



Generative AI in higher education: A scoping review and framework for student motivation and engagement

Edmund De Leon Evangelista ^{1*}

 0000-0002-8050-1360

Syed M. Salman Bukhari ¹

 0009-0009-3541-3124

¹ College of Technological Innovation, Zayed University, Abu Dhabi, UNITED ARAB EMIRATES

* Corresponding author: edmund.evangelista@zu.ac.ae

Citation: Evangelista, E. D. L., & Bukhari, S. M. S. (2026). Generative AI in higher education: A scoping review and framework for student motivation and engagement. *Contemporary Educational Technology*, 18(4), Article ep694. <https://doi.org/10.30935/cedtech/19499>

ARTICLE INFO

Received: 2 May 2026

Accepted: 16 Sep 2026

ABSTRACT

Generative artificial intelligence (GenAI) is increasingly influencing how students write, create, reflect, solve problems, and engage with learning in higher education. However, the evidence remains divided, particularly regarding motivation, engagement, creativity, perception, self-regulation, and psychologically informed learning design. This study presents a PRISMA-ScR-based scoping review of GenAI in higher education from a motivational and psychological perspective. The final corpus included 195 studies published between 2023 and 2025 within a specified eligibility window of January 2021 to December 2025. Records were dual screened, coded using a structured framework, and analyzed across psychological focus, thematic constructs, GenAI tools, study designs, outcomes, contextual characteristics, self-determination theory (SDT) needs, equity, and evidence strength. The screening agreement was 81.9%, with Cohen's $\kappa = 0.642$. Motivation was the most common primary focus (31.3%), followed by perception (26.7%), engagement (18.5%), creativity (15.4%), and general/other (8.2%). The literature was dominated by ChatGPT and general GenAI systems, while survey-based and observational evidence was far more common than experimental evidence, longitudinal research, or studies reporting objective performance outcomes. Among 155 empirical studies, mixed or conditional outcomes were most frequent (46.5%), compared with predominantly positive findings (21.9%). SDT related evidence was most common for competence (42.6%), followed by autonomy (29.7%) and relatedness (22.6%), although direct measurement was rare. Based on these findings, the study proposes the GenAI engagement framework, a literature informed structure that organizes five recurring components of GenAI supported learning without assuming a fixed sequence or causal pathway. The framework is not presented as a validated causal model, but as a structured basis for future research and instructional design. The findings suggest that the educational value of GenAI depends less on the technology itself than on how it is guided, evaluated, and integrated into meaningful learning activities.

Keywords: generative artificial intelligence, student motivation, learner engagement, self-determination theory, higher education, educational psychology, multimodal learning, AI-enhanced learning

INTRODUCTION

Generative artificial intelligence (GenAI) has quickly become part of everyday teaching and learning in higher education. Tools such as ChatGPT, Gemini, Claude, DALL-E, and other multimodal systems are now used for writing, feedback, brainstorming, visualization, reflection, problem solving, and creative activities (Ramesh et al., 2021; Saúde et al., 2024; Zhai, 2023). These tools can help students generate ideas, revise drafts, understand difficult concepts, and explore content in ways that respond more closely to their individual

needs. GenAI is therefore influencing not only how students complete academic tasks, but also how they participate in and experience learning.

Despite this rapid growth, the motivational and psychological effects of GenAI remain insufficiently understood. Much of the existing research focuses on adoption, perceived usefulness, satisfaction, academic writing support, and general attitudes toward ChatGPT and related systems (Bannister et al., 2023; Chan & Hu, 2023; Ivanov et al., 2024; Saúde et al., 2024). Although these studies offer useful early insights, many rely on short-term, self-reported experiences. They provide less evidence about how GenAI shapes motivation, engagement, creativity, self-regulation, and sustained participation over time. The research is also heavily concentrated on text-based conversational tools, while visual, audio, multimodal, and integrated GenAI systems have received less systematic attention in relation to learner experience and instructional design.

This gap matters because learning with GenAI is not simply a technical interaction. It is also a motivational and psychological experience. Immediate feedback may help students feel more capable, while opportunities to explore ideas at their own pace may strengthen their sense of choice and control. GenAI supported revision and reflection may also encourage deeper involvement in learning. However, these benefits are not automatic. Poorly designed or uncritical use may contribute to overreliance, reduced cognitive effort, inaccurate information, academic integrity concerns, unequal access, and weaker human interaction. Understanding GenAI in higher education therefore requires more than describing what these tools can produce. It requires careful examination of how they may support or limit meaningful learning (Kasneji et al., 2023; Sok & Heng, 2024).

Self-determination theory (SDT) offers a useful lens for examining these issues because it explains motivation through three basic psychological needs (autonomy, competence, and relatedness) (Deci & Ryan, 2012). Autonomy concerns the learner's sense of choice and ownership, competence concerns feelings of progress and effectiveness, and relatedness concerns connection and belonging within the learning environment. These needs are directly relevant to GenAI supported learning. The conversational tools may support autonomy by allowing students to direct their own inquiries and may support competence through explanation, feedback, and revision. Relatedness is more complex because interaction with an AI system alone does not create meaningful social connection. Instead, relatedness may be supported when GenAI is integrated into peer discussion, group work, teacher-student interaction, or collaborative reflection. Although some studies draw on SDT or related motivational concepts (Collie & Martin, 2024), literature still lacks a broad synthesis of how these psychological mechanisms are represented across higher education research.

To address this gap, the present study conducts a PRISMA-ScR-based scoping review (SR) of GenAI in higher education from a motivational and psychological perspective. The final corpus includes 195 studies, which are mapped according to their psychological focus, thematic constructs, GenAI tools, study designs, contextual characteristics, outcome directions, evidence strength, and representation of SDT related needs. The review also examines important gaps in literature, including the dominance of ChatGPT focused and survey-based research, limited longitudinal and experimental evidence, weak direct comparison with traditional teaching, uneven attention to equity and accessibility, and the limited direct measurement of psychological constructs.

Building on this synthesis, the study proposes the GenAI engagement framework, a literature informed conceptual framework that organizes GenAI supported learning around five recurring components (initiation, autonomy and curiosity activation, sustained and deep engagement, iterative feedback and reflection, and collaboration or relatedness). The framework links patterns in the reviewed literature with SDT related mechanisms. It is not presented as a statistically validated causal model or as evidence that learners necessarily move through these components in a fixed order. Instead, it offers a structured basis for interpreting current evidence, guiding instructional design, and informing future empirical evaluation.

Research Questions

This study is guided by the following research questions (RQs):

RQ1: What psychological constructs dominate current research on GenAI in higher education, including motivation, engagement, creativity, perception, self-regulation, and related learner experiences?

- RQ2:** Which GenAI tools and study designs dominate the literature, and how are they distributed across the main psychological constructs?
- RQ3:** How are the SDT related needs of autonomy, competence, and relatedness represented and supported in current research on GenAI enhanced learning?
- RQ4:** What conceptual, methodological, contextual, and practical gaps remain in understanding how GenAI may support or hinder sustained learner motivation and engagement in higher education?

By addressing these questions, the study provides a theory informed synthesis of GenAI research in higher education. It also offers a practical conceptual framework for developing more intentional, psychologically supportive, and learner centered GenAI learning environments.

The remainder of the paper is organized as follows. We first review major GenAI tools and their educational affordances. We then present the main psychological constructs and the theoretical role of SDT. After that we describe the search, screening, coding, reliability, and synthesis procedures. Next we present the findings in relation to the RQs. We also discuss the findings, implications, limitations, and future research priorities. After presenting the GenAI engagement framework, we finally conclude the paper.

GenAI TOOLS AND EDUCATIONAL AFFORDANCES

GenAI is increasingly used in higher education for writing, feedback, brainstorming, explanation, visualization, coding, and creative work. Unlike technologies that mainly deliver fixed content, GenAI systems can generate and revise text, images, audio, code, and other outputs in response to learner input. These capabilities create opportunities for more interactive and personalized learning, but educational benefits are not automatic. Their value depends on how technology is integrated into learning activities, whether students remain responsible for evaluating their outputs, and whether its use supports rather than replaces meaningful cognitive effort. Poorly guided use may contribute to inaccurate learning, overreliance, superficial task completion, or reduced critical engagement (Baidoo-Anu & Owusu Ansah, 2023; Kasneci et al., 2023; Lo et al., 2024).

Major GenAI Modalities in Higher Education

GenAI in higher education covers several modalities. The text-based conversational systems such as ChatGPT, Claude, and Gemini are widely used for writing, explanation, feedback, language learning, and coding (Bannister et al., 2023; Chan & Hu, 2023; Zhai, 2023). Visual systems such as DALL-E and Midjourney extend GenAI to design, visual storytelling, and creative ideation (Omran Zailuddin et al., 2024; Ramesh et al., 2021). Audio and speech systems can support transcription, language practice, and alternative forms of expression, while AI copilots and embedded assistants increasingly support planning, drafting, revision, and workflow-related tasks. These modalities differ in their potential educational value and associated risks, and evidence remains considerably more developed for text-based systems than for visual, audio, multimodal, and embedded applications. **Table 1** presents the main categories.

Educational Affordances and Risks

Across these modalities, GenAI may support several aspects of learning. Personalized inquiry and learner choice may strengthen autonomy, while explanations, formative feedback, and opportunities for revision may support competence. GenAI can also assist creativity by helping learners generate, compare, and refine alternatives, and multimodal systems may provide additional ways to access or express learning content. However, these benefits depend on learners actively evaluating and developing the generated material rather than accepting it as a finished answer (Baidoo-Anu & Owusu Ansah, 2023; Kasneci et al., 2023; Kumar et al., 2024; Omran Zailuddin et al., 2024).

However, evidence on educational and motivational effects is mixed. Positive perceptions of usefulness, convenience, or satisfaction do not necessarily indicate deeper engagement or improved learning. For example, a systematic review by Lo et al. (2024) identified both engagement and disengagement in ChatGPT supported learning, including mixed emotional responses and concerns about overreliance and reduced critical thinking. The empirical evidence has also identified potentially adverse outcomes. Abbas et al. (2024) found that greater ChatGPT use was associated with increased procrastination and memory loss and with

Table 1. GenAI modalities and educational affordances

GenAI modality	Example tools	Common educational uses	Potential motivational affordances	Key risks and challenges
Text-based conversational GenAI	ChatGPT, Claude, Gemini, GPT-based systems	Writing, brainstorming, explanation, tutoring, summarization, coding, and feedback	May support autonomy through self-paced inquiry and competence through feedback and revision	Inaccurate outputs, overreliance, weak source verification, and academic integrity concerns
Visual and image GenAI	DALL-E, Midjourney, Stable Diffusion	Visual storytelling, design tasks, concept illustration, and creative ideation	May support creativity, exploration, and multimodal expression	Prompt sensitivity, representational bias, authorship concerns, and limited educational evaluation
Audio, speech, and music GenAI	Speech-to-text, text-to-speech, synthetic voice, and music generation systems	Transcription, pronunciation practice, audio feedback, communication, and creative production	May support participation, language learning, accessibility, and alternative expression	Accuracy, authenticity, consent, accessibility quality, and potential misuse
AI copilots and embedded assistants	Learning management assistants, coding copilots, writing copilots, and productivity integrated GenAI	Planning, drafting, coding, revision, feedback, and workflow support	May support self-regulation, structured learning, and continuous feedback	Privacy, hidden automation, platform dependence, bias, and unclear accountability
General or unspecified GenAI	Broad references to GenAI without a named tool	Conceptual discussion, policy analysis, and broad educational applications	Help identify wider educational and institutional implications	Limited tool specificity, weak comparability, and unclear learning mechanisms

Note. The affordances listed represent potential educational mechanisms rather than consistently demonstrated effects and their value depends on task design, learner behavior, instructional guidance, and the quality of GenAI outputs

poorer academic performance in their university student sample. Other concerns include inaccurate or fabricated information, cognitive offloading, bias, privacy, unequal access, and dependence on AI generated answers (Dempere et al., 2023; Kasneci et al., 2023; Sok & Heng, 2024). These findings suggest that GenAI should not be treated as inherently motivating or beneficial. Its effects are conditional on how, why, and under what instructional conditions it is used.

Implications for Learning Design

Therefore, the educational value of GenAI depends less on access to the technology itself than on the learning activity built around it. GenAI is more likely to support meaningful learning when students are required to question outputs, compare alternatives, verify information, revise their work, explain their decisions, and reflect on what they have learned. In comparison, using GenAI mainly to obtain finished answers may reduce opportunities for independent reasoning and sustained cognitive engagement (Abbas et al., 2024; Kirkwood & Price, 2014; Lo et al., 2024). This learning design perspective connects GenAI use with motivation, engagement, creativity, perception, and self-regulation and provides the basis for examining autonomy, competence, and relatedness through SDT in the next section.

PSYCHOLOGICAL FOUNDATIONS AND THEORETICAL FRAMING

This review examines GenAI in higher education through five related psychological constructs (motivation, engagement, creativity, perception, and self-regulation). SDT provides an additional lens for interpreting how GenAI supported learning may support or limit autonomy, competence, and relatedness. These concepts guide the coding and interpretation of the reviewed literature and inform the GenAI engagement framework.

Key Psychological Constructs in GenAI-Enhanced Learning

The psychological effects of GenAI extend beyond whether students use a particular tool. GenAI may influence why learners begin a task, how deeply they participate, how they develop ideas, how they judge technology, and how independently they manage their learning. Therefore, the five constructs below provide the main psychological categories used in this review.

Motivation

Motivation refers to interests, goals, values, and reasons that influence whether learners begin a task, persist with it, and invest effort in learning (Deci & Ryan, 2012). GenAI may support motivation by reducing initial uncertainty, providing timely assistance, and allowing learners to explore questions at their own pace. However, convenient access to generated answers does not necessarily produce sustained motivation. Excessive reliance may reduce independent effort or encourage learners to use GenAI as a shortcut rather than as support for learning (Abbas et al., 2024; Kasneci et al., 2023). Therefore, motivation is treated as a potentially supportive but context-dependent outcome.

Engagement

Engagement reflects the quality and depth of learner involvement and is commonly understood through behavioral, emotional, and cognitive dimensions (Fredricks et al., 2004). GenAI can make learning more interactive through questioning, feedback, revision, and exploration, but frequent interaction with a tool should not automatically be interpreted as meaningful engagement. The learners may use GenAI repeatedly while remaining cognitively passive or disengaged from the underlying task. Recent evidence similarly indicates that GenAI supported learning can involve both engagement and disengagement depending on how the technology is used (Lo et al., 2024). Therefore, in this review, engagement refers to purposeful and sustained involvement rather than frequency of AI use alone.

Creativity

Creativity involves developing ideas or products that are both novel and useful within a particular context (Runco, 2014). GenAI can support brainstorming, alternative perspectives, visual ideation, design exploration, and multimodal expression (Garcia, 2025; Omran Zailuddin et al., 2024). However, the novelty of AI generated content should not be treated as evidence of learner creativity. Educational value depends on whether learners select, critique, adapt, combine, and develop generated ideas through their own judgment. Therefore, GenAI can support the creative process without replacing the role of the learner in producing and evaluating meaningful creative work.

Perception

Perception refers to how students and educators judge GenAI in terms of usefulness, ease of use, trust, fairness, transparency, and relevance to learning (Chan & Hu, 2023; Ivanov et al., 2024). Such perceptions can influence adoption and continued use, but favorable attitudes should not be interpreted as evidence of improved motivation, engagement, or academic performance. A learner may find GenAI convenient or useful while gaining little educational benefit from its use. Therefore, perception is treated separately from demonstrated psychological or learning outcomes.

Self-Regulation

Self-regulation refers to the ability of the learner to plan, monitor, evaluate, and adjust learning strategies in pursuit of academic goals (Panadero, 2017; Zimmerman, 2002). GenAI may support this process through planning assistance, feedback, alternative explanations, and opportunities for revision. At the same time, dependence on generated responses may weaken independent monitoring, decision making, and cognitive effort when learners accept outputs without evaluation (Abbas et al., 2024; Kasneci et al., 2023). Therefore, GenAI supported self-regulation depends on whether learners remain responsible for setting goals, evaluating information, and deciding how AI feedback should shape their work.

SDT as the Theoretical Lens

SDT proposes that meaningful and sustained motivation is more likely when three basic psychological needs are supported: autonomy, competence, and relatedness (Deci & Ryan, 2012). These needs provide a useful lens for examining why similar GenAI tools may support learning in some situations but contribute to dependence or shallow engagement in others.

- **Autonomy:** Autonomy reflects a learner's sense of choice, ownership, and control. GenAI may support autonomy when learners use it to explore alternatives, ask personalized questions, or make informed

decisions about their work. However, autonomy may be weakened when students become dependent on generated suggestions or feel that AI use limits rather than support their own judgment.

- **Competence:** Competence refers to a learner's sense of capability, progress, and effectiveness. Explanations, scaffolding, feedback, and opportunities for revision may help learners work through difficult tasks. However, receiving an answer or feeling supported by a tool is not the same as developing competence. Learners need to understand, evaluate, and independently apply what they have learned.
- **Relatedness:** Relatedness concerns connection, belonging, and meaningful interaction with other people. Interaction with an AI system alone does not constitute relatedness. GenAI may support it indirectly when generated content is used within peer discussion, teacher student interaction, collaborative critique, group reflection, or co-creation. In this sense, relatedness depends primarily on the social learning design surrounding technology.

SDT therefore provides a theory informed way to distinguish simple technology use from learning experiences that preserve learner choice, support genuine progress, and maintain meaningful human interaction. It also supports the distinction made in this review between directly measured psychological needs and needs that are explicitly discussed or inferred from reported learning mechanisms.

Construct Interdependence and Relevance to the Review

The five psychological constructs are related but conceptually different. Motivation concerns why learners begin and persist with learning, whereas engagement reflects the depth and quality of their participation. Creativity concerns the development and refinement of ideas, perception captures how learners judge GenAI, and self-regulation concerns how they plan, monitor, and adjust their learning. A single GenAI supported activity may influence several of these constructs at the same time, but evidence for one should not automatically be interpreted as evidence for another.

For this reason, the review uses two complementary forms of analysis. Psychological construct mapping identifies the learner-related outcomes explicitly examined in the literature, while the SDT lens provides a theory informed basis for interpreting how GenAI supported learning may support or limit autonomy, competence, and relatedness. These perspectives allow the review to describe what has been studied while examining the possible motivational mechanisms behind those findings. The next section describes how these concepts were operationalized within the review methodology.

METHODOLOGY

This review used a structured SR process to examine how GenAI has been studied in higher education in relation to motivation, engagement, creativity, perception, self-regulation, and related psychological mechanisms. The methodology was designed to map the breadth and characteristics of the evidence while maintaining a clear audit trail from search and screening to coding and synthesis.

Review Design and Reporting Framework

A SR was appropriate because research on GenAI in higher education is rapidly developing and includes diverse tools, psychological constructs, study designs, and educational contexts. The aim was therefore to identify major patterns, theoretical links, and evidence gaps rather than estimate a single intervention effect.

The review was guided by PRISMA-ScR reporting principles and established guidance for conducting SRs (Arksey & O'Malley, 2005; Tricco et al., 2018). The eligible records were first publicly available between 1 January 2021 and 31 December 2025, based on the earliest verifiable online or print publication date. This period captured the main growth of GenAI research in higher education while maintaining a fixed and reproducible eligibility boundary.

The synthesis focused on mapping the distribution and nature of the literature across psychological constructs, GenAI tools, study designs, contexts, outcomes, and SDT related needs. Because the evidence included was methodologically diverse and did not address one common intervention or outcome, a meta-analysis was neither planned nor conducted. The review also did not apply a formal risk-of-bias (RoB)

instrument. Instead, study designs and evidence characteristics were charted to provide a descriptive view of the relative strength and limitations of the evidence base. The review was not prospectively registered, and the absence of prospective registration is acknowledged as a limitation.

Search Strategy and Information Sources

The search was conducted in two stages. The original search was completed in May 2025 using Elicit, an AI-based semantic search platform. It covered five broad areas (motivation and engagement, classroom uses of ChatGPT and GenAI, creativity, practical learning applications, and student perceptions and attitudes). Elicit was used as the initial exploratory source for identifying potentially relevant literature. The five thematic search areas were preserved in the review records. However, the exact wording of the original Elicit prompts was not archived and could not be rebuilt reliably.

To broaden the search and improve transparency, a supplementary API-based search was conducted on 21 June 2026 using OpenAlex, CrossRef, and Semantic Scholar. The source adapted free text queries were used rather than applying one identical Boolean string across all platforms. Eight queries were run in OpenAlex and eight in Semantic Scholar, covering motivation and engagement, SDT-related needs, creativity and multimodal GenAI, equity and accessibility, adverse outcomes, comparisons with traditional teaching, learning analytics and conference-related signals, and disciplinary variation. Five closely related queries were used in CrossRef because of differences in its retrieval structure. The complete query strings, API endpoints, filters, requested fields, and retrieval parameters are provided in [Appendix A](#).

The OpenAlex search used the /works endpoint with publication dates limited to 1 January 2021-31 December 2025 and publication types limited to articles, reviews, and preprints. A maximum of 50 records were retrieved for each query. CrossRef searches used the /works endpoint, were limited to journal articles published between 2021 and 2025 and returned up to 40 records per query. Semantic Scholar searches used the Graph API paper search endpoint and returned up to 50 records per query without an API level publication type restriction. The pagination was not used in any of the three supplementary sources. Semantic Scholar retrieval was affected by rate limiting. Therefore, repeated requests used delays and retry procedures with exponential backoff for HTTP 429 responses.

Python scripts supported retrieval, metadata cleaning, record formatting, relevance filtering, deduplication, and preparation of the audit files. Retrieved records were rechecked against the 2021-2025 publication window and passed through a relevance heuristic requiring both GenAI/large language model terminology and education-related terminology in the title or abstract. Records were de-duplicated within the supplementary search using normalized DOI and, when a DOI was unavailable, normalized title matching. No fuzzy title matching was used. These procedures produced a supplementary pool of 375 unique records after filtering and within-pool deduplication. Automated procedures supported record management only and did not make final eligibility decisions.

Manual checking was used when indexed metadata was incomplete or unclear, primarily to verify publication status, DOI, venue, publication date, and relevance to higher education. It was not treated as an additional database search or as systematic citation chasing. Final inclusion and exclusion decisions were made by human reviewers using the shared eligibility criteria described later.

The original Elicit records and the supplementary API records were then merged into a single candidate pool for cross-source deduplication and screening, as described in the next subsection. Scopus, Web of Science, ERIC, PsycINFO, and systematic searches of AIED and LAK proceedings were not conducted. In addition, raw per query hit counts were not preserved for the supplementary APIs, and the absence of pagination means that the API searches should be interpreted as structured supplementary retrieval rather than exhaustive database coverage. These limitations are acknowledged later.

Eligibility, Record Management, and Study Selection

Studies were eligible when they focused on GenAI or LLM-based tools in higher education and addressed at least one learner related construct, such as motivation, engagement, creativity, perception, self-regulation, autonomy, competence, or relatedness. Eligible publications included peer reviewed journal articles, full peer reviewed conference papers, systematic and narrative reviews, and peer reviewed conceptual or commentary

articles. Standalone non-peer reviewed preprints, conference abstracts, posters, short workshop summaries, and records without sufficient information for verification were excluded. Studies were also excluded when they focused only on general AI, technical system development without a learner related component, K-12 education, workplace training, or settings outside higher education.

Faculty focused and educator focused studies were retained only when they directly examined student learning, instructional design affecting learner motivation or engagement, or educator perspectives clearly linked to student experience. Studies concerned only with faculty workload, institutional policy, or technology adoption without a learner related psychological component were excluded.

Publication eligibility was based on the earliest verifiable public date. The online publication date was used where available, followed by the online first date and then the print date. As a result, papers published online in 2025 but assigned to 2026 print issues remained eligible. Records first made publicly available in 2026 were excluded. The original Elicit search produced 59 unique records, while the supplementary searches contributed 375 records, giving a combined pool of 434. Records were deduplicated using exact normalized DOI matching and, when a DOI was unavailable, exact normalized title matching. DOI prefixes, letter case, punctuation, and spacing were standardized before comparison. Fuzzy title matching was not used.

Where both a preprint and a peer reviewed version of the same study were present, the peer reviewed version was retained, and the preprint was removed as a version duplicate. Standalone preprints without an eligible published version were excluded. This process removed 17 cross-source duplicates and three preprint-publication version pairs, resulting in 20 total duplicate or version removals. The corrected study selection process was as follows:

- 434 records were identified
- 20 duplicates or version pairs were removed
- 414 records were independently screened by title and abstract
- 215 records were excluded at this stage
- 199 records underwent detailed eligibility assessment
- four records were excluded following detailed assessment or quality-control review
- 195 studies were included in the final synthesis

The four detailed or quality control (QC) exclusions comprised one original eligibility exclusion, two standalone preprints, and one secondary school study. The complete selection process is reported in the PRISMA-ScR flow diagram.

Data Charting and Coding Framework

A structured charting framework was developed to describe the characteristics, psychological focus, and evidential contribution of each included study. Bibliographic fields included title, DOI, publication year, publication category, source provenance, and venue. Contextual fields included discipline, participant population, geographic region, institution type, and learning setting.

Following the initial contextual coding, a targeted human validated re-extraction was conducted across all 195 included studies to strengthen the depth and consistency of the contextual analysis. The contextual information was reviewed using the most detailed accessible source for each record (full text PDF methods sections for 48 studies, extended abstract or publisher text for 144 studies, and corpus abstracts for three studies where no more detailed source was available). Participant population, geographic region, institution type, and learning setting were re-examined using predefined coding rules. Each variable was classified as reported, not reported, or not applicable, allowing genuinely absent information and structurally inapplicable variables to be distinguished from reported contextual characteristics. The resulting human validated contextual dataset was used for the final contextual analysis, while the previously locked analytical coding for the other review domains remained unchanged. Each study was assigned one primary psychological focus. This was determined using the following hierarchy:

1. Explicit research aims or RQ
2. Main measured or analyzed outcome

3. Theoretical or analytical framing
4. Principal findings and conclusions
5. Keyword frequency only as supporting evidence when the preceding information remained unclear

Primary focus categories included motivation, engagement, creativity, perception, and general/other. Self-regulation was retained as a multi-label thematic construct rather than a separate primary focus category because it was treated as a learning process that could accompany motivation, engagement, creativity, or perception. This classification was mutually exclusive. Studies were also assigned multi-label thematic codes for motivation, engagement, creativity, perception, self-regulation, and general/other. A study could receive more than one thematic code when several constructs were meaningfully represented. This distinction allowed the review to separate the central focus of a study from its broader psychological coverage.

The dominant GenAI tool was coded using a controlled vocabulary. ChatGPT, general GenAI, general GPT or LLM systems, AI copilots or assistant systems, visual or image GenAI, and other or unspecified tools. Each study received one dominant tool code, even when several tools were discussed. Study design was coded separately from publication category. Design categories included survey, qualitative, mixed methods, experimental, quantitative, case study, conceptual, systematic review, narrative review, and unclear. Publication category described whether the record was a journal article, conference paper, review, or conceptual publication. The review also coded evidence related to the three SDT needs (autonomy, competence, and relatedness). Every positive SDT code was assigned to one evidence level:

- Directly measured using an SDT scale
- Explicitly discussed as an SDT construct
- Indirectly mapped from reported mechanisms
- Conceptual recommendation only

Relatedness required evidence of human social connection, such as peer interaction, teacher student engagement, group work, belonging, or co-creation. Interaction with a GenAI system alone was not coded as relatedness.

The five GenAI engagement framework components were also mapped across the corpus (initiation, autonomy and curiosity activation, sustained and deep engagement, iterative feedback and reflection, and collaboration or relatedness). Deep engagement was coded only when a study described sustained cognitive, behavioral, or emotional involvement, persistence, elaboration, or deep learning (DL). Motivation, positive perception, frequent tool use, or a general mention of engagement was not sufficient.

For empirical studies, outcome direction was classified as predominantly positive, mixed or conditional, predominantly adverse, null or unclear, perception-only, or descriptive without an evaluative outcome. Additional indicators captured self-report evidence, behavioral data, objective academic performance, learning analytics, validated psychological measures, longitudinal designs, and experimental signals.

Comparator coding identified whether GenAI was compared with traditional teaching, tutoring, pre/post-performance, or another digital tool. Comparator studies were further classified as observational or self-reported comparisons, alternative digital tool comparisons, uncontrolled pre/post studies, or conceptual and qualitative comparisons. This prevented broad references to traditional teaching from being interpreted as controlled experiments.

The framework also captured equity, accessibility, and implementation issues, including unequal access, premium versus free services, connectivity, device limitations, disability support, language learning use, accessibility-based language support, resource costs, faculty training, AI literacy, privacy and governance, and cultural bias. Language learning applications were coded separately from accessibility support. Finally, studies were grouped into three descriptive evidence tiers:

- **Tier A:** Experimental or quasi-experimental evidence with stronger causal signals
- **Tier B:** Observational empirical evidence
- **Tier C:** Conceptual articles and systematic or narrative reviews

These tiers were used to describe the overall evidence base rather than to replace a formal RoB assessment. The reviewer procedures, reliability analyses, and consensus process used to produce the final locked coding matrices are described in the next subsection.

Reviewer Reliability, Quality Control, and Consensus

Screening reliability, secondary coding reliability, and contextual re-extraction reliability were assessed separately because they involved different decisions and datasets. For eligibility screening, two reviewers independently assessed all 414 records using the shared inclusion and exclusion codebook. They agreed on 81.9% of decisions, with Cohen's $\kappa = 0.642$. The remaining 75 disagreements were reviewed through consensus discussion before the final inclusion decisions were confirmed.

A later QC audit examined 36 records flagged because of publication type, preprint status, publication date, conference classification, or higher education relevance. Both reviewers independently assessed these records and reached 100% agreement. This audit resulted in metadata correction, removal of ineligible records, and confirmation of the final corpus of 195 studies.

Secondary coding was evaluated using a stratified sample of 50 studies selected with a random seed of 2026. The sample reflected variation in publication year, study design family, tool family, and publication category. Both reviewers applied the same operational codebook independently, and disagreements were documented and resolved before the final matrices were locked. For the single label variables, Cohen's κ was 0.662 for primary psychological focus, 1.000 for dominant GenAI tool, 0.667 for study design, and 1.000 for publication category.

For multi-label coding, the mean Jaccard similarity was 0.864 and the median was 0.875. The macro-average Cohen's κ across variables for which κ could be estimated was 0.887. Variables with no variation across reviewers, including connectivity and device limit coding in the reliability sample, were reported as κ not estimable rather than being assigned an artificial perfect value.

Contextual re-extraction was evaluated separately. The primary reviewer conducted the human validated contextual review across all 195 included studies, and a second reviewer independently verified a stratified sample of 50 studies selected. Across the contextual variables, the macro-average Cohen's κ was 0.945. Institution type, the most interpretive contextual variable, showed Cohen's $\kappa = 0.727$ with 80% agreement. Ten institution type disagreements were subsequently reviewed and resolved through consensus before the final contextual dataset was locked. After reliability assessment, remaining differences were resolved using the shared definitions and consensus rules. The resulting consensus matrices, including the human validated contextual dataset, formed the final locked datasets used for the descriptive analyses, tables, figures, and framework mapping. Screening, secondary coding, and contextual reliability statistics should therefore be interpreted separately because they assess different forms of reviewer agreement.

Evidence Synthesis and Methodological Boundaries

The synthesis combined descriptive mapping with theory informed interpretation. Frequencies and percentages were used to summarize publication years, psychological constructs, GenAI tools, study designs, contexts, outcomes, comparators, equity related themes, and evidence tiers. Because several variables were multi-label, their totals could exceed the number of included studies. The findings were organized around the four RQs. The analysis first examined the main psychological constructs, followed by the distribution of GenAI tools and study designs. It then considered how autonomy, competence, and relatedness were represented and identified the main conceptual, methodological, contextual, and practical gaps in the literature.

SDT was used as an interpretive lens rather than imposed as a direct measurement framework. Therefore, evidence was distinguished between needs that were directly measured, explicitly discussed, indirectly mapped, or presented only as conceptual recommendations. The GenAI engagement framework was developed by mapping the included studies to five recurring components (initiation, autonomy and curiosity activation, sustained and deep engagement, iterative feedback and reflection, and collaboration or relatedness). Study could contribute to more than one component. The mapping identifies recurring patterns in literature but does not establish temporal ordering, transition mechanisms, or causal relationships among

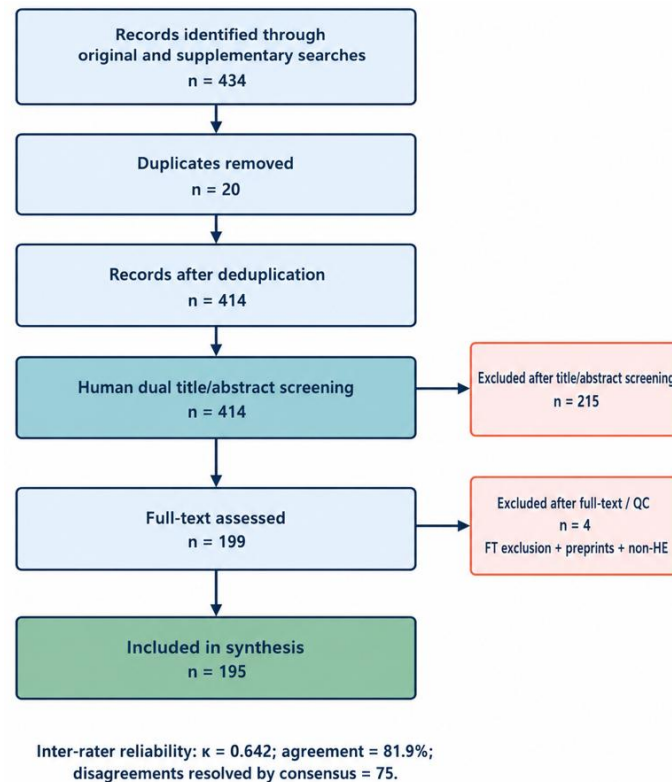


Figure 1. PRISMA-ScR flow diagram for study identification, screening, and inclusion (the authors' own work)

the components. Therefore, the framework is presented as literature informed conceptual structure requiring future empirical validation.

Several methodological boundaries remain. The search did not include direct searches of Scopus, Web of Science, ERIC, PsycINFO, or systematic AIED and LAK proceedings. The conference coverage was therefore incomplete. Secondary analytical coding relied largely on titles, abstracts, and indexed metadata, whereas the targeted contextual variables were subsequently re-extracted and human validated using the most detailed accessible source for each study, as described later. Some SDT and engagement framework classifications also required interpretation, and the review period excluded studies first published after 31 December 2025. These limitations are considered in greater detail in the discussion. The following section presents the findings from the final corpus of 195 studies and organizes them according to the RQs.

RESULTS

This section presents the findings from the final corpus of 195 studies. Results are reported in relation to study selection, corpus characteristics, psychological constructs, GenAI tools, study designs, contextual coverage, outcome direction, SDT related needs, the GenAI engagement framework, equity, and evidence strength.

Study Selection and Reliability

The combined search identified 434 records. After removing 20 duplicate or version records, 414 records remained for title and abstract screening. Of these, 215 were excluded, leaving 199 records for detailed eligibility assessment. Four additional records were excluded during detailed review and QC, resulting in a final corpus of 195 studies. The four detailed or QC exclusions included one record that failed the original eligibility assessment, two standalone non-peer reviewed preprints, and one study conducted only in a secondary school setting. The complete selection process is shown in [Figure 1](#).

Two reviewers independently screened all 414 records. They agreed on 81.9% of decisions, with Cohen's $\kappa = 0.642$. The 75 disagreements were resolved through consensus. A separate QC review of 36 flagged records produced 100% reviewer agreement.

Table 2. Study selection and reviewer reliability results

Measure	Result
Records identified	434
Duplicate or version records removed	20
Records screened	414
Title and abstract exclusions	215
Detailed assessments	199
Detailed or QC exclusions	4
Final included studies	195
Screening agreement	81.9%
Screening Cohen's κ	0.642
Screening disagreements resolved	75
QC records reviewed	36
QC agreement	100%
Secondary-coding sample	50
Primary focus κ	0.662
Dominant tool κ	1.000
Study-design κ	0.667
Publication-category κ	1.000
Mean Jaccard similarity	0.864
Median Jaccard similarity	0.875
Macro-average multi-label κ	0.887
Contextual verification sample	50
Contextual macro-average κ	0.945
Institution-type κ	0.727
Institution-type agreement	80.0%
Institution-type disagreements resolved	10

The secondary coding reliability was assessed using a stratified sample of 50 studies. Cohen's κ was 0.662 for primary psychological focus, 1.000 for dominant GenAI tool, 0.667 for study design, and 1.000 for publication category. For multi-label coding, the mean Jaccard similarity was 0.864, the median was 0.875, and the macro-average κ across estimable variables was 0.887. Variables with no variation between reviewers were reported as not estimable rather than as perfect agreement. **Table 2** presents the selection and reliability results. The contextual re-extraction reliability was assessed separately using a stratified sample of 50 studies. The macro-average Cohen's κ across contextual variables was 0.945. Institution type showed $\kappa = 0.727$ with 80% agreement, and 10 disagreements were resolved through consensus.

Corpus and Study Characteristics

The final corpus included studies published between 2023 and 2025. Publication volume increased across this period, with 30 studies published in 2023, 61 in 2024, and 104 in 2025. More than half of the included literature therefore came from 2025, reflecting the rapid recent growth of GenAI research in higher education. Survey studies were the most common design, accounting for 80 studies (41.0%). Qualitative studies followed with 29 studies (14.9%), while 24 studies (12.3%) were conceptual and 21 (10.8%) used mixed methods. The corpus also included 10 systematic reviews, 10 studies with unclear designs, six narrative reviews, and five studies each using quantitative, experimental, or case-study designs. **Figure 2** presents the study design distribution.

Most included publications were journal articles (139; 71.3%). The remaining records included 24 conceptual publications (12.3%), 17 conference papers (8.7%), and 15 review articles (7.7%). In terms of search provenance, 162 studies were identified only through the supplementary API searches, 20 only through Elicit, and 13 through both sources. **Table 3** presents the main characteristics of the final corpus.

The corpus was recently and strongly dominated by journal-based, survey-based research. Experimental, longitudinal, and other stronger causal designs remained comparatively uncommon.

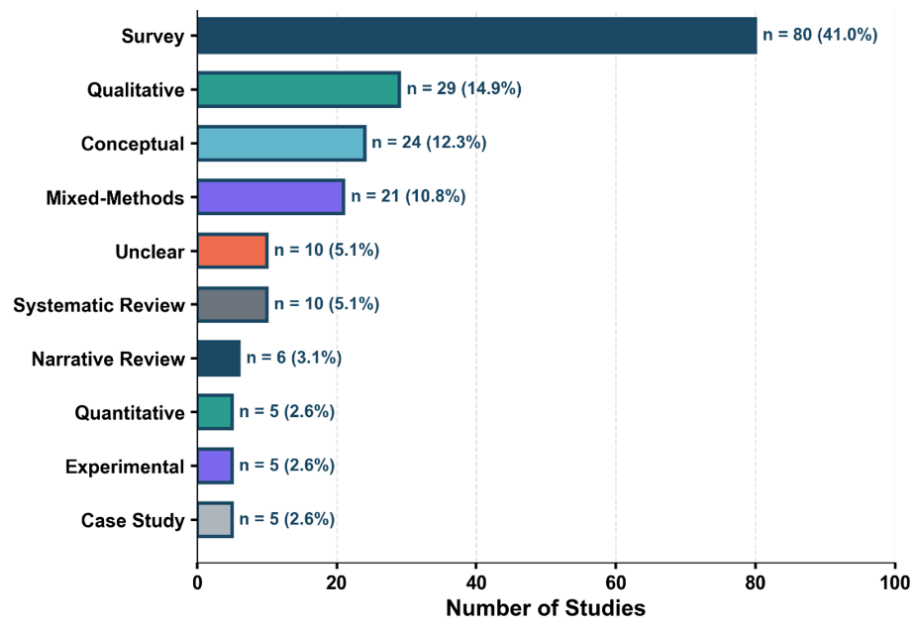


Figure 2. Study designs represented in the included literature (the authors' own work)

Table 3. Characteristics of the included studies (n = 195)

Characteristic	Category	n	Percentage (%)
Publication year	2023	30	15.4
	2024	61	31.3
	2025	104	53.3
Study design	Survey	80	41.0
	Qualitative	29	14.9
	Conceptual	24	12.3
	Mixed methods	21	10.8
	Systematic review	10	5.1
	Unclear	10	5.1
	Narrative review	6	3.1
	Quantitative	5	2.6
	Experimental	5	2.6
	Case study	5	2.6
Publication category	Journal	139	71.3
	Conceptual	24	12.3
	Conference	17	8.7
	Review	15	7.7
Source provenance	API only	162	83.1
	Elicit only	20	10.3
	Elicit and API	13	6.7

Psychological Focus and Thematic Coverage

Each study was assigned one primary psychological focus based on its research aim, main outcome, theoretical framing, and principal findings. Motivation was the most common primary focus, appearing in 61 studies (31.3%). Perception followed with 52 studies (26.7%), engagement with 36 (18.5%), and creativity with 30 (15.4%). Remaining 16 studies (8.2%) were classified as general/other. **Figure 3** presents this distribution.

The multi-label thematic analysis captured broader construct coverage because one study could address several psychological themes. Perception appeared in 96 studies (49.2%), followed by motivation in 71 (36.4%), engagement in 65 (33.3%), creativity in 63 (32.3%), and self-regulation in 26 (13.3%). General/other themes appeared in 16 studies (8.2%). As shown in **Table 4**, the primary focus and thematic results serve different purposes. Primary focus coding identifies the central psychological concern of each study, whereas thematic coding captures additional constructs discussed within the same study. For this reason, thematic totals exceed the corpus size.

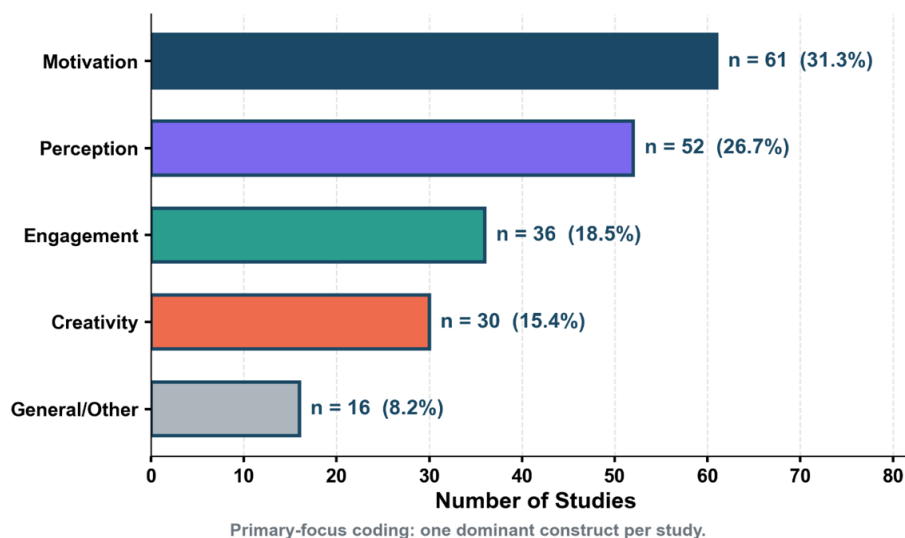


Figure 3. Primary psychological focus of the included studies (the authors' own work)

Table 4. Psychological focus and thematic coverage of the included studies (n = 195)

Construct	Primary focus, n (%)	Thematic coverage, n (%)
Motivation	61 (31.3)	71 (36.4)
Perception	52 (26.7)	96 (49.2)
Engagement	36 (18.5)	65 (33.3)
Creativity	30 (15.4)	63 (32.3)
Self-regulation	Not used as a separate primary category	
General/other	16 (8.2)	16 (8.2)

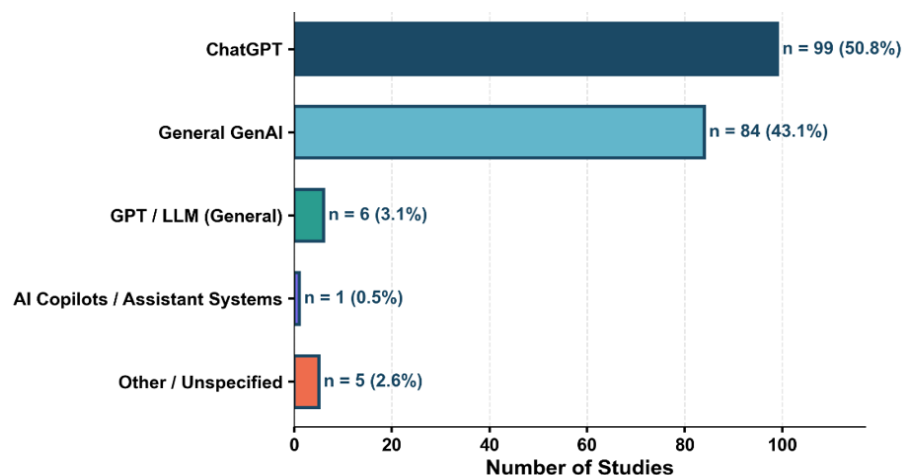


Figure 4. Dominant GenAI tools represented in the included studies (the authors' own work)

These findings show that motivation was the most common central focus, while perception was the most widely represented construct when both primary and secondary themes were considered.

GenAI Tools and Methodological Patterns

The literature was strongly concentrated on text-based GenAI. ChatGPT was the dominant tool in 99 studies (50.8%), while 84 studies (43.1%) examined GenAI in general without identifying one main system. General GPT or LLM systems were dominant in six studies (3.1%), other or unspecified tools in five (2.6%), and AI copilots or assistant systems in one study (0.5%). No study was classified as having visual or image GenAI as its dominant tool after the tool categories were standardized. The absence of visual or image GenAI from the dominant tool category does not mean that such tools were entirely absent. Rather, they were not the main system examined in any study under the final controlled coding rules. The dominant tool distribution is presented in [Figure 4](#) and [Table 5](#).

Table 5. Psychological focus and thematic coverage of the included studies (n = 195)

Dominant GenAI tool	n	Percentage (%)
ChatGPT	99	50.8
General GenAI	84	43.1
GPT/LLM general	6	3.1
Other or unspecified	5	2.6
AI copilots or assistant systems	1	0.5
Visual or image GenAI	0	0.0
Total	195	100

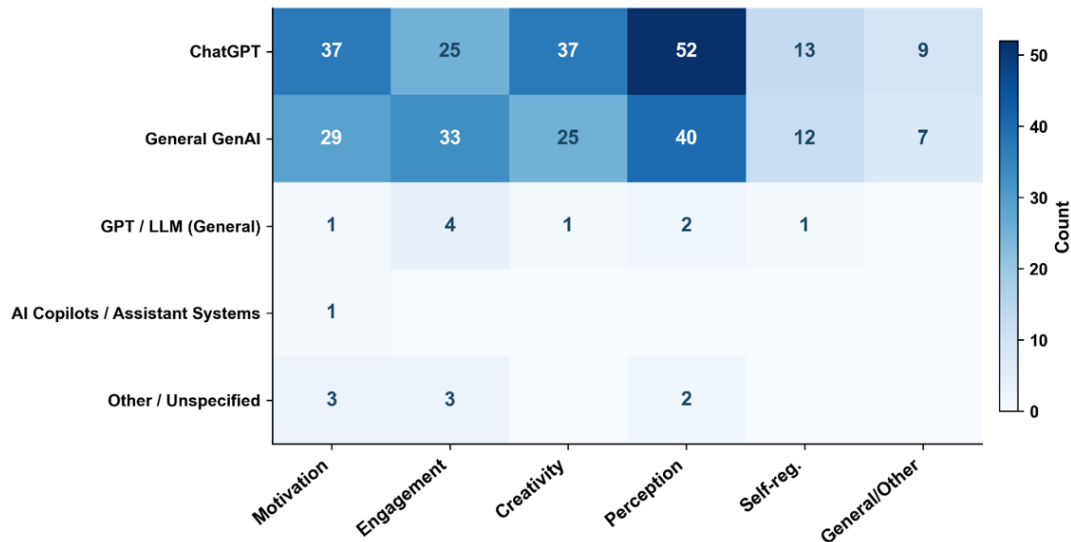
**Figure 5.** Dominant GenAI tools mapped against psychological constructs (the authors' own work)

Figure 5 maps the dominant tool categories against the multi-label psychological constructs. ChatGPT and general GenAI accounted for most studies across motivation, engagement, creativity, perception, and self-regulation. The smaller tool categories contributed only limited evidence across these constructs.

The methodological pattern shown earlier in **Figure 2** was similarly concentrated. Surveys accounted for 80 studies (41.0%), whereas only five studies (2.6%) were classified as experimental. The experimental design category represented the mutually exclusive primary study design classification. Under the broader evidence tier classification, 12 studies contained experimental or quasi-experimental signals. Among the 155 empirical studies, 16 contained longitudinal evidence, nine used validated psychological scales, three included learning analytics evidence, and 17 reported objective academic or performance outcomes. These patterns show that most of the corpus examined perceptions and experiences rather than directly measured changes in learning or behavior.

Context, Outcomes, and Comparative Evidence

Contextual characteristics were re-examined through the human validated contextual re-extraction across all 195 included studies. General or multidisciplinary higher education remained the most common disciplinary category, representing 93 studies (47.7%), followed by engineering or STEM in 27 (13.8%), language learning or linguistics in 21 (10.8%), business, economics, or tourism in 16 (8.2%), and computing or programming in 14 (7.2%). Smaller groups included health or medical education, arts and design, social sciences and humanities, and teacher education. Participant population was reported for 155 studies (79.5%), not reported in two (1.0%), and not applicable to 38 (19.5%). Mixed undergraduate and postgraduate populations were the largest category, appearing in 62 studies (31.8%), followed by mixed student and faculty populations in 49 (25.1%). World region was reported for 71 studies (36.4%) and not reported for 124 (63.6%). Among studies with regional information, Asia was the most frequently represented region with 35 studies (17.9% of the full corpus), followed by North America with 10 (5.1%), Europe with seven (3.6%), and Oceania with six (3.1%).

Table 6. Main contextual characteristics of the included studies (n = 195)

Contextual variable	Category	n	Percentage (%)
Discipline	General/multidisciplinary higher education	93	47.7
	Engineering/STEM	27	13.8
	Language learning/linguistics	21	10.8
	Business/economics/tourism	16	8.2
	Computing/programming	14	7.2
	Health/medical education	9	4.6
	Arts/design/creative disciplines	5	2.6
	Not reported	5	2.6
	Social sciences/humanities	4	2.1
	Education/teacher education	1	0.5
Participant population	Mixed undergraduate/postgraduate	62	31.8
	Mixed students and faculty	49	25.1
	Not applicable	38	19.5
	Mixed students and faculty; institutional/policy documents	19	9.7
	Institutional/policy documents	9	4.6
	Faculty/educators	6	3.1
	Postgraduate students	4	2.1
	Faculty/educators; institutional/policy documents	2	1.0
	Undergraduate students	2	1.0
	Not reported	2	1.0
	Undergraduate students; institutional/policy documents	1	0.5
	Postgraduate students; faculty/educators	1	0.5
World region	Not reported	124	63.6
	Asia	35	17.9
	North America	10	5.1
	Europe	7	3.6
	Oceania	6	3.1
	Middle East	5	2.6
	Africa	5	2.6
	Multi-region/global	2	1.0
	Latin America and Caribbean	1	0.5
Institution type	University/higher education institution	92	47.2
	Not reported	39	20.0
	Not applicable	38	19.5
	Multiple institutions	13	6.7
	Public	5	2.6
	Other higher education setting	4	2.1
	Private	4	2.1
Learning setting	Not reported	119	61.0
	Online	35	17.9
	Not applicable	30	15.4
	Blended/hybrid	9	4.6
	Face-to-face	2	1.0

Note. Contextual variables were human validated across the complete corpus using the most detailed accessible source for each study: *Not reported* indicates that the relevant information could not be established from the accessible source, whereas *Not applicable* indicates that the contextual variable did not meaningfully apply to the publication type

Institution type was reported for 118 studies (60.5%), not reported for 39 (20.0%), and not applicable to 38 (19.5%). Universities or other higher education institutions represented the largest identified category, accounting for 92 studies (47.2%), while 13 studies (6.7%) involved multiple institutions. Learning setting was reported for 46 studies (23.6%), not reported for 119 (61.0%), and not applicable to 30 (15.4%). Online learning was the most frequently identified setting, appearing in 35 studies (17.9%), followed by blended or hybrid settings in nine (4.6%) and face-to-face settings in two (1.0%). **Table 6** presents the revised contextual characteristics.

Outcome direction was assessed across the 155 empirical studies (**Table 7**). Mixed or conditional findings were the largest category, accounting for 72 studies (46.5%). Predominantly positive outcomes were reported in 34 studies (21.9%). Seventeen studies (11.0%) reported perception-only evidence, and another 17 (11.0%) were descriptive without evaluating an outcome. Predominantly adverse findings appeared in 10 studies (6.5%), while five studies (3.2%) reported null or unclear effects.

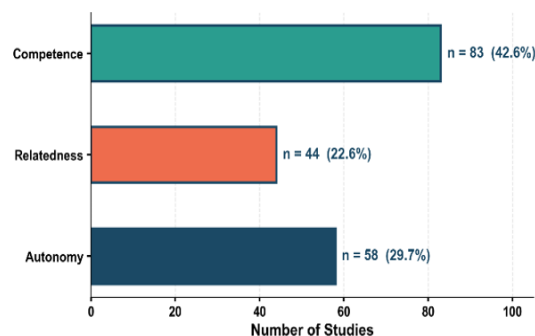
Table 7. Outcome direction across empirical studies (n = 155)

Outcome direction	n	Percentage (%)
Mixed or conditional	72	46.5
Predominantly positive	34	21.9
Perception-only evidence	17	11.0
Descriptive without an evaluative outcome	17	11.0
Predominantly adverse	10	6.5
Null or no clear effect	5	3.2
Total	155	100

Table 8. Profile of comparator evidence (n = 18)

Comparator characteristic	Category	n
Evidence type	Observational or self-reported comparison	10
	Alternative digital tool comparison	5
	Pre/post comparison without a control group	2
	Conceptual or qualitative comparison	1
Comparative result	Inconclusive	12
	Better for GenAI	3
	Mixed	1
	Worse for GenAI	1
	Similar	1
Traditional-pedagogy evidence	Traditional teaching, tutoring, or feedback category	9
	Active controlled or quasi-controlled traditional comparison	0

Note. Comparative-result categories summarize the direction reported across heterogeneous comparator designs and should not be interpreted as estimates of comparative effectiveness

**Figure 6.** Representation of autonomy, competence, and relatedness in the included studies (the authors' own work)

The empirical evidence was also limited in several respects. Twenty-six studies relied only on self-reported evidence, 27 included behavioral evidence, 17 assessed objective academic or performance outcomes, nine used validated psychological scales, three used learning analytics, and 16 contained longitudinal evidence. Only 18 studies included any form of comparison. Ten used observational or self-reported comparisons, five compared GenAI with another digital tool, two used pre/post comparisons without a control group, and one used a conceptual or qualitative comparison. Nine studies referred to traditional teaching, tutoring, or feedback category, but none used an active controlled or quasi-controlled comparison with traditional education. Across the 18 comparator studies, 12 produced inconclusive findings, three favored GenAI, one reported mixed result, one favored the comparator, and one found similar outcomes. **Table 8** presents comparative evidence.

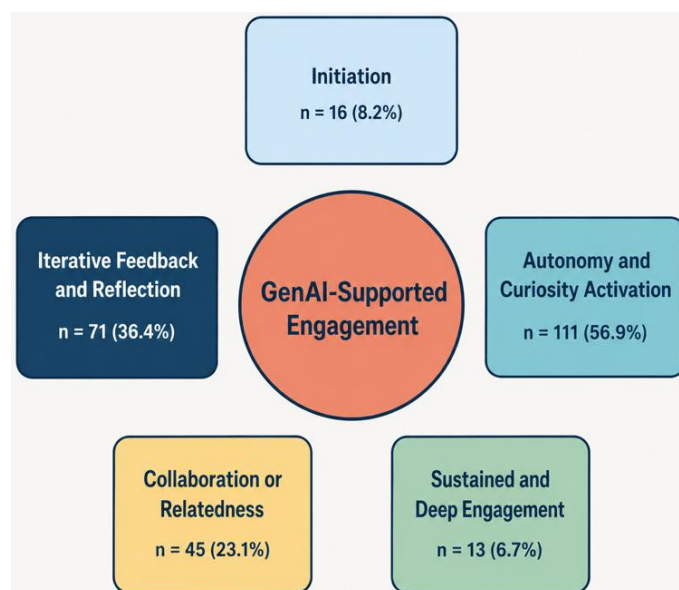
The mixed findings were more common than uniformly positive results. Since no study used an active controlled or quasi-controlled comparison between GenAI and traditional education, the present evidence does not permit conclusions about comparative effectiveness of GenAI relative to traditional teaching.

SDT Related Evidence

SDT related evidence was mapped across all 195 studies. Competence was the most frequently represented need, appearing in 83 studies (42.6%). Autonomy was identified in 58 studies (29.7%), while relatedness appeared in 44 studies (22.6%). **Figure 6** presents the distribution of the three needs.

Table 9. SDT-related needs and evidence levels in the included studies (n = 195)

SDT need	Present, n (%)	Directly measured	Explicitly discussed	Indirectly mapped	Conceptual recommendation
Autonomy	58 (29.7)	1	15	36	6
Competence	83 (42.6)	1	17	61	4
Relatedness	44 (22.6)	1	14	29	0

**Figure 7.** Evidence mapping of the GenAI engagement framework components (component placement does not imply temporal sequence, progression, or causal relationships) (the authors' own work)

Most SDT related evidence was indirect rather than based on formal measurement. Only one study directly measured each of autonomy, competence, and relatedness using an SDT-based scale. Competence was indirectly mapped in 61 studies, autonomy in 36, and relatedness in 29. [Table 9](#) presents the evidence levels.

Autonomy and competence co-occurred in 31 studies, while 17 studies contained evidence related to all three needs. Relatedness was coded only when the study included human social connection, such as peer interaction, teacher-student engagement, group work, belonging, or co-creation. Interaction with a GenAI system alone was not treated as relatedness. These findings show that SDT related mechanisms are visible in the literature, but direct and theory-based measurement remains uncommon.

GenAI Engagement Framework Mapping

The 195 studies were mapped to the five components of the GenAI engagement framework using the refined operational definitions. Autonomy and curiosity activation was the most frequently represented component, appearing in 111 studies (56.9%). Iterative feedback and reflection appeared in 71 studies (36.4%), followed by collaboration or relatedness in 45 studies (23.1%). Initiation was identified in 16 studies (8.2%), while sustained and deep engagement was the least represented component, appearing in 13 studies (6.7%). The collaboration or relatedness component was coded separately from SDT relatedness (the framework component captured collaborative learning processes, whereas the SDT code required evidence specifically related to human connection, belonging, or relational support). [Figure 7](#) presents the final literature-based mapping.

Studies were included only when they described sustained cognitive, behavioral, or emotional involvement, persistence, elaboration, or DL ([Table 10](#)). Motivation alone, positive perception, frequent GenAI use, or a general mention of engagement was not sufficient. Several components appeared together within the same studies. For example, autonomy and curiosity activation co-occurred with iterative feedback and reflection in 50 studies and with collaboration or relatedness in 33 studies. However, these overlaps indicate that the components may coexist within the same studies. They should not be interpreted as evidence of temporal progression, transitions between components, or causal mechanisms. Therefore, mapping supports the presence of the five components in literature, but not a specific sequence among them.

Table 10. Evidence mapped to the GenAI engagement framework (n = 195)

Framework component	N	%	Empirically operationalized or observed	Explicitly discussed	Inferred	Conceptual only
Initiation	16	8.2	11	2	0	3
Autonomy and curiosity activation	111	56.9	70	12	15	14
Sustained and deep engagement	13	6.7	7	0	2	4
Iterative feedback and reflection	71	36.4	35	11	13	12
Collaboration or relatedness	45	23.1	21	10	6	8

Table 11. Equity, accessibility, and implementation themes (n = 195)

Theme	n	Percentage (%)
Language-learning application	43	22.1
AI literacy	37	19.0
Disability support	24	12.3
Privacy and governance	22	11.3
Faculty training	12	6.2
Resource cost	11	5.6
Unequal access	7	3.6
Accessibility oriented language support	7	3.6
Translation or multilingual accessibility	4	2.1
Simplification or literacy support	4	2.1
Premium versus free access	1	0.5
Device limitations	1	0.5
Cultural bias	1	0.5
Connectivity	0	0.0
Low-resource language access	0	0.0
Speech or audio accessibility	0	0.0

Note. Theme counts are non-mutually exclusive (translation or multilingual accessibility and simplification or literacy support are overlapping subcategories of accessibility-based language support and should not be summed)

Equity, Accessibility, and Evidence Strength

Equity, accessibility, and implementation issues were unevenly represented across the corpus (Table 11). AI literacy was the most frequently identified implementation-related theme, appearing in 37 studies (19.0%). Disability support was discussed in 24 studies (12.3%), privacy and governance in 22 (11.3%), faculty training in 12 (6.2%), and resource costs in 11 (5.6%). Unequal access appeared in seven studies (3.6%), while premium vs. free access appeared in one study (0.5%). Device limitations were identified in one study, and no study was coded for connectivity as a main equity concern under the applied definitions. Language learning applications appeared in 43 studies (22.1%). However, only seven studies (3.6%) addressed language support as an accessibility issue. This included translation, multilingual access, simplification, or literacy support. Therefore, language teaching and accessibility were treated as separate categories.

The evidence strength profile showed that 12 studies (6.2%) were classified as tier A, representing experimental or quasi-experimental signals. Tier B included 143 observational empirical studies (73.3%), while tier C included 40 conceptual and review studies (20.5%). Tier A and tier B together formed the empirical outcome corpus of 155 studies (Table 12).

The overall evidence base was therefore dominated by observational studies, with relatively limited experimental, longitudinal, objective performance, and validated psychological evidence. The results identified several consistent gaps across literature.

- The evidence remained strongly concentrated on ChatGPT, general GenAI systems, surveys, and short-term learner perceptions. Visual or image GenAI was not the dominant tool in any included study, while AI copilots or assistant systems were dominant in only one. Other non-text modalities were not separately quantified under the dominant tool coding scheme.
- Stronger empirical designs were uncommon. Only 12 studies showed experimental or quasi-experimental signals, 16 included longitudinal evidence, 17 reported objective academic or performance outcomes, and nine used validated psychological scales.

Table 12. Evidence-strength profile of the included studies (n = 195)

Evidence tier	Description	n	Percentage (%)
Tier A	Experimental or quasi-experimental evidence	12	6.2
Tier B	Observational empirical evidence	143	73.3
Tier C	Conceptual, systematic review, or narrative review evidence	40	20.5
Total		195	100

- Mixed or conditional findings were more common than predominantly positive findings. Direct comparative evidence was limited, and no study used an active controlled or quasi-controlled comparison between GenAI and traditional education. Therefore, comparative effectiveness remains unresolved and represents an important methodological gap in the current literature.
- SDT related needs were usually inferred from mechanisms such as feedback, choice, confidence, or collaboration rather than measured directly. Direct SDT measurement was limited to one study for each psychological need.
- The engagement framework mapping showed uneven support across its components. Autonomy and curiosity and iterative feedback were widely represented, whereas initiation and sustained deep engagement appeared much less often. The evidence supported component level associations but not a fixed or causal sequence.
- Contextual, equity, and accessibility reporting remained limited. Despite the human validated contextual re-extraction, geographic region and learning setting remained unreported in many studies, while access challenges, resource needs, and accessibility outcomes were examined in relatively few studies.

These findings answer the four RQs and provide the evidence base for the following Discussion, which considers their theoretical, methodological, and practical implications.

DISCUSSION

This SR shows that research on GenAI in higher education has grown quickly, but the evidence remains uneven. The final corpus of 195 studies gives substantial attention to motivation, perception, engagement, and creativity, but much of this work is concentrated on ChatGPT, general GenAI systems, and survey-based research. Mixed or conditional outcomes were more common than consistently positive findings, and only a small proportion of studies used experimental, longitudinal, or objective performance measures. These patterns suggest that the field has moved beyond simply asking whether students are willing to use GenAI. However, it has not yet developed a strong evidence base for explaining how GenAI affects sustained motivation, deep engagement, self-regulation, collaboration, or learning performance over time. Therefore, the findings support cautious interpretation and highlight the need to distinguish promising educational possibilities from demonstrated learning effects.

Interpretation Through SDT

SDT provides a useful way to interpret how GenAI may influence learner motivation through autonomy, competence, and relatedness (Deci & Ryan, 2012). Competence was the most frequently represented need, appearing in 83 studies, followed by autonomy in 58 and relatedness in 44. However, only one study directly measured each need using an SDT-based scale. Most evidence was inferred from mechanisms such as feedback, confidence, choice, collaboration, or instructional recommendations. Competence was commonly associated with immediate explanations, feedback, scaffolding, and opportunities for revision. These features may help students feel more capable and make progress when working through unfamiliar tasks. However, feeling supported by a tool is not the same as developing genuine competence. Learning is more likely when students examine the feedback, correct mistakes, explain their reasoning, and apply what they have learned independently.

Autonomy was reflected in self-paced inquiry, personalized questioning, choice of learning direction, and control over revision. GenAI may support autonomy when learners use it to explore alternatives and make informed decisions. It may weaken autonomy when students accept generated content without evaluation or

become dependent on the system for decisions, they should make themselves. Therefore, autonomy depends less on the availability of GenAI and more on whether the learning activity preserves student ownership and judgment. Relatedness requires even more careful interpretation. Interaction with GenAI alone does not create social connection. The 44 studies mapped to relatedness included evidence of peer discussion, teacher student interaction, group work, shared reflection, or co-creation. Therefore, GenAI may support relatedness indirectly when its outputs become the basis for meaningful human interaction. Collaborative critique, peer review, and group interpretation are more likely to support relatedness than isolated chatbot use.

SDT helps explain why the same GenAI tool may support meaningful learning in one context but encourage passive use in another. The difference lies in whether the activity strengthens learner choice, progress, and human connection.

Scholarly Implications

The review highlights an important difference between positive perceptions and demonstrated educational outcomes. Perception was the most widely represented thematic construct, appearing in 96 studies, while surveys were the most common design. These studies provide useful information about acceptance, trust, satisfaction, and intention to use. However, favorable attitudes toward GenAI should not be treated as evidence of improved motivation, engagement, or academic performance. This distinction is supported by the outcome analysis. Mixed or conditional findings accounted for 72 of the 155 empirical studies, compared with 34 studies reporting predominantly positive outcomes. Ten studies reported mainly adverse findings, while others were descriptive, perception focused or showed no clear effect. Therefore, the evidence does not support the view that GenAI is uniformly beneficial. Its effects appear to depend on factors such as task design, learner skills, instructor guidance, discipline, assessment practices, and the purpose for which the technology is used.

The methodological profile also remains limited. Only 12 studies showed experimental or quasi-experimental signals, 16 included longitudinal evidence, 17 assessed objective performance, and nine used validated psychological scales. Most empirical evidence was observational. This makes it difficult to determine whether GenAI causes changes in motivation or learning, whether effects are sustained, or whether early enthusiasm declines as use becomes routine. Direct comparative evidence was similarly weak. Only 18 studies included any comparator, and none used an active controlled or quasi-controlled comparison between GenAI and traditional education. Accordingly, this review cannot determine whether GenAI is more, less, or equally effective than traditional education. The comparator findings should instead be interpreted as evidence of a methodological gap and a priority for future controlled research.

The strong focus on ChatGPT and broad GenAI categories creates another limitation. Visual, audio, multimodal, and embedded assistant systems were rarely the dominant focus of the reviewed studies. As a result, the current literature provides limited insight into whether different modalities support different psychological or learning processes. Visual systems may be particularly relevant to design and creativity, while audio and speech tools may support language learning and alternative forms of participation. These possibilities require direct investigation rather than assumption. The human validated contextual re-extraction improved the depth of contextual characterization, particularly for participant population and institution type. However, reporting remained uneven across the literature. Geographic region was reported in 71 studies (36.4%), while learning setting was reported in 46 (23.6%), indicating that important contextual information was still unavailable for many publications. Therefore, the findings should not be assumed to generalize equally across countries, institutional contexts, delivery settings, disciplines, or student populations.

Practical Implications for Educators and Designers

For educators, the main implication is that GenAI should be used as part of a learning process rather than as a shortcut to a finished answer. Activities are more likely to support learning when students are required to question, compare, verify, revise, and reflect on generated content. For example, students might use GenAI to produce an initial essay plan, identify its weaknesses, compare it with course criteria, revise it, and explain why the final version is stronger. In a design course, students might generate several prototypes, evaluate

bias or feasibility, discuss them with peers, and justify their final decisions. These activities preserve student responsibility while using GenAI as a source of ideas, feedback, or alternatives.

Learning activities should also be designed around psychological needs. Autonomy can be supported by offering meaningful choices in topics, tools, prompts, and revision strategies. Competence can be supported through structured feedback, progressive tasks, opportunities to correct errors, and visible improvement. Relatedness can be strengthened through peer critique, collaborative prompting, group reflection, and shared creation.

Educators should also distinguish assistance from substitution. GenAI can help students begin difficult tasks, but it should not replace analysis, judgment, source verification, or original reasoning. Clear expectations are therefore needed regarding acceptable use, attribution, academic integrity, and responsibility for errors.

For instructional designers and technology developers, useful features may include revision histories, source checking prompts, explanations of generated suggestions, confidence indicators, opportunities for learner choice, and collaborative workspaces. Privacy, transparency, bias, and accessibility should be considered from the beginning rather than added after implementation. The equity findings also show that access cannot be taken for granted. Premium services, device availability, AI literacy, disability support, language accessibility, and institutional resources may shape benefits from GenAI. Therefore, responsible implementation requires staff training, student guidance, accessible alternatives, and policies that do not disadvantage learners who cannot or choose not to use a particular system.

Limitations and Future Research Directions

- **Search coverage and reproducibility:** The supplementary search expanded the original Elicit-based retrieval through OpenAlex, CrossRef, and Semantic Scholar, but Scopus, Web of Science, ERIC, PsycINFO, and systematic searches of AIED and LAK proceedings were not conducted. The review was not prospectively registered. The exact wording of the original Elicit prompts was not preserved. Although the complete supplementary API queries and parameters are provided in Supplementary [Appendix A](#), the API searches were capped at 50 records per query for OpenAlex and Semantic Scholar and 40 for CrossRef without pagination, and raw per query hit counts were not retained. These factors may have limited retrieval completeness and reproducibility. Future reviews should prospectively register their protocols, preserve complete search records, and use broader databases and paginated retrieval strategies.
- **Contextual reporting and source depth:** Contextual characteristics were re-extracted and human validated across all 195 studies using the most detailed accessible source. However, source depth varied, with full text PDF methods sections available for 48 studies, extended abstract or publisher text for 144, and corpus abstracts for three. Therefore, geographic region and learning setting remained unavailable for many studies. Future research should report geographic location, institution type, participant population, discipline, delivery mode, and access conditions more consistently, particularly in underrepresented and resource-constrained settings.
- **Interpretive coding and evidence appraisal:** Independent screening, reliability assessment, human validation, and consensus coding strengthened consistency, but some classifications remained interpretive, particularly indirect SDT mapping, outcome direction, equity themes, and components of the GenAI engagement framework. The review also did not apply a formal RoB instrument, and the evidence tiers should not be interpreted as study level quality ratings. Future reviews should integrate formal quality appraisal and, where feasible, prospectively registered protocols.
- **Study design and comparative evidence:** Much of the available evidence remains observational, short-term, and self-reported, while controlled comparisons with traditional education were absent. Future studies should use longitudinal, experimental, and quasi-experimental designs comparing clearly defined GenAI supported activities with traditional instruction, human tutoring, non-GenAI digital tools, and alternative GenAI approaches. Outcomes should extend beyond perceptions to performance, retention, transfer, critical thinking, creativity, and self-regulation.

- **Measurement, modality, and equity:** Autonomy, competence, and relatedness were rarely measured directly, and behavioral, learning analytics, and objective evidence remained limited. Visual, audio, multimodal, and embedded GenAI systems were also underrepresented. Future studies should combine validated psychological measures with behavioral and performance indicators while examining modality, accessibility, multilingual support, cost, connectivity, AI literacy, cultural bias, privacy, and unequal access to premium systems.
- **Framework validation:** The GenAI engagement framework is a literature informed conceptual structure rather than a validated model of learner progression. The present evidence supports the recurrence and overlap of its five components but does not establish temporal ordering, transition mechanisms, or causal relationships. Future research should test whether the components can be measured reliably, how they interact across learning contexts, whether distinct engagement pathways emerge, and under what conditions they support sustained and meaningful engagement.

The next section presents the GenAI engagement framework as a literature informed structure derived from the patterns identified across the reviewed literature.

GenAI ENGAGEMENT FRAMEWORK: A LITERATURE-INFORMED STRUCTURE

The findings show that research on GenAI in higher education has expanded rapidly, but it still offers an incomplete account of how learners remain meaningfully engaged over time. Much of the evidence focuses on short-term perceptions, motivation, and tool use, while sustained engagement, repeated feedback, collaboration, and direct psychological measurement receive less attention. Literature is also dominated by ChatGPT and broad text-based GenAI systems, with limited evidence on visual, audio, multimodal, and embedded tools.

Related digital learning research also suggests that interactive affordances, feedback, and repeated participation can support sustained engagement. For example, Kazu and Kuvvetli (2023) showed how structured interaction in digital game-based language learning can contribute to vocabulary learning and continued learner participation. Although GenAI differs from digital games, this evidence provides a useful precedent for examining how digital affordances may support motivation and engagement over time.

The GenAI engagement framework brings these findings together by organizing five recurring components identified across the reviewed literature. It does not propose a developmental sequence, transition mechanism, or causal pathway. Rather, the components represent analytically different but potentially overlapping aspects of GenAI supported engagement whose relationships require future empirical testing. It connects the patterns identified in the review with SDT, particularly the needs for autonomy, competence, and relatedness (Deci & Ryan, 2012).

The framework is not presented as a statistically validated model or as proof that learners follow one fixed sequence. Instead, it organizes five recurring components that may occur independently or overlap in different combinations depending on the learner, task, tool, and instructional context. The evidence mapped to these components is reported in [Figure 7](#) and [Table 10](#).

Framework Components

- **Initiation:** Initiation refers to the point at which GenAI helps a learner begin or structure a task. This may include generating starting ideas, clarifying an unfamiliar concept, creating an initial outline, exploring possible questions, or producing an early visual prompt. Only 16 studies were mapped to this component under the refined coding rules. This relatively small number reflects the strict requirement that studies explicitly describe how GenAI supported task entry or early task structuring. Within the framework, initiation is not intended to replace student effort. Its role is to reduce uncertainty and provide a starting point from which the learner can continue independently.
- **Autonomy and curiosity activation:** Autonomy and curiosity activation refers to learner choice, exploration, and self-directed inquiry. It was the most frequently represented component, appearing in 111 studies. This pattern suggests that much of the literature describes GenAI as a tool for personalized questioning, exploration of alternatives, and learner-controlled interaction. However,

having many options does not automatically create autonomy. Autonomy is supported only when students retain meaningful control over what they ask, which outputs they accept, how they revise their work, and which learning direction they follow. It may be weakened when learners follow AI-generated suggestions without evaluation or feel that using GenAI is compulsory.

- **Sustained and deep engagement:** Sustained and deep engagement refers to continued cognitive, behavioral, or emotional involvement in learning. It includes persistence, elaboration, critical examination, DL, and willingness to revise or extend one's thinking. Only 13 studies met the refined definition of this component. Motivation, positive attitudes, frequent use, or a general mention of engagement were not enough. This distinction is important because regular interaction with GenAI does not necessarily mean that students are thinking deeply. Deep engagement requires learners to question, evaluate, connect, and develop the generated material rather than simply accept it.
- **Iterative feedback and reflection:** Iterative feedback and reflection refer to repeated cycles of feedback, evaluation, revision, and self-assessment. This component appeared in 71 studies and is closely linked to competence and self-regulation. GenAI may support this process by providing explanations, suggestions, alternative approaches, or feedback on a draft. However, educational value comes from how learners interpret and act on that feedback. A single generated response is less likely to support learning than a structured cycle in which students examine the advice, decide what is useful, revise their work, and explain what changed.
- **Collaboration or relatedness:** Collaboration or relatedness refers to the use of GenAI within human social interaction. It appeared in 45 studies and included peer discussion, teacher-student interaction, group critique, shared reflection, and co-creation. GenAI interaction alone was not treated as relatedness. The framework assumes that relatedness develops when generated outputs create opportunities for students and educators to communicate, negotiate meaning, compare ideas, and work together. GenAI may support this process, but the social connection comes from the learning design and human interaction surrounding the tool.

Relationship With SDT

The framework connects the five components with autonomy, competence, and relatedness, but these relationships are not one-to-one.

- Autonomy is most clearly reflected in curiosity-driven exploration, meaningful choice, and learner control. It can also appear during initiation when students select how to approach a task or during feedback when they decide which suggestions to accept.
- Competence is closely connected to initiation, feedback, and deep engagement. Early support may help students begin a difficult task, while repeated feedback and revision can strengthen understanding and perceived progress. However, competence requires active learning rather than simple dependence on generated answers.
- Relatedness is mainly represented through collaboration, peer interaction, and shared reflection. It may also strengthen engagement when students feel connected to a learning community and have opportunities to discuss or challenge AI-generated material.

The framework therefore uses SDT to explain why the same tool may produce different outcomes. A GenAI activity may be motivationally supportive when it preserves choice, supports genuine progress, and creates meaningful human interaction. The same activity may lead to shallow engagement when it removes learner responsibility or replaces dialogue and independent thought.

From Evidence Mapping to Learning Design

The framework is intended to help educators move beyond isolated GenAI use. A learning activity should not end when a system generates an answer. Instead, the activity can guide students through exploration, evaluation, revision, reflection, and collaboration.

For example, students in a writing course may use GenAI to generate several possible essay directions. They can then select one, assess the initial outline against course criteria, develop a draft, request feedback,

Table 13. GenAI engagement framework components and their instructional application

Framework component	Main educational role	Main SDT connection	Example instructional use	Possible indicators
Initiation	Help learners begin or structure a task	Autonomy and competence	Generate starting questions, outlines, examples, or initial concepts	Task start delay, number of initial alternatives, learner confidence at task entry
Autonomy and curiosity activation	Supports exploration, choice, and learner-directed inquiry	Autonomy	Compare different prompts, explanations, tools, or learning pathways	Learner choice, prompt diversity, self-directed questions, perceived autonomy
Sustained and deep engagement	Encourages persistence, elaboration, and critical thought	Autonomy and competence	Critique, extend, verify, or apply generated material	Time on task, depth of revision, persistence, cognitive-engagement measures
Iterative feedback and reflection	Supports monitoring, revision, and improvement	Competence and self-regulation	Use repeated feedback cycles and written reflection on changes	Number of revision cycles, revision quality, reflection depth, performance improvement
Collaboration or relatedness	Connects GenAI use with human discussion and co-creation	Relatedness	Peer critique, collaborative prompting, group interpretation, or joint production	Peer interaction, group participation, collaborative-output quality, perceived relatedness

revise the argument, and explain which suggestions they accepted or rejected. Peer review can extend the activity by asking students to compare generated advice and discuss which revisions improved the work.

In a visual-design course, students may use image-generation tools to produce several prototypes. They can critique representational bias, compare design choices in groups, refine their prompts, and justify the outcome. In this process, GenAI supports ideation, but students remain responsible for evaluation and decision-making. **Table 13** summarizes how the framework can be translated into instructional activities and future measurement.

Framework Boundaries and Future Validation

The GenAI engagement framework should be interpreted as a non-sequential conceptual structure rather than a model of learner progression. Its five components may appear individually or in different combinations, and not every learning activity will necessarily include all five. Their relationships may also vary across learners, tasks, tools, and instructional contexts.

The framework also does not assume that every component produces a positive outcome. Feedback may confuse learners when it is inaccurate, autonomy may be reduced through dependence, and collaboration may remain superficial if students do not meaningfully discuss the generated content. Instructional conditions therefore shape whether each component supports or limits learning.

Future research can test the framework by developing reliable measures for its components and examining how they relate across tasks, disciplines, tools, and time. Longitudinal and learning analytics studies could examine prompt development, feedback cycles, revision depth, persistence, peer interaction, and changes in autonomy, competence, and relatedness. Experimental studies could also compare complete framework-based activities with simpler answer-generation tasks.

Therefore, the main contribution of the framework is not to propose a proven direction, but to provide a structured way to design, observe, and test GenAI supported learning while keeping learner agency, reflection, progress, and human connection at the center.

CONCLUSION

This SR examined how GenAI is being studied in higher education through a psychological and motivational lens. Based on a final corpus of 195 studies, the findings show rapid growth in research on motivation, perception, engagement, creativity, self-regulation, and learner experience. However, the evidence remains uneven. Literature is dominated by ChatGPT, general GenAI systems, surveys, and short-

term studies, while experimental, longitudinal, multimodal, comparative, and learning analytics research remains limited. The review makes three main contributions. First, it provides a broader and more transparent synthesis through expanded searching, dual screening, reliability assessment, structured coding, and a final audited corpus. Second, it maps the evidence across psychological constructs, GenAI tools, study designs, contexts, outcomes, SDT-related needs, equity concerns, and evidence strength. Third, it proposes the GenAI engagement framework as a literature informed structure comprising initiation, autonomy and curiosity activation, sustained and deep engagement, iterative feedback and reflection, and collaboration or relatedness. The framework should not be interpreted as a validated causal sequence. Its purpose is to organize the available evidence and provide a practical basis for future research and instructional design. By linking GenAI supported learning with autonomy, competence, and relatedness, it highlights the importance of learner choice, meaningful feedback, reflection, and human interaction. The reviewed evidence suggests that the educational value of GenAI depends substantially on how it is designed, guided, and integrated into learning. Thoughtful use may support agency, creativity, feedback, and collaboration, but these benefits must be balanced against overreliance, shallow engagement, misinformation, academic integrity concerns, bias, privacy, and unequal access. Therefore, future research should move beyond early adoption and perception studies toward stronger, longer-term, and context sensitive evidence on how GenAI can support responsible and psychologically meaningful learning in higher education.

Author contributions: **EDLE:** conceptualization, funding acquisition, methodology, supervision; **SMSB:** data curation, formal analysis, visualization, writing – original draft, writing – review & editing. Both authors approved the final version of the article.

Funding: This article was funded by Zayed University under the Research Incentive Fund (RIF R22013).

Acknowledgments: The authors would like to thank Zayed University for supporting this research.

Ethics declaration: Ethical approval and informed consent were not required because this study reviewed publicly available published literature.

AI statement: Generative AI tools were used only to support language refinement and formatting. All content was reviewed and verified by the authors, who take full responsibility for the manuscript.

Declaration of interest: The authors declared no competing interest.

Data availability: Data generated or analyzed during this study are available from the authors on request.

REFERENCES

- Abbas, M., Jam, F. A., & Khan, T. I. (2024). Is it harmful or helpful? Examining the causes and consequences of generative AI usage among university students. *International Journal of Educational Technology in Higher Education*, 21, Article 10. <https://doi.org/10.1186/s41239-024-00444-7>
- Arksey, H., & O'Malley, L. (2005). Scoping studies: Towards a methodological framework. *International Journal of Social Research Methodology*, 8(1), 19-32. <https://doi.org/10.1080/1364557032000119616>
- Baidoo-Anu, D., & Owusu Ansah, L. (2023). Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. *Journal of AI*, 7(1), 52-62. <https://doi.org/10.61969/jai.1337500>
- Bannister, P., Santamaría Urbieta, A., & Alcalde Peñalver, E. (2023). A systematic review of generative AI and (English-medium-instruction higher education). *Aula Abierta*, 52(4), 401-409. <https://doi.org/10.17811/rifie.52.4.2023.401-409>
- Chan, C. K. Y., & Hu, W. (2023). Students' voices on generative AI: Perceptions, benefits, and challenges in higher education. *International Journal of Educational Technology in Higher Education*, 20, Article 43. <https://doi.org/10.1186/s41239-023-00411-8>
- Collie, R. J., & Martin, A. J. (2024). Teachers' motivation and engagement to harness generative AI for teaching and learning: The role of contextual, occupational, and background factors. *Computers and Education: Artificial Intelligence*, 6, Article 100224. <https://doi.org/10.1016/j.caeai.2024.100224>
- Deci, E. L., & Ryan, R. M. (2012). Self-determination theory. In P. A. M. Van Lange, A. W. Kruglanski, & E. T. Higgins (Eds.), *Handbook of theories of social psychology* (pp. 416-437). SAGE. <https://doi.org/10.4135/9781446249215.n21>
- Dempere, J., Modugu, K., Hesham, A., & Ramasamy, L. K. (2023). The impact of ChatGPT on higher education. *Frontiers in Education*, 8. <https://doi.org/10.3389/feduc.2023.1206936>

- Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of Educational Research*, 74(1), 59-109. <https://doi.org/10.3102/00346543074001059>
- Garcia, M. B. (2025). The paradox of artificial creativity: Challenges and opportunities of generative AI artistry. *Creativity Research Journal*, 37(4), 755-768. <https://doi.org/10.1080/10400419.2024.2354622>
- Ivanov, S., Soliman, M., Tuomi, A., Alkathiri, N. A., & Al-Alawi, A. N. (2024). Drivers of generative AI adoption in higher education through the lens of the theory of planned behaviour. *Technology in Society*, 77, Article 102521. <https://doi.org/10.1016/j.techsoc.2024.102521>
- Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., . . . Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103, Article 102274. <https://doi.org/10.1016/j.lindif.2023.102274>
- Kazu, İ. Y., & Kuvvetli, M. (2023). A triangulation method on the effectiveness of digital game-based language learning for vocabulary acquisition. *Education and Information Technologies*, 28, 13541-13567. <https://doi.org/10.1007/s10639-023-11756-y>
- Kirkwood, A., & Price, L. (2014). Technology-enhanced learning and teaching in higher education: What is 'enhanced' and how do we know? A critical literature review. *Learning, Media and Technology*, 39(1), 6-36. <https://doi.org/10.1080/17439884.2013.770404>
- Kumar, S., Rao, P., Singhania, S., Verma, S., & Kheterpal, M. (2024). Will artificial intelligence drive the advancements in higher education? A tri-phased exploration. *Technological Forecasting and Social Change*, 201, Article 123258. <https://doi.org/10.1016/j.techfore.2024.123258>
- Lo, C. K., Hew, K. F., & Jong, M. S.-Y. (2024). The influence of ChatGPT on student engagement: A systematic review and future research agenda. *Computers & Education*, 219, Article 105100. <https://doi.org/10.1016/j.compedu.2024.105100>
- Omran Zailuddin, M. F. N., Nik Harun, N. A., Abdul Rahim, H. A., Kamaruzaman, A. F., Berahim, M. H., Harun, M. H., & Ibrahim, Y. (2024). Redefining creative education: A case study analysis of AI in design courses. *Journal of Research in Innovative Teaching & Learning*, 17(2), 282-296. <https://doi.org/10.1108/JRIT-01-2024-0019>
- Panadero, E. (2017). A review of self-regulated learning: Six models and four directions for research. *Frontiers in Psychology*, 8. <https://doi.org/10.3389/fpsyg.2017.00422>
- Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., & Sutskever, I. (2021). Zero-shot text-to-image generation. In M. Meila, & T. Zhang (Eds.), *Proceedings of the 38th International Conference on Machine Learning* (pp. 8821-8831).
- Runco, M. A. (2014). *Creativity: Theories and themes: Research, development, and practice* (2nd ed.). Elsevier.
- Saúde, S., Barros, J. P., & Almeida, I. (2024). Impacts of generative artificial intelligence in higher education: Research trends and students' perceptions. *Social Sciences*, 13(8), Article 410. <https://doi.org/10.3390/socsci13080410>
- Sok, S., & Heng, K. (2024). Opportunities, challenges, and strategies for using ChatGPT in higher education: A literature review. *Journal of Digital Educational Technology*, 4(1), Article ep2401. <https://doi.org/10.30935/jdet/14027>
- Tricco, A. C., Lillie, E., Zarin, W., O'Brien, K. K., Colquhoun, H., Levac, D., Moher, D., Peters, M. D. J., Horsley, T., Weeks, L., Hempel, S., Akl, E. A., Chang, C., McGowan, J., Stewart, L., Hartling, L., Aldcroft, A., Wilson, M. G., Garritty, C., . . . Straus, S. E. (2018). PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation. *Annals of Internal Medicine*, 169(7), 467-473. <https://doi.org/10.7326/M18-0850>
- Zhai, X. (2023). ChatGPT for next generation science learning. *XRDS: Crossroads, The ACM Magazine for Students*, 29(3), 42-46. <https://doi.org/10.1145/3589649>
- Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. *Theory Into Practice*, 41(2), 64-70. https://doi.org/10.1207/S15430421TIP4102_2

APPENDIX A: COMPLETE SEARCH STRATEGY AND API QUERIES

Search Overview

The literature search was conducted in two stages. The original search was performed in May 2025 using Elicit and covered five broad thematic areas:

- (1) motivation and engagement,
- (2) classroom uses of ChatGPT and GenAI,
- (3) creativity,
- (4) practical learning applications, and
- (5) student perceptions and attitudes.

The thematic areas were preserved in the review records, but the exact wording of the original Elicit prompts was not archived and therefore could not be rebuilt reliably. To broaden coverage and improve search transparency, a supplementary API-based search was conducted on 21 June 2026 using OpenAlex, CrossRef, and Semantic Scholar. The source adapted free text queries were used because the three platforms differ in their search structure and support parameters. The supplementary searches covered motivation and engagement, SDT-related needs, creativity and multimodal GenAI, equity and accessibility, adverse outcomes, comparisons with conventional teaching, learning analytics and conference-related signals, and disciplinary variation. The search was limited to records published between 2021 and 2025. No explicit language filter was applied at the API level. The post-retrieval processing required records to contain both GenAI/LLM terminology and education-related terminology in the title or abstract before they entered the screening pool.

OpenAlex Search

- **Search date:** 21 June 2026
- **Endpoint:** <https://api.openalex.org/works>
- Eight free text queries were submitted separately

Table A1. OpenAlex search queries

ID	Exact query
Q1	generative artificial intelligence student motivation engagement university
Q2	ChatGPT self-determination theory autonomy competence relatedness higher education
Q3	generative AI creativity visual multimodal university students
Q4	generative artificial intelligence equity accessibility inclusion university
Q5	generative AI adverse negative outcomes overreliance plagiarism university students
Q6	generative AI compared traditional pedagogy control group university
Q7	generative AI AIED learning analytics student engagement conference
Q8	generative AI student perceptions discipline STEM humanities university

Table A2. OpenAlex parameters

Parameter	Setting
Search field	search=<query>
Publication start date	from_publication_date:2021-01-01
Publication end date	to_publication_date:2025-12-31
Publication types	article, review, preprint
Records requested per query	per_page=50
Pagination	Not used

Therefore, a maximum of 50 records were retrieved for each query. Following retrieval, publication year was rechecked and the common relevance heuristic was applied before records were merged and de-duplicated. Raw per query hit counts were not retained in the final archived search summary. In the final supplementary API pool, 194 records identified OpenAlex as one of their source platforms, including records overlapping with CrossRef or Semantic Scholar.

CrossRef Search

- **Search date:** 21 June 2026
- **Endpoint:** <https://api.crossref.org/works>
- Five CrossRef queries were used.

Table A3. CrossRef search queries

ID	Exact query
CR1	generative artificial intelligence student motivation engagement university
CR2	ChatGPT self-determination autonomy competence relatedness education
CR3	generative AI equity accessibility higher education
CR4	ChatGPT negative effects academic integrity overreliance students
CR5	generative AI creativity multimodal DALL-E education

Table A4. CrossRef parameters

Parameter	Setting
Publication start date	from-pub-date:2021
Publication end date	until-pub-date:2025
Publication type	journal-article
Records requested per query	rows=40
Selected fields	DOI, title, author, published, container-title, is-referenced-by-count, abstract, URL, link
Pagination	Not used

The retrieved records were subjected to the same publication year and relevance checks used for the other supplementary sources before being merged into the API pool. Raw per query hit counts were not preserved in the final archived search summary. In the final supplementary pool, 99 records identified CrossRef as one of their retrieval sources, including overlaps with other APIs.

Semantic Scholar Search

- **Search date:** 21 June 2026
- **Endpoint:** <https://api.semanticscholar.org/graph/v1/paper/search>
- Eight Semantic Scholar queries were used.

Table A5. Semantic Scholar search queries

ID	Exact query
Q1	generative AI ChatGPT student motivation engagement higher education
Q2	generative AI self-determination theory autonomy competence relatedness students
Q3	generative AI creativity DALL-E multimodal higher education learning
Q4	generative AI equity accessibility higher education digital divide
Q5	ChatGPT negative effects overreliance academic integrity higher education students
Q6	ChatGPT compared traditional teaching feedback, higher education randomized
Q7	generative AI artificial intelligence in education learning analytics student engagement
Q8	generative AI student perceptions discipline STEM humanities higher education

Table A6. Semantic Scholar parameters

Parameter	Setting
Records requested per query	limit=50
Publication filter	None applied at API level
Requested fields	paperId, title, authors, year, abstract, externalIds, venue, citationCount, url, openAccessPdf, isOpenAccess, publicationTypes, journal
Pagination	Not used
Delay between requests	3.5 seconds
Maximum retries	4
HTTP 429 handling	Exponential backoff
Failed requests	Returned an empty result set after retry failure
Records requested per query	limit=50

Semantic Scholar retrieval was affected by API rate limiting. Therefore, a dedicated retry script was used to repeat Semantic Scholar retrieval, adding 88 records during the retry stage. After final merging and deduplication, 94 records in the supplementary pool identified Semantic Scholar as a retrieval source.

Common Post-Retrieval Processing

Records retrieved from OpenAlex, CrossRef, and Semantic Scholar underwent common processing before screening. First, publication years were checked against the predefined eligibility window of 2021-2025.

A post retrieval relevance heuristic was then applied. Records were retained only when their title or abstract contained evidence of both:

- GenAI, ChatGPT, large language model, or closely related terminology
- Education or higher education related terminology

The heuristic was used to remove clearly irrelevant API results before human screening. It did not make final inclusion or exclusion decisions. Duplicate records within supplementary API pool were identified using:

- Normalized DOI matching where a DOI was available
- Normalized exact-title matching when no DOI was available

Letter case, punctuation, spacing, and DOI formatting were standardized before comparison. The fuzzy title matching was not used. After post-retrieval filtering and within-API deduplication, the supplementary search produced 375 unique candidate records.

Table A7. Record flow before screening

Stage	Records, n
Original Elicit records after within-source deduplication	59
Supplementary API records after filtering and deduplication	375
Total records identified before cross-source deduplication	434
Cross-source duplicates or version pairs removed	20
Records retained for dual screening	414

Integration With the Original Search

The supplementary API pool was merged with the original Elicit corpus before screening. Of the 375 API records, 18 were already represented in the Elicit corpus (17 through exact DOI matches and one through a title match between a preprint and its peer reviewed journal version). The 17 DOI matched records were removed during cross-source deduplication. The title matched preprint was removed as one of three preprint publication version pairs identified during quality control; the other two version pairs occurred entirely within the API pool. Therefore, 20 records were removed in total before screening (17 cross-source duplicates and three version pairs), leaving 414 records for dual screening.

Human Screening and Search Boundaries

API retrieval and automated processing were used only to identify and organize candidate records. The final inclusion and exclusion decisions were made by human reviewers using the shared eligibility criteria described in the main manuscript. The manual checking was used when necessary to verify publication status, DOI, venue, publication date, and higher education relevance when indexed metadata were incomplete or ambiguous. The manual checking was not treated as an additional independent database search or as systematic backward or forward citation chasing. Scopus, Web of Science, ERIC, PsycINFO, and systematic searches of AIED and LAK conference proceedings were not conducted.

Search Reproducibility and Known Limitations

Several limitations of the archived search record should be noted. The exact wording of the original May 2025 Elicit prompts was not preserved, only the five thematic search areas remain available. The raw hit counts for individual OpenAlex, CrossRef, and Semantic Scholar queries were also not retained in the final archived search summary. In addition, the supplementary API searches did not use pagination. OpenAlex and Semantic Scholar retrieval were capped at 50 records per query, while CrossRef retrieval was capped at 40 records per query. Therefore, supplementary searches should be interpreted as structured search expansion rather than exhaustive retrieval of every record indexed by each platform.

Semantic Scholar retrieval was additionally affected by rate limiting and required a separate retry procedure. No exact time of day was archived for the supplementary search, although the search date of 21 June 2026 was preserved. No prospective protocol registration was identified for the review. These limitations are also reported before in the main manuscript.

