ from adults in prior knowledge, self-regulation and AI literacy. A meta-analysis of ChatGPT studies from November 2022 to December 2024 reported a larger effect for middle and high school students ($g = 0.928$) than for undergraduates ($g = 0.538$), but the difference was not significant (Doo & Park, 2026). Second, school-focused reviews either cover studies only up to 2024 (Zhang et al., 2026) or focus on a single subject, such as middle school science (Gunsaldi et al., 2025). Third, and most importantly, existing reviews rarely distinguish immediate performance from retention. They therefore cannot say whether GenAI-supported gains last.

This review addresses these gaps. It focuses exclusively on adolescents aged 12–18, includes studies published up to September 2026, and treats learning performance, knowledge retention and HOT as separate outcome domains.

Research questions

The purpose of this systematic review is to synthesise empirical evidence published since 2023 on the effects of GenAI-assisted learning among secondary school students aged 12–18. The review seeks to answer three research questions:

RQ1: What effects does generative AI-assisted learning have on the learning performance of secondary school students aged 12–18, and how do these effects vary by age group, subject area and instructional design?

RQ2: What evidence exists regarding the effects of generative AI-assisted learning on knowledge retention among secondary school students aged 12–18, particularly in studies employing delayed post-tests?

RQ3: How does generative AI-assisted learning influence the higher-order thinking skills of secondary school students aged 12–18, including critical thinking, problem solving and creativity?

The characteristics of the included studies are reported descriptively at the start of the Results section to contextualise the answers to these questions.



METHODOLOGY

This review followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 statement (Page et al., 2021). We formulated a search strategy, set inclusion and exclusion criteria, screened records against them, coded the included studies and synthesised their findings. The approach mirrors recent PRISMA-based reviews of K-12 education research (e.g., Kidd et al., 2026).

Inclusion criteria

- Empirical studies that implemented a GenAI tool (e.g., ChatGPT, Gemini, Copilot, a custom LLM chatbot or a GenAI image generator) as part of students' learning.
- Participants were secondary school students aged 12–18, or enrolled in grades 6–12 (including middle, junior high, high and secondary vocational schools).
- At least one outcome measured learning performance, knowledge retention or higher-order thinking.
- An experimental, quasi-experimental or pre-test/post-test design.
- Peer-reviewed journal articles or conference papers published in English between January 2023 and September 2026.

Exclusion criteria

- Studies of primary school pupils, university students, pre-service or in-service teachers, or mixed samples whose secondary results were not reported separately.
- Correlational, cross-sectional, case-study or qualitative-only designs.
- Studies measuring only perceptions, engagement, motivation, usability or emotional outcomes.
- Studies in which the AI tool could not be identified as generative, or in which GenAI was not part of the manipulated condition.
- Conceptual papers, frameworks, protocols, reviews, meta-analyses, theses and preprints.

Search strategy

Search terms were built from three concept blocks: generative AI (e.g., "generative AI", ChatGPT, "large language model"), secondary-school population (e.g., "secondary school", "high school", "middle school", adolescents, K-12) and outcomes (e.g., "learning outcomes", achievement, retention, "critical thinking", "problem solving"). Terms within a block were combined with OR and blocks were combined with AND. ERIC was chosen for its education focus, IEEE Xplore for educational technology research, and Google Scholar to capture studies in journals indexed by neither. Because Google Scholar returns thousands of results ranked by relevance, one query was run per research question and the first results page of each query was screened. All searches were run on 29 September 2026.



Table 1. Search sources and records identified.

Source	Query focus	Filters applied	Records
ERIC	GenAI AND secondary population AND learning outcomes	Publication date since 158 2022	
IEEE Xplore	Same three blocks, searched in All Metadata	None at export (all content types exported)	441
Google Scholar, query 1	GenAI AND secondary AND performance or achievement	Since 2022, first results	6
Google Scholar, query 2	GenAI AND secondary AND retention or delayed post-test	Since 2022, first results	5
Google Scholar, query 3	GenAI AND secondary AND critical thinking, HOT or problem solving	Since 2022, first results	6
Total			616

Screening and selection

Records from ERIC (NBIB export) and IEEE Xplore (CSV export) were merged in a spreadsheet. Duplicates were identified by normalised title and removed ($n = 4$), together with one record published before 2023. Titles and abstracts of the remaining 594 records were screened against the criteria. The 76 records that passed this stage were assessed for eligibility using their full abstract, author keywords and database descriptors (e.g., ERIC education-level descriptors). Google Scholar records were checked against the database records (two duplicates) and the remaining 15 were assessed in the same way. Reasons for exclusion were recorded at the eligibility stage. The selection process is summarised in Figure 1.



Of 616 records identified, 37 studies met the inclusion criteria

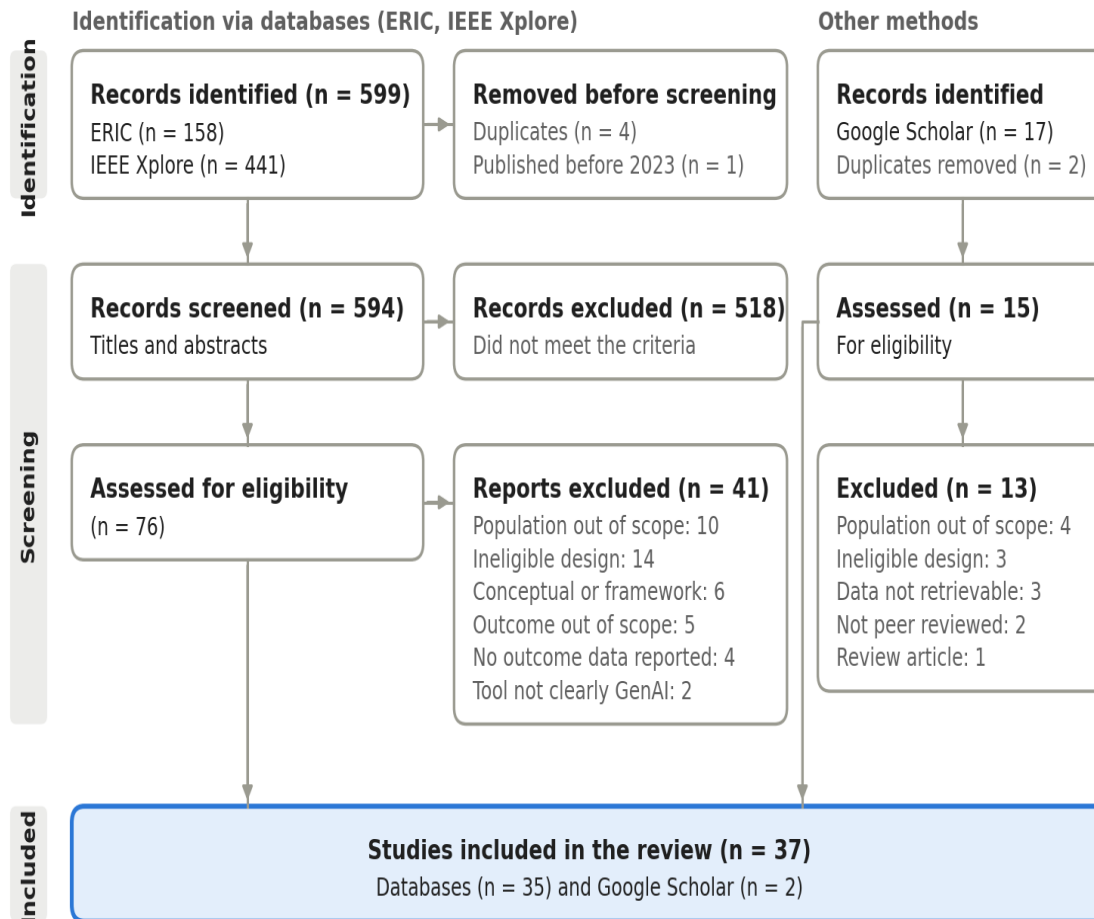


Figure 1. PRISMA 2020 flow diagram of study identification, screening and inclusion.

Data extraction and coding

Data were extracted into a coding sheet with categories derived from the research questions (Table 2). Outcome findings were coded by direction of effect for each outcome domain: **positive** (a statistically significant advantage for the GenAI condition or a significant pre-post gain), **null** (no significant difference), **negative** (a significant disadvantage for the GenAI condition) or **mixed** (advantages on some measures or subgroups and none or disadvantages on others). When the retrieved record did not report a detail, it was coded as not reported (n.r.).



Table 2. Coding scheme.

Category	Codes
Publication year	2023, 2024, 2025, 2026
Country or region	Country as stated or as indicated by the authors' affiliation; grouped by region
School level	Lower secondary (grades 6–9, about ages 12–15); upper secondary (grades 10–12, including vocational, about ages 15–18); unspecified
Subject area	Programming and computing; science and STEM; language and literature; mathematics; AI literacy; social sciences and humanities; cross-curricular
GenAI tool	General-purpose chatbot; custom or purpose-built GenAI system; multiple GenAI tools; image generator
Study design	Randomised controlled trial; controlled non-randomised (quasi-experimental or comparative); single-group pre-post
Outcome domain	Learning performance (RQ1); knowledge retention with a delayed test (RQ2); higher-order thinking (RQ3)
Direction of effect	Positive; null; negative; mixed

Synthesis

The included studies used different outcome measures, and many records did not report the statistics needed to compute effect sizes. A meta-analysis was therefore not appropriate. Findings were synthesised narratively for each research question, supported by vote counting based on the direction of effect (McKenzie & Brennan, 2019). Direction of effect was also cross-tabulated with study design and school level to examine whether findings depended on the rigour of the evidence.

Results

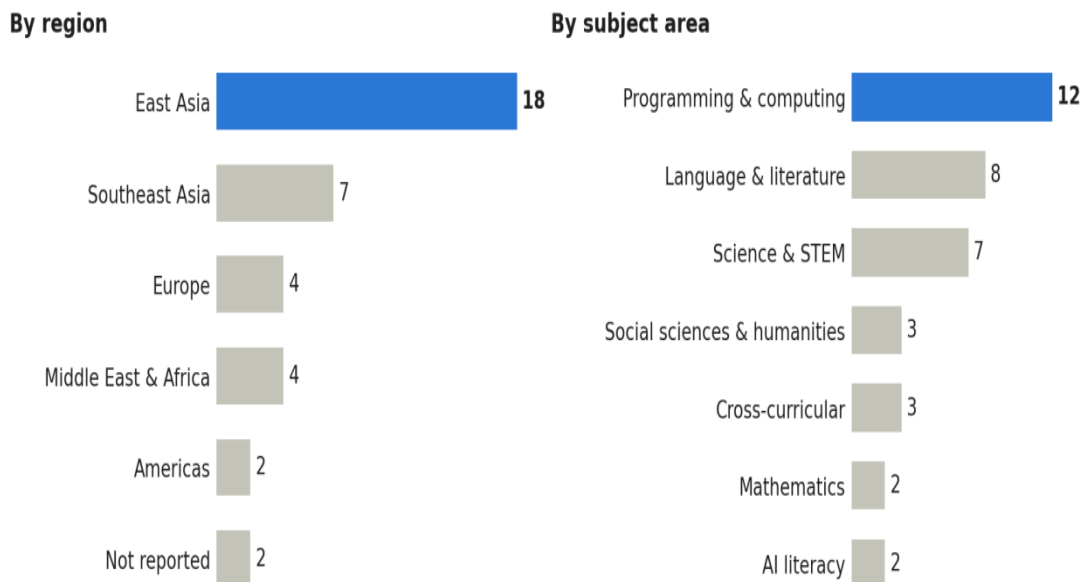
Characteristics of included studies

Thirty-seven studies met the inclusion criteria (Table 3). The evidence base is very recent: none appeared in 2023, 2 appeared in 2024, 25 in 2025 and 10 in the first nine months of 2026. Classroom intervention studies with adolescents therefore took more than a year to follow the public release of ChatGPT.



Evidence is concentrated in East Asia and in programming classes

Number of included studies (n = 37); bars share one scale



East Asia: China 13, Taiwan 4, Korea 1

Figure 2. Distribution of the 37 included studies by region and by subject area.

The studies were conducted in 17 identified countries, but the distribution was uneven (Figure 2). China (13 studies) and Taiwan (4) together contributed almost half of the evidence, and 25 of the 37 studies came from East and Southeast Asia. Only four studies came from Europe, four from the Middle East and Africa, and two from the Americas. Programming and computing was the most frequent subject (12 studies), followed by language and literature (8) and science and STEM (7). Mathematics, the core subject of most secondary curricula, appeared in only two studies.

Upper secondary students (grades 10–12, including secondary vocational schools) were studied more often (21 studies) than lower secondary students (grades 6–9; 14 studies); two studies did not report the level. General-purpose chatbots were the most common tool (24 studies; ChatGPT was named in 15). Ten studies used a custom or purpose-built GenAI system, such as a dialogic feedback agent, a retrieval-augmented tutor or a multi-agent programming assistant. Most studies used controlled non-randomised designs (27), seven used a single group with a pre-test and post-test, and only three were randomised controlled trials (RCTs). Sample sizes were reported for 27 studies and ranged from 13 to 371 students (median 75; 2,380 students in total).



Table 3. Included studies (n = 37). Direction of effect: + positive, 0 null, – negative, ± mixed, — not measured; n.r. = not reported in the retrieved record.

Study	Country	Level (grade/age)	N	Subject	GenAI tool	Design	Per f.	Ret .	HO T
Aissa & Mhiri (2025)	Tunisia	Upper (secondary)	n.r.	Programmi ng	ChatGPT (general)	Quasi- experiment al	+	—	+
Aldous et al. (2025)	Qatar	Lower (middle school)	18	Science and STEM	Custom LLM chatbot (RAG)	One-group pre-post	+	—	—
Bear et al. (2026)	Mexico	Upper (ages 16–18)	68	Language (French FL)	Multiple GenAI tools	Pre-post, mixed methods	+	—	+
Borjigin & Shen (2025)	Singapor e	Lower (middle school)	n.r.	Science and STEM	ChatGPT (general)	Quasi- experiment al, mixed methods	+	—	+
Chrysopoulos & Moussiades (2025)	Greece	Upper (high school)	18	Programmi ng	Custom LLM tutor	Quasi- experiment al (pilot)	+	—	—
Chu et al. (2025)	Taiwan	Upper (high school)	60	Language (literature)	GenAI image generator	Quasi- experiment al	+	—	—
Cui & Li (2026)	China	Upper (secondary vocational)	75	Cross- curricular	GenAI assistant (general)	Quasi- experiment al	—	—	±
Farhane (2025)	Morocco	n.r. (secondary)	n.r.	Language (EFL writing)	ChatGPT (general)	Quasi- experiment al, delayed post-test	+	+	—
Fütterer et al. (2026)	Germany	Lower (grades 7–9)	371	Science & language	Custom GenAI (SRL prompts)	RCT	0	—	—
Gong et al. (2025)	China	Upper (high school)	172	Programmi ng	GenAI dialogic	Quasi- experiment	+	—	+



Study	Country	Level (grade/age)	N	Subject	GenAI tool	Design	Per f.	Ret .	HO T
					feedback	al			
Guo Q. et al. (2025)	China	Lower (middle school)	n.r.	Science and STEM	LLM (general)	Quasi- experiment al	+	—	+
Guo R. et al. (2025)	China	Upper (high school)	83	Programmi ng	GenAI assistant (general)	Pre-post, mixed methods	—	—	+
Guo (2025)	Z. China	Upper (high school)	100	Language (argumenta tive writing)	GenAI assistant (general)	One-group pre-post	+	—	±
Hariyanti al. (2026)	et Indonesia	Upper (grade 11, vocational)	66	Computing /IT	ChatGPT (general)	Quasi- experiment al	+	—	+
Huang Chen (2025)	& Taiwan	Lower (grade 7)	160	Science and STEM	Custom GenAI chatbot	RCT (2×2 factorial)	0	—	—
Huang L. al. (2026)	et China	Upper (high school)	80	Science and STEM	Custom GenAI chatbot + lab tools	Quasi- experiment al	—	—	+
Hussein et al. (2026)	n.r.	Upper (grade 12)	61	Programmi ng	ChatGPT (general)	Quasi- experiment al	—	—	±
Li et al. (2025)	n.r.	Upper (grade 11)	93	Social studies (SDGs)	ChatGPT (general)	Quasi- experiment al	—	—	+
Li X. et al. (2025)	Canada	Upper (high school)	n.r.	Language (text analysis)	ChatGPT + prompt framework	Experimen tal (details n.r.)	+	—	±
Li H.-J. et al. (2025)	China	Upper (secondary vocational)	92	Programmi ng	GenAI assistant (general)	Quasi- experiment al	+	—	+
Liu et al. (2026)	Thailand	Lower (junior high)	64	Language (Chinese reading)	ChatGPT (general)	Quasi- experiment al	+	—	—



Study	Country	Level (grade/age)	N	Subject	GenAI tool	Design	Per f.	Ret .	HO T
Mai (2026)	Vietnam	Upper (ages 16–17)	42	Language (EFL writing)	ChatGPT (general)	Quasi-experimental	±	—	—
Mottl & Musilek (2024)	Czech Republic	Upper (upper secondary)	n.r.	History	ChatGPT (general)	Comparative, 1-month delayed test	0	0	—
Nam et al. (2026)	South Korea	Lower (grade 7)	50	Language (EFL writing)	ChatGPT (general)	Comparative pre-post	0	—	—
Njoku & Njoku (2025)	Nigeria	n.r. (secondary)	n.r.	Mathematics	ChatGPT (general)	Experimental (details n.r.)	+	—	—
Öner (2025)	Türkiye	Lower (grade 6)	13	Mathematics	ChatGPT (general)	One-group pre-post	+	—	—
Panchu & Suamuang (2025)	Thailand	Upper (upper secondary)	26	AI literacy	ChatGPT (general)	One-group pre-post	+	—	+
Pianon et al. (2025)	Thailand	Lower (grade 9)	18	AI literacy	Multiple GenAI tools	One-group pre-post	+	—	—
Rahadi et al. (2025)	Indonesia	Upper (vocational high school)	n.r.	Vocational subject	ChatGPT (general)	Quasi-experimental	+	—	—
Sui et al. (2026)	Taiwan	Lower (grade 8)	94	Science and STEM	Custom GenAI scaffolds	RCT	—	—	±
Tang et al. (2025)	China	Lower (middle school)	131	Computing /IT	GenAI assistant (general)	Quasi-experimental	±	—	—
Tian et al. (2025)	China	Upper (senior high)	60	Programming	GenAI assistant (general)	Quasi-experimental	+	—	+
Yang et al. (2025)	Taiwan	Upper (high school)	153	Programming	ChatGPT (general)	Quasi-experiment	—	—	—



Study	Country	Level (grade/age)	N	Subject	GenAI tool	Design	Per f.	Ret .	HO T
al									
Zhang H. & Li (2024)	China	Lower (junior high)	n.r.	Programmi ng	GenAI assistant (general)	Quasi- experiment al	+	—	+
Zhang & Zhang (2025)	China	Lower (junior high)	n.r.	Cross- curricular PBL	LLM (general)	Quasi- experiment al	+	—	—
Zhao et al. (2025)	China	Upper (grade 10)	85	Programmi ng	Custom multi- agent GenAI	Quasi- experiment al	+	—	+
Zhu et al. (2025)	China	Lower (middle school)	127	Humanities /assignmen ts	Custom LLM chatbot	Quasi- experiment al	—	—	0

RQ1: Effects on learning performance

Most studies reported that GenAI improved learning performance, but the pattern weakens as the evidence becomes more rigorous. Thirty-one studies measured learning performance. Of these, 23 reported positive effects, 4 found no difference, 2 reported mixed results and 2 reported negative effects (Figure 3).

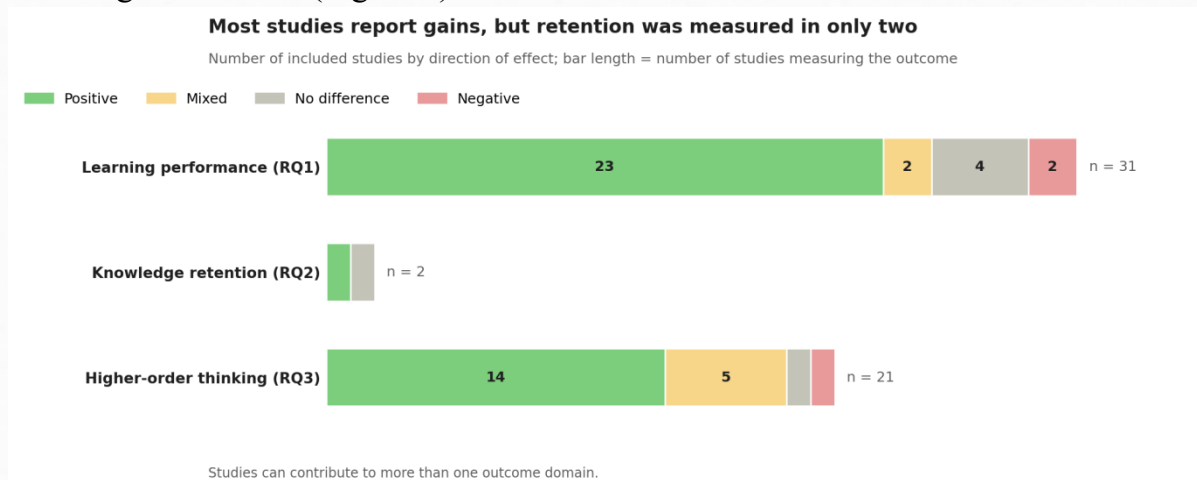


Figure 3. Direction of reported effects for each outcome domain. Positive findings dominate performance and higher-order thinking, while retention was examined in only two studies.

By subject area. Programming produced the most consistent gains. Eight of the ten programming or computing studies that measured performance reported positive effects, including studies of staged dialogic feedback (Gong et al., 2025), multi-agent collaborative



programming (Zhao et al., 2025), a progressive guided teaching model (Tian et al., 2025) and a hint-based tutor (Chrysopoulos & Moussiades, 2025). The two exceptions are instructive. High school students who used ChatGPT in regular programming lessons reported lower flow and self-efficacy and achieved lower scores than peers taught conventionally (Yang et al., 2025). In an information technology course, GenAI without teacher supervision did not improve knowledge mastery over traditional instruction (Tang et al., 2025). In language learning, six of eight studies reported gains in reading, writing or communicative competence (e.g., Liu et al., 2026; Chu et al., 2025; Bear et al., 2026). ChatGPT feedback and a rule-based AI digital textbook improved Korean seventh graders' writing equally (Nam et al., 2026). Vietnamese students using ChatGPT improved fluency and structure faster, but their essays became formulaic (Mai, 2026).

Performance with the tool is not the same as learning. Middle school students who completed assignments with an LLM chatbot submitted more work and obtained higher assignment scores than control students. Yet they scored significantly lower on a subsequent knowledge test (Zhu et al., 2025). This study is the clearest example in the review of GenAI raising the quality of products while lowering what students learned.

By instructional design. Positive studies typically embedded GenAI in an explicit pedagogical model that limited answer-seeking: step-by-step hints rather than solutions, reflection prompts or teacher-structured tasks (e.g., Borjigin & Shen, 2025; Hariyanti et al., 2026; Zhang & Zhang, 2025). Two studies compared GenAI conditions directly. In the first, teacher supervision of GenAI use significantly increased engagement and knowledge mastery compared with unsupervised use (Tang et al., 2025). In the second, a GenAI chatbot used alone in a science game did not improve learning and produced the most fragmented inquiry behaviour; instructional videos drove the learning gains (Huang & Chen, 2025). Conversely, a large RCT found that GenAI tools purpose-built to support self-regulated learning were no better than standard ChatGPT for domain knowledge (Fütterer et al., 2026).

By school level. Positive effects were reported in 13 of 16 upper secondary studies and 8 of 13 lower secondary studies. This difference is confounded with design, because both RCTs that measured performance involved lower secondary students.

By study design. The direction of effect depended strongly on study design (Figure 4). All six single-group pre-post studies reported positive effects. Among 23 controlled non-randomised studies, 17 did so. Neither RCT that measured learning performance found a GenAI advantage (Fütterer et al., 2026; Huang & Chen, 2025).

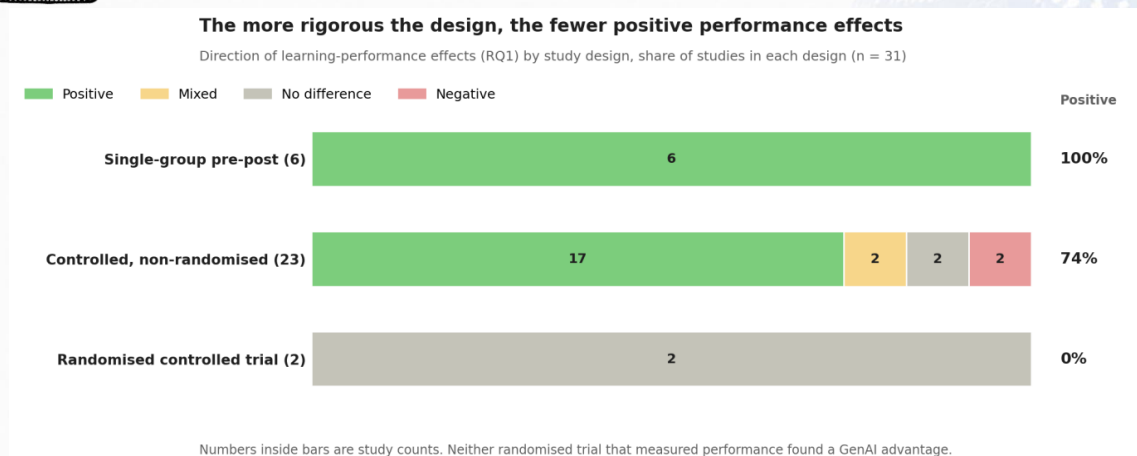


Figure 4. Direction of learning-performance effects by study design. The share of positive findings falls from 100% in single-group studies to 74% in controlled non-randomised studies and 0% in randomised trials.

RQ2: Effects on knowledge retention

Evidence on retention is almost absent: only 2 of the 37 studies administered a delayed post-test. In an upper secondary history experiment, students learned new content from text, video or ChatGPT and were retested about one month later (Mottl & Musílek, 2024). Video produced the best immediate acquisition but the least durable retention. Text and ChatGPT both produced good results, with no clear advantage for ChatGPT. In a quasi-experiment with Moroccan secondary students, ChatGPT served as a formative feedback tool for paragraph writing (Farhane, 2025). Pre-test, post-test and delayed post-test assessments indicated statistically significant effects, and the delayed test was used specifically to examine whether writing gains were retained.

Two further studies offer indirect evidence on durability. In an RCT with eighth graders, cognitive GenAI scaffolding improved argumentation on the first task, but the gains did not persist across later tasks. Metacognitive scaffolding, in contrast, produced continued improvement by the final task (Sui et al., 2026). The finding by Zhu et al. (2025) that GenAI-assisted assignment scores did not translate into test performance also suggests that immediate gains may overstate learning.

Several studies with delayed tests were identified but excluded because they did not meet the eligibility criteria. They include an AI-versus-teacher feedback study with fifth-grade students that measured writing achievement and retention (Baz & Hasirci Aksoy, 2025), and a correlational survey of self-reported retention among grade 11 students (Lanes et al., 2026). Retention is thus being studied, but not yet in rigorous designs with adolescents aged 12–18.

RQ3: Effects on higher-order thinking

Twenty-one studies measured at least one higher-order thinking skill. Fourteen reported positive effects, five reported mixed effects, one found no effect and one reported a negative effect (Figure 3). The results differed sharply by type of skill.



Computational thinking and structured problem solving showed the most consistent gains. All five studies measuring computational thinking or programming problem solving reported improvements (Guo Q. et al., 2025; Guo R. et al., 2025; Tian et al., 2025; Zhang H. & Li, 2024; Zhao et al., 2025). Related gains were reported for design thinking in project-based learning (Hariyanti et al., 2026), problem solving in STEM projects (Borjigin & Shen, 2025) and logical reasoning in programming (Aissa & Mhiri, 2025). Scientific reasoning, argumentation and questioning also improved when a GenAI chatbot was combined with digital laboratories in physics (Huang L. et al., 2026).

Critical thinking produced more variable results. ChatGPT increased high school students' critical thinking tendencies compared with a search-engine condition, and the approach worked best for students who already had high critical thinking levels (Li et al., 2025). Dialogic GenAI feedback improved critical thinking most when it was introduced during debugging rather than code writing (Gong et al., 2025). In a 12th-grade programming course, critical thinking improved only when students collaborated with ChatGPT individually rather than in teams (Hussein et al., 2026). A ChatGPT prompt framework improved several analytical skills, but gains in critical thinking were less pronounced (Li X. et al., 2025). In argumentative writing, GenAI improved language and style but had inconsistent effects on critical thinking, depending on students' baseline thinking and AI skills (Guo Z., 2025). The only negative finding came from EFL argumentative writing: ChatGPT users showed over-reliance and less critical engagement, while the control group improved originality and argumentation (Mai, 2026).

Creativity improved only under specific conditions. Among low-achieving secondary vocational students, GenAI alone did not improve creativity. All four creativity dimensions improved only when the teacher provided structured support alongside GenAI (Cui & Li, 2026). Hussein et al. (2026) found no creativity gains, whereas Li H.-J. et al. (2025) reported gains in the 5C competencies, with creativity gains most evident among students with diverging and assimilating learning styles. Finally, the one study that tested transfer found no difference between chatbot-assisted and control students in their ability to apply knowledge in new contexts (Zhu et al., 2025).

DISCUSSION

This review examined 37 empirical studies on GenAI-assisted learning among adolescents aged 12–18. Three main findings emerge. The evidence base is young and geographically narrow. Reported effects on immediate performance and computational thinking are mostly positive, but they weaken as research designs become more rigorous. And almost nothing is yet known about whether GenAI-supported learning is retained.

Immediate performance: promising but fragile

The high proportion of positive findings for learning performance (23 of 31 studies) is consistent with meta-analyses reporting moderate-to-large average effects (Doo & Park, 2026; Gu & Yan, 2025; Liu et al., 2025). The pattern across designs, however, calls for caution. Every



single-group study reported gains, whereas neither RCT that measured domain knowledge found a GenAI advantage. Single-group pre-post designs cannot separate the effect of GenAI from ordinary instruction, maturation or practice with the test. Non-randomised comparisons of intact classes are vulnerable to teacher and class effects, and novelty effects may inflate results in short interventions. Many studies were small, and ten records did not report sample size. Publication bias is also plausible in a field where positive results are easier to publish.

The studies that reported null or negative results share an important feature. Students used a general-purpose tool with little structure (Yang et al., 2025; Mai, 2026), GenAI was compared with a strong alternative scaffold (Huang & Chen, 2025; Nam et al., 2026), or learning was measured separately from task performance (Zhu et al., 2025). The last point is critical. GenAI reliably improves what students produce while the tool is available. Whether it improves what students can do without the tool is a different question, and the few studies that separated the two found no gain or a loss. This pattern mirrors the high school mathematics field experiment by Bastani et al. (2025) and is consistent with the cognitive offloading account (Risko & Gilbert, 2016).

Pedagogy matters more than the tool

Across outcome domains, the benefits of GenAI depended on how it was embedded in teaching. Teacher supervision (Tang et al., 2025) and teacher-provided structure (Cui & Li, 2026) were the conditions under which GenAI produced gains in knowledge and creativity. This finding agrees with the meta-analytic contrast between students with teacher support ($g = 1.426$) and without it ($g = 0.077$) reported by Gu and Yan (2025). The most successful designs limited direct answer-giving through hints, staged feedback, reflection prompts or progressive guidance (Chrysopoulos & Moussiades, 2025; Gong et al., 2025; Tian et al., 2025). The type of scaffold also mattered: metacognitive scaffolding sustained argumentation gains where cognitive scaffolding did not (Sui et al., 2026). By contrast, simply building a purpose-designed tool was not enough (Fütterer et al., 2026). For adolescents, whose self-regulation is still developing, the teacher's orchestration of GenAI appears to be the active ingredient.

Higher-order thinking: structured skills respond, open-ended skills are harder

The HOT findings divide along the structure of the skill. Computational thinking and programming problem solving improved in every study that measured them. These skills involve decomposition, debugging and iteration, which map well onto the dialogic and code-explaining strengths of LLMs. Critical thinking and creativity showed mixed results, and the only negative HOT finding concerned argumentative writing. Here GenAI can supply fluent arguments that replace, rather than support, students' own reasoning (Mai, 2026). Several studies suggest that GenAI amplifies existing dispositions: students with high critical thinking levels benefited most (Li et al., 2025), and gains depended on baseline thinking and AI skills (Guo Z., 2025). Without deliberate design, GenAI may therefore widen gaps between stronger and weaker thinkers.



The retention gap

The most important gap concerns retention. Only two studies used a delayed post-test, and neither provides strong evidence that GenAI-supported learning lasts longer than conventional learning. This is a serious limitation for a technology whose main risk, according to learning science, is that it reduces the effortful processing that underpins long-term memory (Bjork & Bjork, 2011). Until retention is measured routinely, positive immediate effects should be interpreted as short-term performance gains rather than as evidence of durable learning.

Implications for practice

For teachers and school leaders, the evidence supports guided rather than open use of GenAI. Teachers can configure or prompt tools to give hints and questions rather than complete answers, and can stage GenAI support at points such as debugging and revision. Students should be required to demonstrate learning without the tool, through in-class tests, oral explanations or delayed quizzes. Reflection and metacognitive prompts should be built into GenAI activities, and teachers should monitor students with weaker self-regulation more closely. For policymakers, the concentration of evidence in East Asia and in programming classes means that findings should not be generalised uncritically to other subjects and contexts, including Central Asian school systems, where no eligible study was found.

Limitations

This review has several limitations. First, only three sources were searched, and Google Scholar screening was limited to the first results page of each query. Relevant studies indexed only in Scopus, Web of Science or regional databases may have been missed. Second, eligibility and coding were based on titles, abstracts, author keywords and database descriptors rather than full texts. Some design details, sample sizes and effect directions may therefore be incomplete, and several studies could not be assessed because their outcome data were not available in the retrieved record. Third, screening and coding were performed by one reviewer, so inter-rater reliability could not be established. Fourth, restricting inclusion to English-language, peer-reviewed publications excluded preprints, theses and non-English studies, some of which measured retention. Fifth, vote counting ignores effect size and sample size and cannot estimate the magnitude of effects. Finally, risk of bias was not assessed with a formal tool, although the cross-tabulation by design partly addresses evidence quality.

Future research

Four priorities follow from this review. Researchers should (1) use randomised or cluster-randomised designs with active control conditions; (2) include delayed post-tests of at least several weeks to measure retention; (3) directly compare guided and unguided GenAI use, including tool-free assessments of learning; and (4) extend research to mathematics, the humanities, lower secondary students and under-represented regions, including Central Asia. Reporting of sample sizes, intervention duration and effect sizes should also improve so that future meta-analyses of adolescent learners become possible.



Conclusion. This systematic review synthesised 37 empirical studies of GenAI-assisted learning among secondary school students aged 12–18, published between 2023 and 2026. Most studies reported that GenAI improved immediate learning performance and structured higher-order skills such as computational thinking, particularly when it was embedded in teacher-guided, scaffolded designs. The evidence is nevertheless young, concentrated in East Asia and programming classes, and dominated by non-randomised designs. The strongest designs found no advantage on domain knowledge, and only two studies measured retention. GenAI should therefore be treated as a promising but unproven support for adolescent learning. Its value depends less on the tool itself than on the pedagogy around it. The next generation of research should test whether guided GenAI use produces learning that students retain and can demonstrate without the tool.

References

1. Aissa, N. and Mhiri, H.S. (2025) 'Assessing the effectiveness of ChatGPT for enhancing programming skills among second-year computer science students in Tunisia', in 2025 11th International Conference on Control, Decision and Information Technologies (CoDIT). IEEE, pp. 1–6. Available at: <https://doi.org/10.1109/CoDIT66093.2025.11321829>.
2. Aldous, K.K., Salminen, J., Jung, S.-g., Azem, J.Y., Medina, J., Khan, S.M., Alabed, A. and Jansen, B.J. (2025) 'Investigating the usability of an educational AI chatbot by middle school teachers and students for enhanced learning', in 2025 3rd International Conference on Foundation and Large Language Models (FLLM). IEEE, pp. 142–152. Available at: <https://doi.org/10.1109/FLLM67465.2025.11390874>.
3. Anderson, L.W. and Krathwohl, D.R. (eds) (2001) A taxonomy for learning, teaching, and assessing: a revision of Bloom's taxonomy of educational objectives. New York: Longman.
4. Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakçı, Ö. and Mariman, R. (2025) 'Generative AI without guardrails can harm learning: evidence from high school mathematics', Proceedings of the National Academy of Sciences, 122(26), e2422633122. Available at: <https://doi.org/10.1073/pnas.2422633122>.
5. Baz, M.A. and Hasirci Aksoy, S. (2025) 'The effect of feedback on informative text writing: AI or teacher?', Open Praxis, 17(3), pp. 611–631.
6. Bear, M.I.N., Rodriguez, A.M. and Dominguez, C.T.S. (2026) 'AI-enhanced learning environments to foster communicative competence in French: a mixed-methods study with STEAM students', in 2026 IEEE Engineering Education World Conference (EDUNINE). IEEE, pp. 1–6. Available at: <https://doi.org/10.1109/EDUNINE62390.2026.11546761>.
7. Bjork, E.L. and Bjork, R.A. (2011) 'Making things hard on yourself, but in a good way: creating desirable difficulties to enhance learning', in Gernsbacher, M.A., Pew, R.W., Hough, L.M. and Pomerantz, J.R. (eds) Psychology and the real world: essays illustrating fundamental contributions to society. New York: Worth Publishers, pp. 56–64.



8. Borjigin, A. and Shen, Z. (2025) 'AI-powered meta-reflection for self-learning in STEM: a middle school case study on smart home energy management', in 2025 IEEE Region 10 Symposium (TENSYPMP). IEEE, pp. 1–6. Available at: <https://doi.org/10.1109/TENSYPMP63728.2025.11144984>.

9. Chrysopoulos, M. and Moussiades, L. (2025) 'Enhancing high school programming education through LLM-based hint generation', in 2025 International Workshop on Artificial Intelligence and Education (WAIE). IEEE, pp. 437–443. Available at: <https://doi.org/10.1109/WAIE67422.2025.11381301>.

10. Chu, H.-C., Hsu, C.-Y. and Wang, C.-C. (2025) 'Effects of AI-generated drawing on students' learning achievement and creativity in an ancient poetry course', Educational Technology & Society, 28(2), pp. 295–309.

11. Cui, Y. and Li, M. (2026) 'Can Gen AI improve creativity in students with low academic performance? A quasi-experimental study based on Chinese samples', SAGE Open, 16(2). Available at: <https://doi.org/10.1177/21582440261449529>.

12. Doo, M.Y. and Park, Y. (2026) 'A meta-analysis of ChatGPT's influence on learning achievement', International Review of Research in Open and Distributed Learning, 27(1), pp. 265–289.

13. Farhane, H. (2025) 'The impact of ChatGPT as a formative feedback tool on Moroccan secondary school students' paragraph writing skills: a quasi-experimental design', Journal of English Language Teaching and Applied Linguistics.

14. McKenzie, J.E. and Brennan, S.E. (2019) 'Synthesizing and presenting findings using other methods', in Higgins, J.P.T., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M.J. and Welch, V.A. (eds) Cochrane handbook for systematic reviews of interventions. 2nd edn. Chichester: Wiley, pp. 321–347.

15. Miao, F. and Holmes, W. (2023) Guidance for generative AI in education and research. Paris: UNESCO.

16. Mottl, J. and Musílek, M. (2024) 'Information technology as a tool for teaching history with a focus on artificial intelligence and audiovisual material', in 2024 47th MIPRO ICT and Electronics Convention (MIPRO). IEEE, pp. 278–282. Available at: <https://doi.org/10.1109/MIPRO60963.2024.10569216>.

17. Nam, Y., Lee, J.M. and Park, H. (2026) 'Improving L2 writing through artificial intelligence: comparing AIDT and ChatGPT feedback', English Teaching, 81(1), pp. 3–36.

18. Njoku, C. and Njoku, I.O. (2025) 'Effect of chat generative pre-trained transformer (ChatGPT) on secondary school students' academic achievement and interest in geometry', African Journal of Science, Technology and Mathematics Education.

19. Öner, A.T. (2025) 'Worked-out examples with ChatGPT-based explanations: a statistics activity', Journal of Inquiry Based Activities, 15(2), pp. 146–169.