Co-funded by the European Union

ADMIT

Use of LLM tools within higher education: Report 2

EADTU
JYVÄSKYLÄN YLIOPISTO UNIVERSITY OF JYVÄSKYLÄ
UNED
HELLENIC OPEN UNIVERSITY
ANADOLU UNIVERSITY
FernUniversität in Hagen
Open Universiteit
Universitat Oberta de Catalunya
FernUni.ch UniDistance.ch
UNIVERSITÀ TELEMATICA INTERNAZIONALE UNINETTUNO
The Open University
OPEN UNIVERSITY OF CYPRUS www.ouc.ac.cy

The research reported in this document was carried out in the ADMIT EU funded project. Proposal ID 101134520 under the call ERASMUS-2023-PCOOP-ENGO

Citation:

Bektik, D., Edwards, C., Whitelock, D.& Antonaci, A.(2025). Use of LLM tools within higher education: Report 2 (Research Report No. 2.2). Zenodo.DOI: 10.5281/zenodo.17747509

Co-funded by the European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them.

Executive Summary

This systematic literature review synthesises 147 peer-reviewed studies published between October 2024 and July 2025 on the use of Large Language Models (LLMs) in higher education. Building on the earlier D2.1 report (Bektik et al., 2024), which provided an early-stage snapshot of potential benefits and emerging concerns, this deliverable consolidates and extends the evidence base. Compared with D2.1, it examines a larger corpus (147 vs. 112 studies), provides clearer thematic organisation aligned to three research questions, and integrates new insights on institutional policy, inclusion, and emerging practices.

The review followed a PRISMA-guided process, which produced a dataset of 539 records initially identified from ACM Digital Library, EBSCO, IEEE Xplore, Scopus, and Web of Science. After screening and de-duplication, 147 articles met our inclusion criteria (empirical, conceptual, and review studies addressing LLMs in higher education). Thematic categories did not emerge inductively but were structured in alignment with the research questions. Within each question, however, the synthesis of evidence generated sub-themes that reflect patterns across the literature. This ensures that findings are both systematic and directly responsive to the guiding RQs.

Key developments since D2.1 include:

- **A shift from alarm to adaptation**: universities are no longer debating whether to acknowledge LLMs but are beginning to redesign assessment and pedagogy around them.
- **Growth in policy responses**: institutional guidelines have begun to appear, though many remain reactive or vague, confirming D2.1 predictions but also revealing gaps in implementation.
- **Expanded attention to equity**: while still underdeveloped, there is greater recognition of accessibility, linguistic diversity, and inclusion as critical issues.
- **Broader methodological scope**: the literature includes more empirical work than in 2024, though it is still dominated by small-scale and exploratory studies.

Findings overview
Across the 147 studies, AI's most visible impact is in assessment and academic integrity (61 papers), followed by curriculum and pedagogy (24) and institutional policy (19). Several studies show how LLMs have driven redesigns of exams and coursework (Agostini & Picasso, 2024; Arum et al., 2025). Others highlight the need for staff development, transparency, and frameworks for academic integrity (Mariyono, 2025). LLMs also show potential for enhancing writing, tutoring, and feedback, but risks remain: over-reliance, AI "hallucinations," and low AI literacy continue to undermine critical thinking. Equity-focused research suggests that LLMs could benefit multilingual learners and students with disabilities, yet current models remain limited in linguistic scope and often reproduce social bias. Digital divides in access are a persistent concern.

Top 5 Key Insights from the Review (Oct 2024–July 2025)

Insight	Summary	Change
1. **LLMs are becoming embedded in learning practice**	Educators and students are using LLMs for tutoring, feedback, and content creation — particularly in writing and STEM tasks — with mixed outcomes.	In D2.1 use was tentative; now studies show everyday integration.
2. **Academic integrity and critical thinking remain vulnerable**	There is growing concern that LLM use may encourage over-reliance, reduce independent thought, and challenge conventional assessment formats.	In D2.1 integrity risks were speculative; now evidence shows concrete cases of AI-enabled cheating and redesign responses.
3. **Equity and inclusion are underexplored but critical**	LLMs could support underserved learners, but most research lacks attention to accessibility, bias, and the digital divide.	D2.1 called for equity research; this review shows only modest progress (16/147 studies), confirming the gap.
4. **Policy responses are emerging but inconsistent**	Institutions are moving from AI bans to guidelines, but few policies are co-created with users or formally evaluated for impact.	Policies predicted in D2.1 are now appearing, but they remain uneven, reactive, and rarely co-designed with staff/students.
5. **The research base is growing, but still limited**	There is a visible increase in empirical studies since D2.1 (Bektik et al., 2024), but the field remains dominated by short-term, exploratory work and lacks cross-cultural perspectives.	Compared to D2.1, the field has grown by 31% but still lacks longitudinal and cross-cultural studies.

Together, these developments suggest that higher education is moving from a phase of **speculation and concern** (2022–2024) into one of **adaptation and cautious integration** (2024–2025).

The comparison with last year shows clear movement: institutions are more pragmatic, researchers more empirical, and policy frameworks beginning to materialise. Yet progress remains uneven. Academic integrity, AI literacy, inclusion, and long-term outcomes remain pressing challenges. The literature continues to expand rapidly, but robust, cross-disciplinary, and cross-cultural studies are still scarce.

Table of contents

Background

The main objective of this deliverable is to explore the latest *published developments* in research and innovation relating to the use of Large Language Models (LLMs) in higher education. Furthermore, it covers research, innovation and innovative *practices* related to evolving opportunities and challenges of LLMs and their *implications for the design* of teaching and learning in higher education. It also assesses the implications of LLMs for diversity, inclusion, and accessibility within higher education, identifying challenges associated with these tools and highlighting good practices. Finally, it investigates the institutional policies and strategies related to the use of LLMs in higher education, identifying key components for developing educational and ethical frameworks and guidelines for best practice and implementation.

Research Questions

This review examines the three following Research Questions, which remain the same as for the previous report:

1. *What opportunities and challenges do LLMs present for teaching and learning practices in higher education?*
2. *How do LLMs impact diversity, inclusion, and accessibility in higher education?*
3. *What guidelines and institutional policies are being established to ensure the responsible and effective use of LLMs in higher education?*

These research questions aim to explore the multifaceted implications of LLMs in higher education, focusing on innovation, diversity and inclusion, and institutional policies. They provide a comprehensive framework for investigating how LLMs can be effectively and ethically integrated into higher education to enhance teaching and learning practices.

Methods

Following up on Bektik, D. et al (2024), a search strategy for this report has been revised and it integrates an improved query based on advisory board comments and partner feedback. The new query expands the generative AI set (Claude, Cohere, ERNIE Bot, Gemini, LLaMA, Mistral, PaLM, and Vicuna), includes emerging concepts such as multimodal AI, AI copilots, instruction tuned and agentic AI, and adds educational/impact related terms like AI literacy, adaptive learning, educational equity and learning design. Wildcards, truncation (*) are used to capture word variations (e.g., *educat* retrieves education, educational, educating). Searches were conducted across five major bibliographic databases: EBSCO Education Source, Web of Science, Scopus, ACM Digital Library, and IEEE Xplore. The final search took place on 30 July 2025.

All references analysed in this review were retrieved directly from these academic databases. Each record included complete bibliographic metadata (title, author, year, source, DOI/ISBN). Because our corpus was built exclusively from established bibliographic databases, the risk of fabricated or “hallucinated” references identified in recent discussions of generative AI and scholarly publishing (Haider et al., 2024) does not apply in this case.

See below, Table 1, that summarises for what is used in the search query for each concept group based on the above-mentioned adaptations.

Table 1 Search query terms for each concept group

Concept group	Example terms (combine with OR within group)
Generative AI & models	"generative AI" OR "genAI" OR "generative artificial intelligence" OR "large language model*" OR LLM OR "generative pre-trained transformer" OR "foundation model*" OR "transformer model*" OR "multi-modal AI" OR "instruction-tuned model*" OR "agentic AI" (agentic AI is an AI system that plans, acts and learns autonomously) OR "AI copilot*" OR "AI agent*" OR "AI assistant*" OR chatbot* OR "AI writing tool*" OR "AI tutor" OR "intelligent tutor*" OR "automated feedback" OR "intelligent tutoring system*" OR "AI-powered tool*" OR "AI-powered assistant*".
Named AI platforms / tools	ChatGPT OR GPT OR GPT-4 OR GPT-3.5 OR Gemini OR Bard OR PaLM OR "Google Gemini" OR "Gemini Pro" OR "Gemini 2" OR Claude OR "Claude 3" OR "Claude 2" OR LLaMA OR LLaMA 2 OR LLaMA 3 OR LLaMA 4 OR Mistral OR Cohere OR Vicuna OR Alpaca OR ERNIE Bot OR OpenAssistant OR Grok OR Perplexity OR Copilot OR Anthropic OR "OpenAI" OR "Amazon Bedrock".
Educational setting	"higher education" OR "tertiary education" OR universit* OR college* OR "postsecondary education" OR "academic institution*" OR HEI* OR campus OR "post-secondary" OR "undergraduate" OR "graduate" OR "degree program" OR STEM OR humanities OR "social sciences".
Stakeholders / participants	student* OR educator* OR teacher* OR professor* OR faculty OR instructor* OR lecturer* OR academic staff OR administrator*.
Pedagogical & contextual concepts	teach* OR learn* OR educat* OR instruct* OR pedagogy OR "course design" OR curriculum OR "learning design" OR "teaching design" OR assessment OR "adaptive learning" OR "AI literacy" OR "digital literacy" OR "learning outcome*" OR "academic performance" OR "student engagement" OR "adaptive learning" OR "personalized learning" OR "responsive AI" OR "automated feedback" OR "learning analytics" OR "adaptive system*".

Inclusion, equity, ethics & integrity	inclusi* OR divers* OR equit* OR accessibility OR “educational equity” OR bias OR fairness OR “AI ethics” OR ethic* OR “ethical implication*” OR “responsible AI” OR privacy OR governance OR policy OR “AI governance” OR “AI policy” OR “academic integrity” OR plagiarism OR cheating.
Impact / perceptions	perception* OR attitude* OR adoption OR accept* OR policy OR implication* OR outcome*.

And here is the general search string, which is adapted for each database

(("generative AI" OR genAI OR "generative artificial intelligence" OR "large language model*" OR LLM OR "foundation model*" OR "transformer model*" OR "multi-modal AI" OR "instruction-tuned model*" OR "agentic AI" OR "AI copilot*" OR "AI agent*" OR "AI assistant*" OR chatbot* OR "AI writing tool*" OR "AI tutor" OR "intelligent tutoring system*" OR "AI-powered tool*" OR "AI-powered assistant*")

OR

(ChatGPT OR GPT OR GPT-4 OR GPT-3.5 OR Gemini OR Bard OR PaLM OR Claude OR "Claude 3" OR "Claude 2" OR LLaMA OR "LLaMA 2" OR "LLaMA 3" OR "LLaMA 4" OR Mistral OR Cohere OR Vicuna OR Alpaca OR "ERNIE Bot" OR OpenAssistant OR Grok OR Perplexity OR Copilot OR Anthropic OR OpenAI OR "Amazon Bedrock"))

AND

("higher education" OR "tertiary education" OR universit* OR college* OR "postsecondary education" OR "post-secondary education" OR "academic institution*" OR HEI* OR campus OR undergraduate OR graduate OR STEM OR humanities OR "social sciences")

AND

(inclusi* OR divers* OR equit* OR accessibility OR "educational equity" OR bias OR fairness OR ethic* OR "ethical implication*" OR "AI ethics" OR "responsible AI" OR privacy OR governance OR policy OR "AI policy" OR "AI governance" OR "academic integrity" OR plagiarism OR cheating OR "AI literacy" OR "adaptive learning" OR "learning outcome*" OR "academic performance" OR "student engagement" OR "learning design" OR assessment OR curriculum OR "personalized learning")

A structured rapid literature review approach (Smela et al., 2023) was employed to identify relevant studies, involving a search of peer-reviewed literature databases. The initial methodology for these reports were initially based on Rapid Evidence Assessment (REA) to account for the anticipated limited number of peer-reviewed publications due to the rapid adoption of the technology. However, as peer-reviewed publications proliferated beyond initial expectations, and increased massively in numbers especially for this reporting period, a systematic literature review was preferred over the REA approach to comprehensively capture the considerable volume of available articles. This review followed the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) statement (Moher et al, 2009) when selecting relevant articles, see Figure 1 below. The final search was conducted on 30 July 2025. Searches were carried out across five major databases—EBSCO Education Source, Web of Science, Scopus, ACM Digital Library, and IEEE Xplore—using adapted search queries tailored to each platform.

Academic articles published between 1 October 2024 and 31 July 2025 (the date of the final search) were reviewed. Non-academic articles, i.e. grey literature, (e.g., articles from mass media) were not included this round due to available increased number of returned articles. The inclusion criteria were that articles had to discuss GenAI in the field of higher education, with no constraints on any specific educational contexts. The 33 literature review papers identified were used as background references. In addition, only English-language articles were included in this review. Table 2 summarises the inclusion and exclusion criteria for article selection.

Table 2 Inclusion and exclusion criteria for studies selection.

Criteria	**Inclusion**	**Exclusion**
Subject	Discuss GenAI in the field of higher education	Do not discuss GenAI in the field of higher education (i.e. school settings, primary/secondary etc.), workplace etc.
Article type	Academic, peer reviewed articles	Social media & Grey Literature (e.g. blogposts, news articles, websites)
Time period	1 October 2024 and 31 July 2025	Articles outside the time period
Language	English	Non-English

Figure 1 that represents the PRISMA flow diagram summarises the identification, screening, exclusion and inclusion decisions made during systematic search. The counts reflect the numbers reported for each database, the duplicates removed, and the records excluded using inclusion–exclusion criteria (non higher education contexts or outside the Oct 2024–Jul 2025 window).

Key numbers used in the diagram

- **Records identified via database searching (n = 539):**
  - EBSCO (25), IEEE Xplore (272), Scopus (150), Web of Science (72), ACM (20).
- **Duplicates removed (n = 48):** After combining all sources, 48 duplicate records (based on title and DOI) were removed, leaving **491** unique records for screening.
- **Records screened by title/abstract (n = 491):** These were assessed against inclusion criteria.
- **Records excluded (n = 27):** Eleven records were about K–12 or industry/workplace contexts rather than higher education, and sixteen had missing or out of range publication dates. These were removed.
- **Records included for qualitative synthesis (n = 464):** The remaining papers (Oct 2024–Jul 2025, peer reviewed, English language, focused on generative AI in higher/tertiary education) proceed to qualitative analysis.

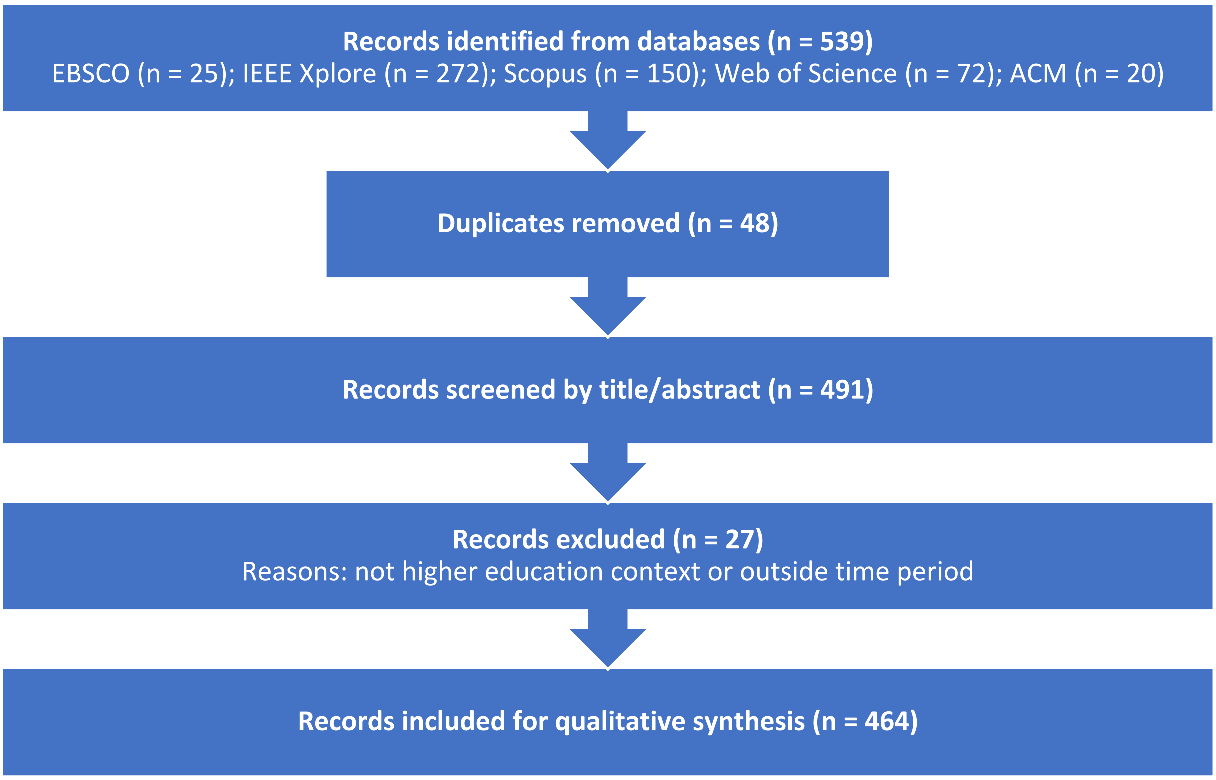

Figure 1-Initial selection of items for Systematic Reviews and Meta-Analysis (PRISMA) flow diagram of article selection

Selection process

1. **Identification (n = 539)** – Results were combined from five databases: EBSCO (25 records), IEEE Xplore (272), Scopus (150), Web of Science (72) and ACM Digital Library (20), for a total of 539 records.

2. **Duplicate removal (n = 48)** – Titles and DOIs were normalised to detect identical references across databases. Forty-eight duplicate records were removed, leaving 491 unique entries for screening.
3. **First screening (n = 491)** – Titles and abstracts were assessed against the initial inclusion/exclusion criteria (focus on generative AI in higher/tertiary education, English language, publication between 1 Oct 2024 and 31 Jul 2025). Eleven papers were excluded because they addressed K12 contexts, industry applications or other non HE settings, and sixteen had missing or outof-range publication dates. That left 464 records.
4. **Second screening for AI & higher education context (n = 464)** – A keyword based heuristic was applied to the remaining titles and abstracts. We required evidence of both (a) generative AI–related terms (e.g., “generative AI,” “large language model,” “ChatGPT”) and (b) a higher education context (e.g., “university,” “college,” “higher education”). Entries just mentioning K12 or industry settings, or lacking clear AI or higher education keywords, were excluded. This removed 276 more records, leaving 188 for qualitative synthesis. After deduplicating titles again (some abstracts appeared in multiple databases), 168 unique papers remained.
5. **Full text availability (n=147)** – Full texts for 168 unique papers were downloaded for further analysis, yet 21 were not available at the time which made the sample size 147 papers.

The updated flow diagram, see figure 2, captures these steps, showing how the initial pool of 539 records was narrowed to 147 unique studies that clearly discuss generative AI within higher education contexts during the specified timeframe.

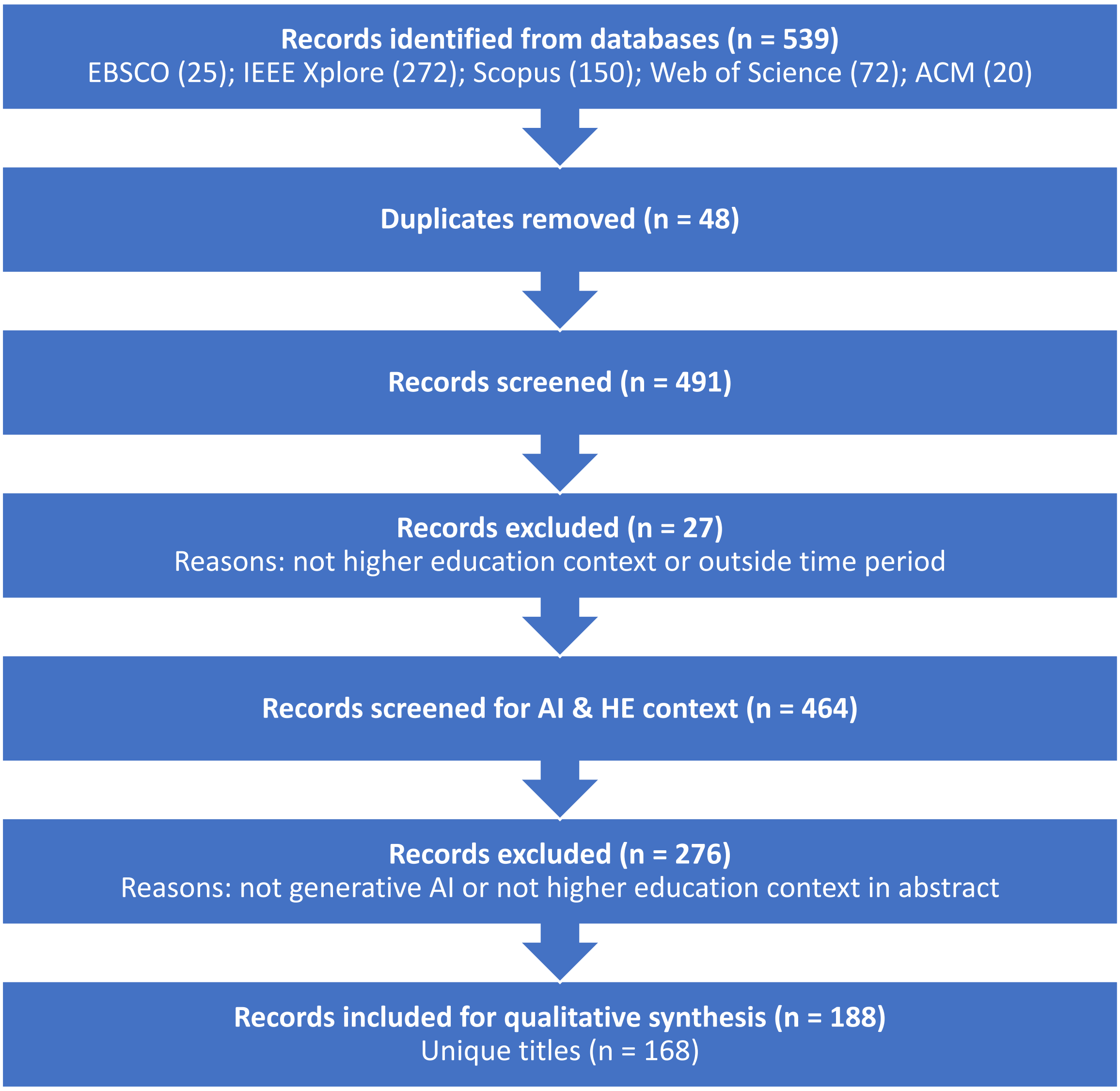

Figure 2-Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) flow diagram of article selection

Introduction

The emergence of large language models (LLMs) such as OpenAI's GPT series (e.g. ChatGPT) have rapidly transformed the landscape of higher education since 2022. Educators and researchers are grappling with how these generative AI tools can be leveraged to enhance teaching and learning, while also mitigating potential risks. On the one hand, LLMs offer unprecedented opportunities for personalised tutoring, content generation, and efficiency in academic tasks. On the other hand, they present new challenges related to academic integrity (e.g. plagiarism and "AI cheating"), reliability of information, and ethical use. Furthermore, questions have arisen about whether LLMs will benefit or harm goals of diversity, equity, and inclusion in education – for instance, could they *bridge* gaps for learners with disabilities or non-native language speakers, or might they *amplify* biases and widen digital divides? Simultaneously, universities worldwide are beginning to formulate guidelines and policies to discipline the use of AI tools at institutional level, aiming to ensure responsible and effective integration into curricula.

Considering these developments, this report presents a systematic literature review of recent studies (147 articles published in late 2024 to mid 2025) on LLMs' effects in higher education. More specifically the report presents the study characteristics and trends, thematic coding, thematic synthesis, population characteristics and methodological coding. Then the report specifically addresses three key research questions under relevant headings.: **(a)** What opportunities and challenges do LLMs present for teaching and learning practices in higher education? **(b)** How do LLMs impact diversity, inclusion, and accessibility in higher education? **(c)** What guidelines and institutional policies are being established to ensure the responsible and effective use of LLMs in higher education? The goal is to synthesise current findings and perspectives, and to identify prevailing trends, gaps, and recommendations in the scholarly literature. The report then compares the results with our previous review (Bektik et al, 2024) and concludes with gaps and areas for further research.

Study Characteristics and Trends

Geographically, the research is globally distributed, with notable contributions from Asia and Europe, as well as North America (Table 3). About one-third of studies had an explicit international or multi-country scope (e.g. global surveys or collaborative analyses), and around 10% did not specify a particular regional context (often conceptual papers or general reviews). The earlier deliverable (Bektik et al, 2024) did not code study location in a systematic way, so no direct year-on-year comparison is possible. However, this year's findings establish a baseline that can support future trend analyses.

Table 3 Regional distribution of studies (by study context or author affiliation), this report N=147

	Number of studies	
Region context	**This report (N=147)**	**% of total**
Asia (e.g. China, India, Middle East)	35	24%
Europe (e.g. UK, Spain, Germany)	25	17%

North America (USA, Canada)	14	10%
South America (e.g. Ecuador, Chile)	4	3%
Africa (e.g. South Africa, Egypt)	3	2%
Oceania (Australia, etc.)	3	2%
Multiple countries / International	44	30%
Not specified (global or no specific locale)	11	7%

In terms of research methods, about two-thirds of the studies were empirical in nature, while the remaining one-third were conceptual or review papers without new data (see Table 4). Among the empirical works, there was a roughly even split between quantitative, qualitative, and mixed methods approach. Approximately 28% of studies used quantitative designs (e.g. surveys of students or experiments measuring learning outcomes), about 15% were qualitative (e.g. interviews, focus groups, or discourse analyses), and roughly 21% employed explicit mixed-methods (combining surveys or log data with interviews or qualitative coding). The conceptual corpus (36% of studies) included literature reviews, opinion pieces, and framework proposals addressing LLM implications. This diversity in methods reflects the interdisciplinary interest in LLMs – spanning education research, computer science, and ethics – and the field's early exploratory stage. Studies also varied in scale: sample sizes ranged from small qualitative studies with a dozen participants to large surveys of 500+ students. For example, some experimental studies involved a single class of 20–30 students or a handful of instructors, whereas the largest survey (Hussain et al., 2024) gathered responses from 700 university students. Most participant-based studies focused on students (often undergraduates), though several investigated educators' perspectives (e.g. faculty attitudes toward using ChatGPT in teaching), and a few examined both groups. A minority of papers analysed institutional documents or policies rather than surveying individuals (e.g. Christidis et al., 2025, who reviewed university policy documents).

Table 4. Methodological approaches of included studies

Methodology category	**Number of studies (N=147)**
Quantitative (surveys, experiments, analytics)	41
Qualitative (interviews, case studies, content analysis)	22
Mixed-methods (combined quantitative + qualitative)	31
Conceptual/commentary[2] or Literature Review	53

[2] "Conceptual/commentary" includes essays, proposed frameworks, and literature reviews without a primary empirical study. Percentages roughly: ~28% quantitative, ~15% qualitative, ~21% mixed, ~36% conceptual (total exceeds 100% due to rounding).

Another important trend is the LLM tools examined. Unsurprisingly, ChatGPT (OpenAI) was the focal point of most studies – it was explicitly mentioned in the titles or focus of at least half the papers and discussed in nearly all. A handful of studies included or focused on other generative AI systems: for instance, a few technical evaluations compared ChatGPT with Google's Bard and Microsoft's Bing Chat on academic tasks (Williams, 2024), or examined GPT-4 versus earlier models (Nikolic et al., 2024). A small number of articles discussed open-source LLMs like BLOOM or LLaMA, or future systems (e.g. Google's Gemini) in passing, but these are not yet commonly studied in educational settings. By and large, ChatGPT (based on GPT-3.5 and GPT-4) has been the prototype for investigating LLM implications in higher education. This reflects its widespread public availability and impact since late 2022.

Of the 147 included studies, 92 (62%) were empirical, while 55 (38%) were conceptual or review-based. Among the empirical studies, 54 relied on student surveys, with sample sizes ranging from 12 to 648 participants (median 130). A further 21 studies used semi-structured interviews, typically involving 10–25 participants. Fifteen applied mixed methods combining surveys and interviews. Most studies (71) focused exclusively on students, while 13 targeted academic staff, and only 8 included multiple stakeholder groups (students, faculty, administrators). Disciplinary foci were often narrow: 18 studies examined engineering cohorts, 11 focused on health and nursing, and 9 on language/communication fields. Multi-disciplinary or institution-wide samples were rare (n=7).

Finally, the topical coverage of the literature can be mapped to our three research question themes. It is found that nearly all studies address issues related to teaching and learning practices (opportunities or challenges), since this is the broad umbrella of "LLMs in education." A substantial subset also look into questions of policy/guidance, while relatively fewer directly focus on diversity and inclusion aspects. Table 5 provides an overview of how many studies in our sample explicitly engaged with each theme. (Note that many papers span multiple themes; for example, a study might discuss teaching benefits of ChatGPT *and* raise academic integrity policy concerns.)

Table 5. Coverage of major themes in the literature

Theme (Research Question)	Number of studies*	Illustrative references
Opportunities and Challenges for Teaching & Learning	115	e.g., Zhou et al. (2025a); Banerjee et al. (2025); Ahmed et al. (2024); Nikolic et al. (2024); Gadekallu et al. (2025)
Diversity, Inclusion & Accessibility Impacts	16	e.g., Gadekallu et al. (2025); Chedrawi et al. (2025); Ulla et al. (2024)
Guidelines and Institutional Policies for LLM Use	44	e.g., Christidis et al. (2025); Wilson (2025); Dai et al. (2024); Dabis & Csáki (2024); Barus et al. (2025)

**Note: These categories are not mutually exclusive. Many studies address multiple themes, so counts sum to over 147.*

Thematic coding

To make sense of the heterogeneous research questions, a thematic analysis was performed based on the *title*, *research question*, *summary* and *main findings* fields for each paper. We assigned one or more of nine themes based on keyword matching shown in table below. Papers could be coded into multiple themes if they addressed more than one issue.

We used an inductive–deductive thematic coding approach. First, we coded each paper according to its explicit research focus (based on titles, abstracts, and methods). We developed a coding frame iteratively, informed by both the three research questions and the recurring categories found in the literature (e.g., adoption, assessment, policy). Each paper could be assigned to more than one theme where relevant. This process generated nine thematic categories (Table 6), which offer finer-grained resolution than the three overarching research questions.

Table 6. Thematic coding categories and definitions

Theme	Description and typical keywords	Links to RQs
Adoption & perception	Studies examining how students or staff adopt, use or perceive GenAI (keywords: adoption, perception, attitude, intention, usage, engagement).	RQ1 (Opportunities & Challenges for Teaching & Learning)
Teaching & learning	Work exploring GenAI as a pedagogical tool, including prompt engineering, instructional design, and effects on student learning outcomes.	RQ1
Assessment & academic integrity	Research focusing on assessment design, plagiarism/cheating detection, feedback, grading and the impact of GenAI on academic integrity.	RQ1 & RQ3
Policy & governance	Articles analysing institutional policies, guidelines,	RQ3 (Policies & Guidelines)

	frameworks or governance for GenAI in HE.	
Ethics & societal impact	Papers discussing bias, privacy, fairness, security, hallucinations, misuse or societal implications.	RQ2 (Inclusion/Accessibility) & RQ3
Teacher & professional development	Research on educators' experiences, training, professional development or changes in academic roles due to GenAI.	RQ1 & RQ3
Domain specific studies	Studies applying GenAI to particular disciplines (e.g., computing, medicine, law, language) or professional contexts.	RQ1
Student support & well-being	Research on GenAI for advising, mental health support, motivation and well-being.	RQ1 & RQ2
Technology evaluation & performance	Work benchmarking GenAI models, detecting hallucinations, evaluating capabilities, or developing AI tools.	RQ1

These thematic categories provide granularity within the overarching research questions. For example, adoption, teaching & learning, and assessment all fall primarily under RQ1; ethics and student support also speak to RQ2; and policy, governance, and professional development map onto RQ3. This structure allowed us to synthesise findings at both a fine-grained level (themes) and at the higher level of the three guiding research questions.

Compared with Bektik et al 2024's Deliverable D2.1 (Bektik et al., 2024), which grouped studies into five broad themes—teaching & learning, assessment & academic integrity, adoption & perceptions, policy & governance, and ethics—the present review (D2.2) expands the coding to nine categories (Table 6). This reflects the diversification of the research agenda between late 2024 and mid-2025. While the five core themes from D2.1

(Bektik et al., 2024) remain central, four additional categories have emerged in this period: Teacher & Professional Development, Student Support & Well-being, Domain-Specific Studies, and Technology Evaluation & Performance. These new themes capture issues that were only briefly mentioned or absent in the earlier report, such as staff training, well-being and motivation, discipline-specific applications, and benchmarking of LLM models. This evolution highlights both continuity across reviews and the field's shift towards more applied, context-specific investigations.

Table 7. Comparison of thematic coding between Deliverable D2.1 and D2.2

Theme	*Coverage in D2.1 (Oct 2024 cut-off)*	*Coverage in D2.2 (Oct 2024–Jul 2025)*
Adoption & Perception	✓ Focused section on student/educator attitudes and use	✓ Still strong theme; expanded with cultural comparisons
Teaching & Learning	✓ Prominent theme (pedagogical integration, opportunities)	✓ Expanded with experiments, quasi-experiments, domain applications
Assessment & Academic Integrity	✓ Central theme (plagiarism, cheating, exam design)	✓ Remains core; now includes AI detection tools and redesign strategies
Policy & Governance	✓ Institutional responses, policy gaps, guidelines	✓ Still key; more emphasis on governance frameworks and AI literacy workshops
Ethics & Societal Impact	✓ Covered bias, fairness, privacy, sustainability	✓ Continues; expanded on hallucinations, IP, societal trust
Teacher & Professional Development	✕ Mentioned only briefly (educator readiness)	✓ Now distinct theme, many staff interviews and PD programmes

Student Support & Well-being	✗ Absent	✓ New: tutoring, advising, motivation, anxiety, well-being
Domain-Specific Studies	✗ Discussed only in passing (e.g., language, computer science)	✓ Elevated to distinct theme with strong discipline-specific evidence
Technology Evaluation & Performance	✗ Mentioned indirectly (LLM capabilities)	✓ New: benchmarking, hallucination detection, system building

Figure 3 shows the distribution of papers across the thematic categories. The most common themes were Assessment & academic integrity (84 papers), Teaching & learning (71 papers), Technology evaluation & performance (52 papers), and Adoption & perception (48 papers). Substantial numbers also addressed Ethics & societal impact (48 papers), Policy & governance (45 papers), Domain-specific studies (53 papers), Teacher & professional development (27 papers), and Student support & well-being (20 papers). A small number of papers (n = 4, 3%) could not be clearly assigned to any of these nine themes—for example, highly general reviews of AI in education or preliminary policy commentaries. These are grouped under an "Other" category.

Because many papers addressed more than one theme, the counts sum to more than 147.

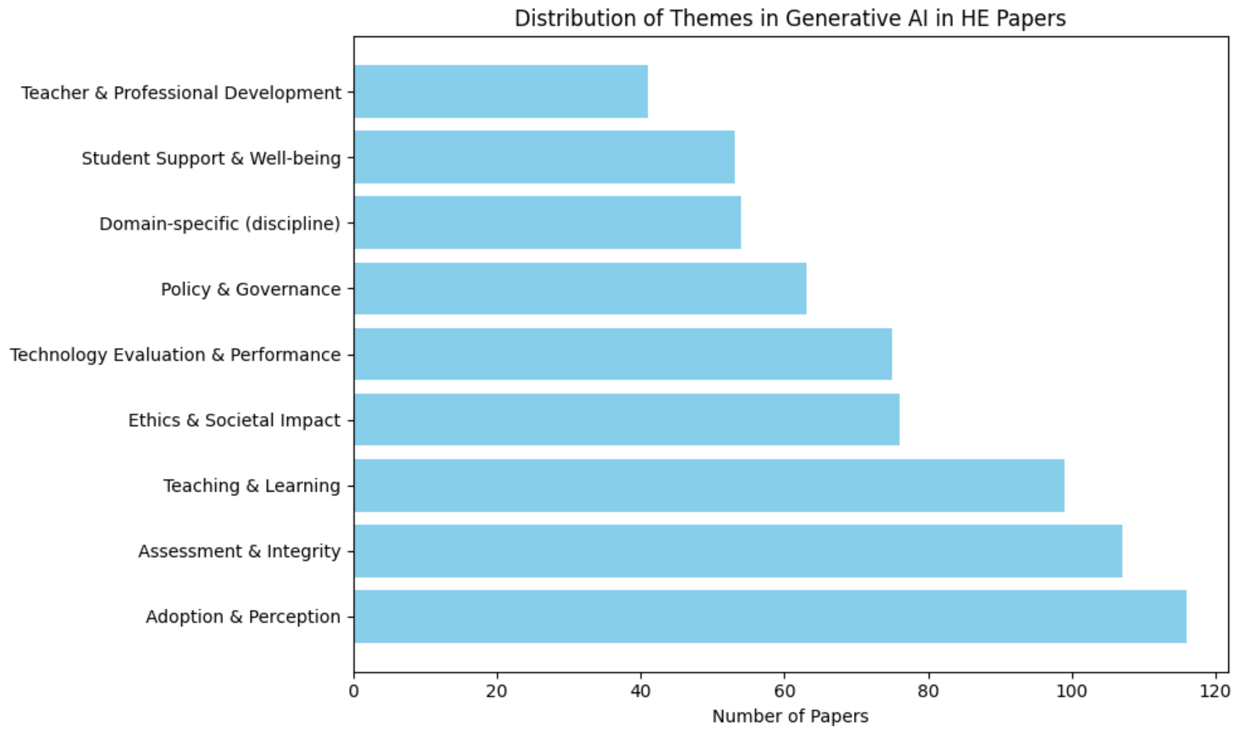

Figure 3. Distribution of Themes Across Papers

Methodological coding of papers reviewed

Each paper's methodology was coded into mutually exclusive categories based on the reported approach (keywords such as survey, experiment, mixed methods, etc.). Many studies reported more than one methodological element (e.g., surveys combined with interviews).

In this review (D2.2), the most common approaches were surveys (n=67, often cross-sectional online questionnaires) and technology development or evaluation studies (n=67, e.g., building chatbots or detection systems). Other substantial categories included qualitative studies (n=64, interviews, focus groups, content analysis), literature or scoping reviews (n=33), and experimental or quasi-experimental designs (n=23). Mixed methods studies (n=20), policy analyses (n=13), and case studies (n=10) also featured, while bibliometric analyses (n=2) and meta-analyses (n=2) were comparatively rare.

Figure 4 visualises these distributions. Compared with the first review (D2.1, Bektik et al, 2024), which reported that most papers were conceptual or descriptive (~70%) with relatively few empirical studies (~30%), this review shows a clear shift towards empirical and applied research designs. While exact methodological coding was not provided in D2.1 (Bektik et al., 2024), the contrast highlights a growing trend towards surveys, qualitative fieldwork, and technology evaluations in the most recent literature. This comparison highlights that while D2.1 (Bektik et al., 2024) was dominated by conceptual reflections, the present review (D2.2) captures a much richer methodological spread, including more empirical studies across multiple designs.

Importantly, however, even within the "empirical" set, sample sizes were often modest, designs were predominantly cross-sectional, and very few studies employed advanced statistical modelling or robust experimental controls. Despite the number of survey-based papers, the field still lacks quantitative depth and statistical sophistication.

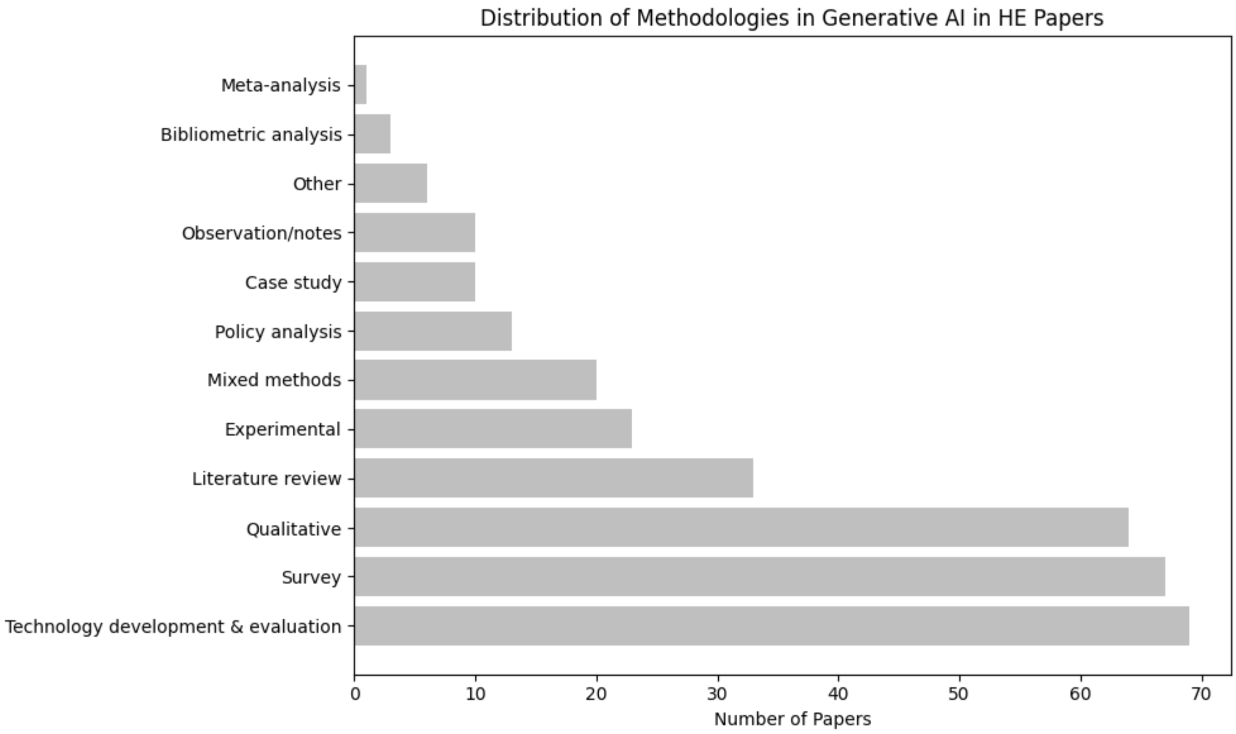

Figure 4[3]. Distribution of Methodologies Across Papers

The methodological distribution also helps us to understand the strength and limitations of evidence underpinning the thematic findings reported in the Results section. For example, ethics and integrity themes were primarily supported by qualitative interviews (n=24) and surveys (n=16), which provide rich but often context-specific insights rather than generalisable conclusions. Assessment-focused papers drew more heavily on technology development and evaluation studies (n=18) and experiments (n=10), reflecting the applied orientation of this theme. Perceptions and adoption studies were dominated by surveys (n=26), often cross-sectional and descriptive, which helps capture attitudes but rarely involves advanced statistical modelling. Inclusion and equity, by contrast, rested largely on small-scale qualitative or case study research (n=7), highlighting the exploratory and underdeveloped nature of this area.

Taken together, these methodological patterns suggest that the literature is still in an exploratory phase. Despite the large number of surveys, relatively few studies employed robust quantitative designs such as controlled experiments, quasi-experiments, or inferential statistical modelling. This helps explain why most reported findings are descriptive (e.g., adoption rates, perceptions, or single-cohort outcomes) rather than predictive or causal. The reliance on small-scale qualitative studies also contributes to valuable depth but limits generalisability. In sum, the quality of evidence across themes is uneven: ethics, assessment, and adoption themes are well populated but methodologically

[3] Note: No equivalent figure was provided in D2.1. The first report grouped studies broadly into "conceptual/descriptive" versus "empirical," estimating ~70% conceptual and ~30% empirical. The comparison here is therefore indicative rather than directly parallel, but it highlights the clear trend toward empirical and applied designs in D2.2

narrow, while inclusion and policy-related themes remain underexplored both in volume and methodological diversity.

Table 8. Methodologies by Theme in Generative AI in Higher Education Studies (n = 147)

Methodology	Ethics & Integrity (n=48)	Assessment (n=84)	Perceptions & Adoption (n=48)	Learning Outcomes (n=28)	Instruction & Curriculum (n=24)	Inclusion & Equity (n=16)	Policy & Governance (n=45)
Survey (n=67)	16	18	26	12	10	4	8
Qualitative (n=64)	24	14	12	14	8	6	6
Tech dev/evaluation (n=67)	8	35	6	6	4	2	4
Literature review (n=33)	6	4	2	2	6	2	12
Experimental/quasi-exp (n=23)	2	16	2	6	2	0	2
Mixed methods (n=20)	2	6	4	4	2	2	3
Policy analysis (n=13)	2	0	0	0	2	1	12
Case study (n=10)	0	2	0	2	1	3	2
Bibliometric/meta (n=5)	0	1	0	0	1	0	3

As Table 8 shows, ethics and integrity were most often investigated through qualitative approaches (24 studies) and surveys (16 studies), providing rich but often context-specific insights. Assessment studies leaned more heavily on technology development/evaluation (35) and experimental designs (16), reflecting the applied orientation of this theme. Perceptions and adoption studies were overwhelmingly survey-based (26 of 48), capturing descriptive attitudes but offering little in the way of causal inference. Inclusion and equity remains the least developed theme (16 studies), drawing mainly on qualitative or case study methods (9 combined). Policy and governance, meanwhile, rests largely on conceptual or documentary analysis, with 12 policy analyses and 12 literature reviews. In sum, while the thematic distribution reflects active experimentation across areas, the evidence base remains methodologically narrow: surveys dominate but lack advanced statistical modelling, qualitative work offers depth but limited generalisability, and large-scale or longitudinal experimental research is largely absent.

This methodological distribution also helps contextualise the thematic findings presented in the next section. For example, the predominance of survey-based studies explains why much of the evidence on adoption and perception relies on self-reported attitudes rather than observed behaviours, while the limited number of experimental or quasi-experimental designs constrains what can be inferred about learning outcomes and assessment quality. Similarly, the small number of policy analyses underscores why the policy and governance theme is often descriptive rather than evaluative.

Results

Adoption & perception (48 papers)

Research on adoption and perception explores how students and staff engage with GenAI and what factors influence their behaviour (Ahmed et al., 2024; Polyportis et al., 2024; Hussain et al., 2024; Xing, 2024; Hsiao et al., 2024; Quezada-Sarmiento et al., 2025; Acosta-Enriquez et al., 2024; Subhani et al., 2025; Klidas et al., 2025; Karkoulian et al., 2024; Tossell et al., 2024; Heil et al., 2025; Chung et al., 2025; Obed et al., 2025; Gao et al., 2025; Chen et al., 2025c; Jin et al., 2024; Moisan et al., 2025; Alghazo et al., 2025; Ma et al., 2024). The majority of studies relied on descriptive cross-sectional surveys (n=54), often informed by frameworks such as the Unified Theory of Acceptance and Use of Technology (UTAUT). These surveys typically measured determinants like performance expectancy, effort expectancy, attitudes, motivation, and social influence, but few employed advanced statistical analyses beyond correlations or regressions. For example, several studies reported that students' intention to use ChatGPT was positively associated with performance expectancy and self-efficacy, and negatively with perceived risk (e.g., Arum et al., 2025). Mixed methods studies (n=15) supplemented survey findings with interviews or focus groups, showing that students valued GenAI as a learning aid for brainstorming, writing support, and information retrieval, but expressed persistent concerns about accuracy, bias, and privacy.

Geographical context also shaped perceptions, but findings come from separate country-based studies rather than direct cross-country comparisons. For instance, research in China (Li et al., 2025a) and Jordan highlighted gaps between familiarity with ChatGPT and effective adoption, pointing to the need for responsible-use training. A study in Peru similarly stressed the importance of institutional guidance. One paper from Bulgaria found that students considered ChatGPT use less ethically acceptable than peers in other reported contexts, although these comparisons were not derived from harmonised datasets. Caution is therefore warranted in interpreting cultural differences: current evidence reflects isolated case studies rather than systematic cross-national analyses.

Overall, adoption studies indicate a generally positive attitude towards GenAI, tempered by ethical concerns and a strong demand for institutional clarity and support.

Teaching & learning (71 papers)

Seventy-one papers examined how GenAI tools support teaching and student learning (Bai et al., 2024; Ahmed et al., 2024; Wang et al., 2024; Banerjee et al., 2025; Qiu, 2024; Hsiao et al., 2024; Yusuf et al., 2024; Wa et al., 2025; Wu et al., 2025; Liu et al., 2025; Karkoulian et al., 2024; Chen et al., 2024; Tossell et al., 2024; Heil et al., 2025; Zhang et al., 2025; Shahzad et al., 2024; Tsz et al., 2024; Amr et al., 2024; Wang et al., 2025; Gadekallu et al., 2025). Experimental and quasi-experimental designs (n=23) integrated ChatGPT into programming courses, language learning, physical education, and nursing education. Most found improvements in problem-solving, writing quality, and student motivation when GenAI was used as a supplement rather than a replacement for traditional instruction. For instance, one study in computer science reported that students who used ChatGPT to debug code performed significantly better than controls (Arum et al., 2025). Case studies (n=10) evaluated prompt-engineering strategies, showing that teacher-guided prompts generated deeper critical engagement than unguided use. Literature reviews (n=33) highlighted common applications such as scaffolding learning activities, generating practice questions, and providing personalised feedback, but cautioned against overreliance. Concerns repeatedly noted included diminished critical thinking and superficial learning, indicating that while pedagogical potential is strong, risks remain if use is unstructured.

Assessment & academic integrity (84 papers)

Academic integrity emerged as a central concern across the dataset (Ahmed et al. (2024); Hussain et al. (2024); Banerjee et al. (2025); Dai et al. (2024); Wilson (2025); Boadu et al. (2025); Campo et al. (2025); Hsiao et al. (2024); Quezada-Sarmiento et al. (2025); Acosta-Enriquez et al. (2024); Yusuf et al. (2024); Subhani et al. (2025); Klidas et al. (2025); Karkoulian et al. (2024); Tossell et al. (2024); Heil et al. (2025); Chung et al. (2025); Cong et al. (2024); Williams (2024); Dabis et al. (2024)). The release of ChatGPT has made it difficult to detect AI generated content, and instructors worry that existing plagiarism policies are insufficient. Research in this theme comprises three strands: (1) assessing GenAI's performance on assignments and exams, (2) designing assessments that reduce cheating opportunities, and (3) developing AI detection tools and frameworks. For example, some studies fed exam questions or student assignments into ChatGPT and evaluated the plausibility of its answers, showing that GenAI can produce superficially correct responses while occasionally fabricating references or hallucinatory content. Others explored dual anonymous marking exercises where markers attempted to distinguish student work from ChatGPT generated text. Detection studies built multistage systems combining IP monitoring with behavioural analysis or employed AI powered detectors to flag suspicious submissions.

Pedagogical papers advocated redesigning assessments to emphasise higher order cognitive skills. An open access review stressed that while AI can enhance personalised learning and automate feedback, it also threatens academic integrity and requires rethinking exam design and assessment strategies. The same review argued that assessments focusing on complex reasoning and real-world problem-solving are harder for AI systems to mimic and recommended institutions deploy advanced AI detection tools, develop ethical AI policies and provide training to students and staff. Many papers echoed these recommendations, calling for assessments that require originality, reflection, collaboration or multimodal outputs, and for clearer guidance on acceptable AI use.

Policy & governance (45 papers)

Policy oriented studies analysed institutional responses to GenAI and proposed governance frameworks (Polyportis et al. (2024); Hussain et al. (2024); Dai et al. (2024); Wilson (2025); Christidis et al. (2025); Yusuf et al. (2024); Klidas et al. (2025); Dabis et al. (2024); De et al.

(2025); Zhang et al. (2025); Rana (2024); Amr et al. (2024); Wang et al. (2025); Obed et al. (2025); Ardito et al. (2024); Bannister et al. (2025); Riaz et al. (2024); Licht (2025); Zlotnikova et al. (2025); Chen et al. (2025a)). Literature reviews and scoping reviews observed that GenAI has become a priority issue for universities, prompting the formation of AI task forces, the provision of template policy language and AI literacy workshops. However, the absence of universal guidelines is a notable gap; even comprehensive AI ethics laws like Asimov's three laws do not directly apply to GenAI in education, and banning GenAI entirely is both impractical and counterproductive. Surveys of policy documents showed that universities emphasise academic integrity and assessment design but vary widely in their guidance for responsible GenAI use. A recent systematic review concluded that consensus is emerging around four thematic areas—learning objectives, teaching & learning activities, curriculum development and institutional support—and stressed the need for explicit institutional support so educators can responsibly use GenAI.

Ethics & societal impact (48 papers)

Ethical debates ran through 48 papers, many overlapping with other themes (Zhou et al. (2025a); Ahmed et al. (2024); Hussain et al. (2024); Xing (2024); Hsiao et al. (2024); Quezada-Sarmiento et al. (2025); Acosta-Enriquez et al. (2024); Karkoulian et al. (2024); Dabis et al. (2024); De et al. (2025); Shahzad et al. (2024); Barea et al. (2023); Obed et al. (2025); Chedrawi et al. (2025); Licht (2025); Zlotnikova et al. (2025); Chen et al. (2025a); Isiaku et al. (2024); Chen et al. (2025c); Carlos et al. (2025)). A consistent concern was bias in AI training data, with implications for discriminatory or culturally inappropriate outputs (e.g., Chen, 2025). Privacy and intellectual property anxieties were also widespread; students in multiple studies expressed reluctance to upload sensitive work to commercial platforms. Broader critiques addressed labour exploitation in AI supply chains (RSIS review, 2025) and environmental costs of large-scale model training. Several conceptual analyses (n=15) proposed adopting principles of fairness, transparency, and accountability, aligned with international AI ethics debates. A minority of studies examined societal-level risks, such as misinformation or "deepfake" misuse (Beyond a Reasonable Doubt, 2025). Overall, ethics research in higher education remains more normative than empirical, with limited evidence on how institutions operationalise ethical commitments.

Teacher & professional development (27 papers)

Many studies investigated how GenAI affects academic staff. Interviews with lecturers from North America, Europe and Asia revealed varied levels of familiarity with AI but a common recognition that improving AI literacy is essential (Liu et al. (2025); Karkoulian et al. (2024); Chen et al. (2024); Peters (2025); Wang et al. (2025); Obed et al. (2025); Chen et al. (2025b); Villagrán et al. (2024); Gao et al. (2025); Oh (2025); Usher (2025); Li et al. (2025a); Zhou et al. (2025b); Ouyang et al. (2024); Song et al. (2024); Agostini et al. (2024); Ulla et al. (2024)). Faculty participants desired guidance on integrating AI into course objectives and assessments, and they expressed concern about workload and resource implications. Some papers described professional development programmes, such as workshops on prompt engineering or AIassisted grading. However, training opportunities remain uneven across institutions. Research also noted the need for disciplinespecific support and secure access to GenAI tools.

Domain specific studies (53 papers)

The dataset included domain specific investigations in disciplines (Zhou et al. (2025a); Wang et al. (2024); Banerjee et al. (2025); Qiu (2024); Quezada-Sarmiento et al. (2025); Wa et al. (2025); Tossell et al. (2024); Williams (2024); Tsz et al. (2024); Barea et al. (2023); Rivas-Echeverr et al. (2025); Wang et al. (2025); Licht (2025); Chen et al. (2025a); Chen et al. (2025b); Morjaria et al. (2024); Villagrán et al. (2024); Cubillos et al. (2025); Oh (2025); Albuquerque et al. (2024)). Those indicated were: business, engineering, language learning, law, medicine, computer programming, and social sciences. In computing education, experiments evaluated the efficacy of ChatGPT as a programming tutor, while in medical and nursing education GenAI was used to generate case scenarios, simulate clinical reasoning and provide writing assistance for research proposals. Legal and business education papers examined the potential of GenAI to draft contracts, summarise judgments or analyse financial statements. Results were generally promising but underscored the need for domain specific rubrics and human oversight, particularly where factual accuracy and ethical implications are critical.

Student support & wellbeing (20 papers)

Twenty papers examined GenAI in student support and well-being roles (Hsiao et al., 2024; Wa et al., 2025; Liu et al., 2025; Klidas et al., 2025; Heil et al., 2025; Gadekallu et al., 2025; Allen et al., 2024; Carlos et al., 2025; Jin et al., 2024; Zeb et al., 2025; Tesi et al., 2025; Yeung, 2025; Gonzo et al., 2025; Arum et al., 2025; Pwanedo et al., 2025). While the earlier deliverable (Bektik et al., 2024) already discussed tutoring and learning support via chatbots, the present review shows this area broadening to include well-being applications. Recent studies explored AI chatbots providing motivational coaching, study advice, and writing feedback (e.g., Chen et al., 2024). Others addressed technology-related anxiety and overreliance, linking GenAI use to students' sense of trust and psychological strain (e.g., Cong et al., 2024). Survey evidence suggested that students appreciate the convenience and personalised feedback but also worry about dependency and the erosion of peer or teacher interactions. Collectively, these papers emphasise that GenAI may supplement, but should not replace, human advising or counselling services. Importantly, mental health and well-being as explicit foci did not appear in the earlier review (Bektik et al., 2024), marking a new direction in 2025.

Technology evaluation & performance (52 papers)

Papers in this theme focused on evaluating GenAI tools, detecting hallucinations and benchmarking performance (Bai et al. (2024); Hussain et al. (2024); Subhani et al. (2025); Tossell et al. (2024); Cong et al. (2024); Williams (2024); Shahzad et al. (2024); Rivas-Echeverr et al. (2025); Wang et al. (2025); Allen et al. (2024); Chen et al. (2025a); Villagrán et al. (2024); Ji et al. (2025); Chen et al. (2025c); Usher (2025); Li et al. (2025a); Thüs et al. (2024); Moisan et al. (2025); Zhou et al. (2025b); Torenvliet et al. (2024)). Researchers tested ChatGPT's ability to answer exam questions, provide correct programming solutions or generate accurate translations. Many found that GenAI outputs are often plausible but occasionally wrong or fabricated; this "hallucination" problem underscores the importance of critical evaluation. Metaanalyses and bibliometric studies synthesised existing work and identified research trends. Technology development papers proposed new AI enabled detection or tutoring systems and assessed their accuracy using metrics such as accuracy, ROC curves or effect sizes.

In the sections below, findings from the literature for each research question theme is synthesised, drawing on representative studies.

Opportunities and Challenges of LLMs for Teaching and Learning (RQ1)

The literature reveals a dynamic mix of enthusiasm and caution regarding the use of LLMs in higher education classrooms.

Consistent with last year's report, generative AI tools are being used to automate routine teaching tasks and personalise learning. Many papers describe AI-assisted content creation, lesson planning and automated feedback that free educators to focus on interaction. For students, generative AI can provide instant language support, help with coding and data analysis, and foster independent study. Some of the new studies present empirical evidence: for example, experimental trials of AI-powered coding tutors showed improved programming scores for first-time users (Wa et al., 2025; Cubillos et al., 2025), and several studies report gains in writing or language proficiency (Tsz et al., 2024; Ouyang et al., 2024).

Opportunities identified by researchers include enhanced learning support, increased efficiency in content creation, and novel pedagogical approaches. For example, Bai et al. (2024) demonstrated in a case study that integrating ChatGPT into a writing course (via a "Write-Curate-Verify" strategy) enabled rapid generation of high-quality scenario-based learning materials, which in turn improved students' intrinsic motivation and performance. Similarly, several studies reported that LLMs can serve as on-demand tutors or assistants: students have used ChatGPT to get instant explanations of complex concepts, feedback on writing, or practice quiz questions, often with positive effects on their learning confidence (Tossell et al., 2024; Heil et al., 2025). In an experiment in China, Wang et al. (2024) found that students who were taught prompt engineering skills (how to effectively query and interact with LLMs) showed significantly better outcomes in a flipped classroom setting – the LLM provided more relevant and accurate information for those who knew how to prompt it well. This suggests a key opportunity to train students and educators in effective AI interaction, thus maximising benefits. Other noted advantages of LLMs include the ability to generate multiple examples or analogies for a concept (saving instructor time), to serve as a conversational partner for practicing skills (e.g. language learning dialogues), and to help instructors redesign assessments and course materials with AI input (Agostini & Picasso, 2024; Song et al., 2024).

At the same time, **challenges and risks** are a major focus of the literature, and concerns highlighted in the 2024 report persist. By far the most frequently discussed challenge is academic integrity. Ethical and academic integrity issues are a prominent theme (56 of 147 articles), with many papers warning that uncritical use of LLMs may undermine students' critical thinking and creativity. Plagiarism detection remains problematic; current detection tools are unreliable, prompting calls for assessment redesign. Many authors voice concern that tools like ChatGPT make it easier for students to plagiarise or cheat on assignments and exams by generating essays, code, or solutions that bypass learning. A number of surveys have probed student attitudes and self-reported behaviours surrounding this issue. For instance, Campo et al. (2025) found significant correlations between ChatGPT use and plagiarism behaviours, with gender, age, and prior academic performance influencing patterns of misuse. In studies across multiple regions, students showed varying perceptions of whether AI use constitutes cheating (Karkoulian et al., 2024; Acosta-Enriquez et al., 2024). The ease of generating answers has led to what some call an "arms race" in assessment: educators feel pressure to redesign exams and assignments to be "ChatGPT-proof" (e.g. more oral exams, in-class writing, personalised tasks) and to develop better AI-detection tools (Nikolic et al., 2024; Ardito et al., 2024). However, detection of AI-generated text is itself challenging – current detectors yield false positives and can be defeated by simple paraphrasing – so relying on them is not a panacea (Newton, 2025). Thus, the presence of LLMs is pushing a re-examination of what and how we assess in higher education, with calls for more authentic assessments that emphasize process and higher-order thinking (Boadu et al., 2025; Martin et al., 2025).

Beyond plagiarism, **hallucinations and accuracy issues** in LLM outputs pose another challenge for teaching and learning. ChatGPT and similar models can produce text that sounds fluent and authoritative but contains factual errors or completely fabricated information. Multiple studies pointed out the risk of students relying on AI-generated content that is incorrect or biased, potentially impeding learning or spreading misinformation (Zhou et al., 2025a; Qiu, 2024). For example, in a comparative evaluation, Williams (2024) tested ChatGPT on biomedical exam questions: while ChatGPT overall gave detailed answers, it sometimes produced plausible-sounding but incorrect explanations, especially on more advanced topics. Students without enough prior knowledge could be misled by such confident but wrong answers. This highlights the importance of students developing critical evaluation skills when using LLMs – they must learn to verify AI-provided information against reliable sources (Albuquerque et al., 2024). Several empirical studies document mixed or negative effects on learning outcomes: Cong et al. (2024) found associations between ChatGPT use and reduced student life satisfaction and academic performance, while others report that high-ability students sometimes perform worse when relying heavily on AI assistance. This underscores the need to teach AI literacy—how to formulate effective prompts, critically evaluate AI output and understand its limitations. Some educators have experimented with using ChatGPT's mistakes as teachable moments; for instance, having students critique or fact-check an essay written by the AI can build their critical thinking (Tesi et al., 2025). Nonetheless, the consensus is that unverified use of LLM outputs is risky, and both students and teachers need awareness of these pitfalls.

A further set of challenges revolves around the **limitations of current LLM capabilities and the need for AI literacy**. Several papers noted that effectively using LLMs requires new skills (prompt crafting, understanding where the AI excels vs. fails). Students who treat ChatGPT as an all-knowing oracle may get poor results or become overly dependent on it, potentially harming the development of their own expertise (Skalka et al., 2025; Chen et al., 2025c). Educators are concerned about over-reliance on AI: if students use LLMs to do the heavy lifting for writing or coding tasks, they might bypass the deep learning that comes from struggle and practice. Empirical evidence on learning outcomes is beginning to accumulate – for example, Heil et al. (2025) observed that students' perceived impact of GenAI tools on learning varied significantly based on individual AI competence. Some instructors in qualitative studies expressed worry that students might lose skills (like writing or problem-solving) if AI always provides a quick answer (Ji et al., 2025). On the other hand, others argue that the nature of required skills will evolve – using AI effectively might itself become a core competency, and educators should focus on teaching how to collaborate with AI as a tool (Yeung, 2025). This debate represents a broader pedagogical challenge: universities must reconsider learning objectives in each discipline given that generative AI is now part of the knowledge ecosystem.

In summary, the literature portrays a nuanced picture of LLMs in higher education. Opportunities such as personalised tutoring, scaffolding student creativity, and efficient content generation are reported, with evidence of improved motivation and learning in certain contexts (Bai et al., 2024; Wang et al., 2024; Wa et al., 2025). Concurrently, significant challenges are acknowledged: threats to academic integrity (Campo et al., 2025; Obed et al., 2025), inaccuracies and biases in AI outputs (Zhou et al., 2025a; Qiu, 2024), and the need to upskill students and staff to use LLMs critically (Skalka et al., 2025; Tesi et al., 2025). Many studies conclude with a call for balanced integration – leveraging LLMs' benefits while implementing safeguards and new teaching strategies to address the challenges (Zlotnikova et al., 2025; Mariyono et al., 2025). This balance ties closely into institutional responses, discussed further under guidelines and policies.

Impacts on Diversity, Inclusion, and Accessibility (RQ2)

A smaller but important thread in the literature addresses how LLMs might affect diversity, equity, inclusion (DEI), and accessibility in higher education. Relatively few studies (approximately 11% of the literature sample, n≈16/147) explicitly centre on these issues, about the same proportion as last year's review, indicating an ongoing gap in the current research focus. Nonetheless, several key points emerge regarding both the positive potential of LLMs for inclusion and the risks that these tools could exacerbate biases or inequalities.

On the positive side, some authors have suggested that LLMs could serve as equalisers by providing personalised support to students who might otherwise be left behind. For example, generative AI might help level the playing field for students with disabilities or those who are non-native English speakers. Gadekallu et al. (2025) conducted a review on the role of GPT-based tools in supporting students with learning disabilities. They noted evidence that tools like ChatGPT can help students with dyslexia or other difficulties by rephrasing complex texts, generating study summaries, or providing practice questions in a low-stakes, judgment-free manner. Similarly, LLMs can offer 24/7 assistance, potentially benefiting students who lack access to human tutoring or who feel uncomfortable seeking help in person. There is optimism that if properly designed, AI tutors could be more patient and adaptive to individual learner needs than overburdened instructors, thereby aiding inclusion. Chedrawi et al. (2025) explored the role of AI agents in fostering inclusivity for students with special needs in higher education, showing promising applications. For instance, an LLM could simplify the language of readings for a student with limited English proficiency, or provide step-by-step explanations for a first-generation college student who might not have the same academic preparation as peers. Ulla et al. (2024) examined how GenAI can foster inclusive language classrooms from a critical pedagogy perspective, showing potential benefits for multilingual learners. Rivas-Echeverr et al. (2025) developed an LLM-based chatbot for legal assistance that could serve diverse user populations.

However, the literature also raises red flags about biases and inequities associated with LLMs. A crucial concern is that LLMs trained on vast internet datasets encode societal biases – related to gender, race, ethnicity, culture, etc. – which could then be reflected in their outputs and interactions with students. Barea (2023) provided a striking analysis of gender and racial biases in GPT-3's responses. Using a technofeminist critical discourse analysis, they showed that GPT-3's generated text often subtly reinforced stereotypes (for example, associating certain professions or academic abilities with specific genders or ethnic groups). If such biased content were presented to students (say, via an AI tutor's examples or explanations), it could marginalise or alienate underrepresented student groups. Li et al. (2025b) performed a scoping review of societal biases in ChatGPT and warned that without intervention, LLMs might inadvertently perpetuate discrimination in educational content – for instance, by using language or examples that are not culturally inclusive or by less accurately answering questions about minority perspectives due to gaps in training data. These findings underline that diversity and fairness issues are inherent technical challenges with current LLMs.

Furthermore, unequal access to LLMs themselves can be an inclusion issue. Not all students have equal internet bandwidth, or the latest devices required to use AI tools effectively. Some institutions (or countries) have banned or restricted ChatGPT, meaning students in those contexts cannot benefit from it, potentially widening a gap between those with AI access and those without. Arum et al. (2025) found that ChatGPT early adoption in higher education showed significant variation in student usage, instructional support, and educational equity implications. Valdivieso (2025) examined generative AI tools in Salvadoran higher education, highlighting challenges of balancing equity in Global South contexts. Jin et al. (2024) provided a global perspective on institutional adoption policies, revealing regional disparities in readiness and access. This "digital divide" could lead to an imbalance where only some students gain AI-assisted learning advantages. Additionally, even within a classroom, if instructors permit AI use, students with more tech familiarity or better devices might gain more from it, possibly disadvantaging others (Zhang et al., 2025).

Accessibility for students with disabilities is another angle being explored. While AI could provide innovative accommodations (like converting text to simpler language, or acting as a study companion that responds to voice prompts for a student with a visual impairment), there is also worry that LLM tools are not yet fully accessible themselves. For example, current chat interfaces may not be optimised for screen readers or may require visual CAPTCHAs, posing barriers. Gadekallu et al. (2025) and Chedrawi et al. (2025) discussed these aspects, though clear empirical evidence remains sparse. Nonetheless, the potential for LLMs to aid students with special needs is frequently mentioned as a future research direction.

In summary, the impact of LLMs on diversity, inclusion, and accessibility is recognized as a critical but under-studied area. The existing literature points out that LLMs carry latent biases, which could negatively impact educational equity if unaddressed (Barea, 2023; Li et al., 2025b). On the other hand, there is cautious optimism that, if these biases can be mitigated and access broadened, LLMs might become powerful tools for inclusive education – offering tailored support to those who need it most (Gadekallu et al., 2025; Chedrawi et al., 2025; Ulla et al., 2024). Many authors explicitly call for more research on this front, urging studies that examine LLM use among diverse student populations (across genders, cultures, ability levels) and that evaluate outcomes such as sense of belonging, engagement, and performance gaps. In our dataset, only 16 papers (out of 147) substantially addressed DEI concerns, making it a clear gap that future work should fill. We return to this in the Discussion, especially regarding the need for responsible AI design and bias mitigation strategies in educational LLM applications.

Guidelines and Institutional Policies for Responsible LLM Use (RQ3)

In the first review (D2.1, Bektik et al., 2024), only a small subset of studies (n=20) explicitly addressed policy, governance, or institutional guidelines for generative AI in higher education. In the present corpus, this has increased to 44 studies, signalling a marked shift towards examining both institutional and governmental responses. This change reveals a progression from early conceptual reflections towards more applied analyses of concrete policy frameworks, institutional guidelines, and AI literacy initiatives.

University guidelines

Recent studies show that many universities that initially banned generative AI tools have moved towards policies of "responsible use." For example, Dai et al. (2024) provide a scoping review of university policies across Asia, highlighting wide variation in institutional readiness and the balance between innovation and integrity. In the UK, Wilson (2025) tracks the evolution of institutional guidelines, showing a shift from prohibition to cautious integration. Similarly, Christidis et al. (2025) analyse Swedish higher education, illustrating how formal policies are being translated into practice. Across these cases, the emphasis is on academic integrity, transparency, and fairness, but systematic evaluation of policy effectiveness is still lacking.

National/Regional regulatory frameworks

At the national and regional level, regulatory frameworks are beginning to shape higher education practices. Ahmed et al. (2024) map the global landscape of regulatory and governance issues, noting that the EU's AI Act stands out for classifying educational AI applications as "high risk," imposing significant compliance obligations on universities. By contrast, the UK's "pro-innovation" approach and US federal guidance emphasise flexibility rather than prescriptive regulation. Comparative regional perspectives also emerge: Quezada-Sarmiento et al. (2025) discuss ethical governance challenges in Latin America, while Acosta-Enriquez et al. (2024) document how Ecuadorian universities are navigating ethical and regulatory debates around AI. Mahrishi et al. (2024) examine global initiatives towards regulatory frameworks for AI in higher education. These findings suggest that institutions in the EU are under greater compliance pressure, whereas those in the UK, US, and Global South retain more discretion, albeit with less clarity.

AI literacy and capacity-building

A notable development since D2.1 (Bektik et al., 2024) is the growing emphasis on AI literacy and professional development as essential complements to policy frameworks. Studies documented staff workshops, student orientations, and short courses that aimed to (a) clarify acceptable use, (b) build prompt-crafting and critical-reading skills, and (c) reduce privacy and accuracy risks when using third-party tools. Most initiatives were small-scale pilots; formal evaluations of learning impact and policy compliance outcomes were still limited. Skalka et al. (2025) examined AI literacy structure and factors influencing student attitudes across Central European universities. Qu et al. (2024) demonstrate disciplinary differences in AI literacy, showing that students in STEM fields are more confident adopters, while those in the humanities report greater uncertainty and ethical hesitation. Tesi et al. (2025) explored how AI literacy and self-regulated learning relate to student writing performance. Collectively, these studies indicate that AI literacy is increasingly viewed as a prerequisite for responsible use, though most initiatives remain small-scale and unevenly distributed.

Patterns and gaps

Overall, the literature shows rapid but uneven growth in policy and governance responses to generative AI. Compared with the previous review (D2.1, Bektik et al., 2024), which found only 20 relevant studies, this review includes 44, signalling a major expansion of interest in institutional and regulatory frameworks.

Several consistent patterns emerge:

Variation in institutional responses. Universities range from outright bans to permissive "responsible use" policies. Studies from Asia (Dai et al., 2024) and Sweden (Christidis et al., 2025) highlight wide gaps in readiness and the absence of consistent, formalised guidelines. In the UK, Wilson (2025) documents a shift from prohibition to cautious integration. Transparency requirements, such as policies mandating students declare AI use in assignments, are emerging in some faculties (Nikolic et al., 2024; Peters, 2025).

Alignment with international frameworks. National and regional approaches differ significantly. The EU's AI Act imposes binding compliance requirements for education as a "high-risk" sector (Ahmed et al., 2024), whereas the UK emphasises a pro-innovation strategy and the US favours flexible guidance. Broader governance issues in Latin America

(Quezada-Sarmiento et al., 2025) and Ecuador (Acosta-Enriquez et al., 2024) show that universities elsewhere often lack clear direction. At the international level, reviews recommend alignment with UNESCO and OECD guidelines on transparency, accountability, and fairness (Mahrishi et al., 2024; Zlotnikova et al., 2025).

AI literacy as a policy complement. Across contexts, there is recognition that policies alone are insufficient without parallel training and literacy initiatives. European studies (Skalka et al., 2025), disciplinary divides in student confidence (Qu et al., 2024), and explorations of literacy-performance relationships (Tesi et al., 2025) all point to the importance of embedding AI literacy in institutional strategies.

Despite this progress, **important gaps remain**. Most studies are descriptive or policy-analytic, with little empirical evaluation of how policies affect student behaviour, teaching practice, or academic integrity in practice. Newton (2025) provides a pragmatic risk assessment approach, but direct assessment of policy outcomes remains limited. Similarly, although there are calls for policies to be co-created with staff and students (Barus et al., 2025), systematic evaluations of participatory approaches are still lacking.

Taken together, the evidence suggests that higher education is moving towards a model of cautious adoption coupled with ethical oversight, where responsible-use policies, regulatory alignment, and literacy initiatives evolve in tandem. However, the effectiveness of these measures remains underexplored, making this a key area for future research.

Common elements of emerging institutional policies on LLM use

We can expect the next couple of years to bring more standardised policies and possibly sector-wide principles. The conversation around policy is very much active. Typical policy elements discussed in the current corpus, reflected across multiple studies (n=44 in this review), are summarised below:

Table 9 Summary of Policy Elements addressed in the studies identified

Policy Element	Studies Addressing	Description
Academic Integrity Clauses	35	Clear statements that unauthorised use of AI for graded work is considered misconduct, analogous to plagiarism, unless explicitly permitted by the instructor
Disclosure Requirements	25	Policies requiring students to disclose any AI use in submissions (via footnotes, "AI usage" statements, or honour code pledges)
Assessment Re-Design Guidance	30	Recommendations for instructors to adopt assessment formats less vulnerable to AI misuse—such as oral exams, in-class work, personalised projects

Detection/Verification Guidance	18	Institutions exploring AI-detection software, often with caveats about reliability
AI Literacy/Training Provisions	22	Policies linking rules to training—workshops for faculty, tutorials for students, or dedicated institutional resources
Data Privacy and Security Cautions	15	Mainly in EU contexts, policies caution against uploading sensitive data to commercial AI platforms

These elements illustrate how institutions are striving to balance control and support. As noted by multiple studies (Dai et al., 2024; Wilson, 2025; Christidis et al., 2025), the key challenge lies in calibrating policies: too lax risks undermining academic standards, too strict risks stifling beneficial innovation. The emerging consensus is that policy development will remain iterative, adapting as both technologies and institutional practices evolve.

Comparison of last year's report (Jan 2022 – Oct 2024) with the current review (Oct 2024 – Jul 2025)

Volume and maturation of the evidence base

The first review (D2.1, Bektik et al., 2024) identified 112 sources (including grey literature) covering the early emergence of generative AI tools such as ChatGPT, Bard/Gemini and other large language models. In this updated review (D2.2), database searches conducted between October 2024 and July 2025 retrieved over 500 records. After removing duplicates and applying inclusion and exclusion criteria, 168 articles met the eligibility criteria. Of these, 21 full texts could not be accessed, leaving 147 unique peer-reviewed papers in the final coded dataset Grey literature was excluded in this round to ensure comparability and quality of evidence.

This reflects a substantial increase in volume within just nine months, underscoring how rapidly the scholarly literature on generative AI in higher education is expanding.

Despite this rapid growth, the overall evidence base is still in an early stage of maturation. Many studies remain exploratory, dominated by small-sample surveys or descriptive case studies. However, compared with D2.1 (Bektik et al., 2024), this corpus includes more empirical evaluations, such as field experiments, quasi-experimental designs, and a small but growing number of meta-analyses. For example, one study evaluated an LLM-powered "CodeTutor" across a semester-long programming course, while another synthesised results on learning outcomes using Bloom's taxonomy as a framework. These developments point to a gradual but significant increase in methodological diversity, though robust large-scale evaluations remain rare.

Tools and platforms

Last year's report noted the dominance of ChatGPT and generic "LLM" references. That pattern persists: ChatGPT/GPT was explicitly mentioned in 65 of the 147 studies while only a handful referred to Gemini/Bard, Claude or Llama. Emerging opensource models (Mistral, LLaMA 2/3) and multimodal tools remain underrepresented in the peer reviewed literature. The focus continues to be on text generation rather than multimodal AI, despite the rapid evolution of multimodal models in industry.

Thematic shifts

- **Assessment and academic integrity.** The 2024 review highlighted general concerns about plagiarism and critical thinking. In the 2025 corpus, assessment has become one of the most frequently addressed topics (44 of 147), reflecting a shift from speculative worries to concrete investigations. Studies test AI assisted grading rubrics, examine how ChatGPT affects exam scores, and propose frameworks for AIdriven assessments. The concern about academic integrity remains high, but there is greater emphasis on designing assessments that leverage AI responsibly rather than banning it outright.

- **Ethics and policy.** Ethical issues and policy responses were major themes in both years. New work builds on last year's discussion of privacy, bias and fairness. Some recent papers analyse universities' updated guidelines, noting a move from blanket prohibitions toward "responsible use" policies and AI literacy training for staff and students. There is continued interest in regulatory developments such as the EU AI Act and national guidelines, but little empirical evaluation of their educational impact.

- **Learning outcomes and evidence.** The earlier report concluded that few studies had measured learning outcomes directly. In the latest literature, more papers report quantitative outcomes: some show improved performance in coding and writing tasks when students use AI assistants, while others (e.g., Wecks et al.) find reduced exam scores for AI users. However, these studies remain small-scale and often lack control groups; systematic reviews continue to call for robust, longitudinal research.

- **Diversity, inclusion and accessibility.** The 2024 report viewed GenAI as a potential equaliser but warned about bias and access inequalities. In this review (D2.2), only 16 papers explicitly addressed diversity, inclusion, or accessibility, compared with just a handful noted in D2.1 (Bektik et al., 2024). Most remain conceptual or design-oriented, with little empirical testing. As in last year's report, there is scant evidence that the diversity gap is narrowing: few studies focus on learners with disabilities or on non-Western cultural contexts. Geographic coverage clusters around the UK (n=7), China (n=5), USA/Canada, Indonesia, and Ecuador, with regions such as Africa and the Middle East scarcely represented.

Geographical and disciplinary spread

STEM fields dominate the findings which is similar to last year. New studies continue to explore AI's use in programming, engineering and medicine, while humanities and social sciences remain underrepresented. Some papers broaden the disciplinary scope (e.g., law, economics, language learning), but cross-disciplinary and cross-cultural comparisons are still lacking. The majority of studies do not specify a geographic context; when mentioned, they cluster around a few countries, mirroring the 2024 report.

Methodological and knowledge gaps

Many of the gaps identified in last year's report persist:

- **Small samples and descriptive designs.** Few studies move beyond convenience samples or single-course case studies; large-scale, multi-institutional trials are still rare.
- **Critical thinking effects.** Concerns about reduced critical thinking and overreliance on AI remain, and the new evidence is mixed: some studies report enhanced reflective skills, while others document declines in exam performance.
- **Bias and non-English contexts.** The dominance of English language corpora and Western cultural norms continues, with little progress on culturally sensitive or multilingual LLMs.
- **Inclusion and accessibility.** Research into how generative AI serves students with disabilities or different socio-economic backgrounds remains sparse.

Overall assessment

Compared to the 2024 report, the literature from Oct 2024 to Jul 2025 shows **rapid growth** and **slightly greater methodological diversity**, with more empirical evaluations of learning outcomes and a stronger focus on assessment design. Ethical and policy discussions have shifted from raising alarms to proposing frameworks and training programmes. Nevertheless, most of the substantive gaps—rigorous experimental evidence, largescale studies, inclusive and cross-cultural research—remain unresolved. The field is still in its early stages; while optimism about GenAI's potential persists, there is growing recognition that careful design, ethical oversight and AI literacy are essential to realise its benefits responsibly.

Response to the Research Questions

RQ1. Opportunities and challenges for teaching and learning

LLMs are becoming increasingly embedded in teaching practice, particularly in writing-intensive and STEM contexts. Studies report improved efficiency and formative learning gains when students use LLMs for feedback, tutoring, and collaborative learning (Bai et al., 2024; Wa et al., 2025; Ouyang et al., 2024). In writing contexts, research demonstrates potential benefits for essay development and revision processes (Tossell et al., 2024; Peters, 2025; Song et al., 2024), while STEM disciplines show promise in programming education and problem-solving support (Banerjee et al., 2025; Cubillos et al., 2025; Wu et al., 2024).

At the same time, challenges are significant. Academic integrity concerns are widespread, with studies documenting increased potential for AI-assisted misconduct (Campo et al., 2025; Nikolic et al., 2024; Obed et al., 2025). Over-reliance on LLMs may reduce critical thinking and independent problem-solving skills (Heil et al., 2025; Ji et al., 2025; Cong et al., 2024), while limited AI literacy restricts effective and ethical use among both students and educators (Qu et al., 2024; Skalka et al., 2025; Tesi et al., 2025; Chen et al., 2025c). Additional concerns include LLM hallucinations affecting learning accuracy (Qiu, 2024) and the need for substantial pedagogical redesign to integrate these tools effectively (Boadu et al., 2025; Martin et al., 2025).

The evidence suggests that LLMs offer substantial opportunities for educational innovation, but realising benefits depends on redesigning assessment practices, embedding AI literacy in curricula, and developing appropriate pedagogical frameworks for human-AI collaboration.

RQ2. Impacts on diversity, inclusion, and accessibility

Evidence on equity impacts remains sparse, with approximately 11% of studies (n≈16/147) explicitly addressing diversity, inclusion, or accessibility dimensions. Potential benefits identified include support for students with disabilities through AI agents and adaptive technologies (Gadekallu et al., 2025; Chedrawi et al., 2025), assistance for multilingual learners in language development (Ulla et al., 2024; Dang et al., 2024), and tools for addressing diverse learning needs (Rivas-Echeverr et al., 2025).

However, significant risks have been documented. Studies highlight concerns about reinforcing existing biases through AI-generated content (Barea, 2023; Li et al., 2025b), privileging English-language users and Western educational contexts (Albuquerque et al., 2024; Valdivieso, 2025), and potentially exacerbating digital divides between well-resourced and under-resourced institutions (Ahmed et al., 2024; Zeb et al., 2025). Cross-regional analyses reveal uneven readiness for equitable AI integration, with some contexts showing stronger equity safeguards than others (Jin et al., 2024; Quezada-Sarmiento et al., 2025; Arum et al., 2025).

More empirical work is required, particularly longitudinal studies, to determine whether LLMs ultimately narrow or widen educational inequalities across different student populations and institutional contexts.

RQ3. Guidelines and institutional policies

Institutions worldwide are moving from outright bans toward responsible-use guidelines that emphasise academic integrity, transparency, and AI literacy development (Wilson, 2025; Christidis et al., 2025; Dai et al., 2024). This policy evolution is documented across diverse contexts including the UK (Wilson, 2025), Sweden (Christidis et al., 2025), Asia-Pacific regions (Dai et al., 2024), and Latin America (Quezada-Sarmiento et al., 2025).

Common policy elements emerging across institutions include: disclosure requirements for AI use in academic work (Nikolic et al., 2024; Peters, 2025); training programmes for students and staff on responsible AI use (Zlotnikova et al., 2025; Dabis & Csáki, 2024); academic integrity clauses and detection mechanisms (Campo et al., 2025; Obed et al., 2025); and frameworks for ethical AI integration in curricula (Yusuf et al., 2024; Mariyono et al., 2025).

Several studies provide detailed analyses of institutional responses. Dabis & Csáki (2024) document the first wave of policy responses from higher education institutions globally, while Rana (2024) reviews policies specifically from research-intensive universities. Jin et al. (2024) offer a global perspective on adoption policies, and Barus et al. (2025) examine how governance frameworks are being shaped by student perceptions.

However, critical gaps remain. Few studies evaluate the effectiveness of these frameworks in practice (Newton, 2025; Gonzo et al., 2025). Governance approaches remain at an early stage, with implementation often reactive rather than proactive and fragmented across institutional units (Ahmed et al., 2024; Mahrishi et al., 2024). There is also limited evidence on how policies translate into changed classroom practices or student behaviours (Bannister et al., 2025; Ouyang et al., 2024).

Discussion

This review extends the earlier D2.1 (Bektik et al., 2024) report by offering a more detailed and thematically organised synthesis of how large language models (LLMs) are influencing higher education practice and policy. Whereas D2.1 provided an early-stage snapshot of potential benefits and emergent risks, this review captures a clearer picture of implementation in real educational settings and a rapidly expanding scholarly response.

One of the most striking shifts is the move from alarm to adaptation. D2.1 documented significant anxiety about threats to academic integrity; in contrast, this review finds increasing engagement with redesigning assessment and pedagogy. Assessment was one of the most common themes in the present corpus (44 of 147 studies), reflecting a transition from speculative concerns to concrete investigations. Institutions are beginning to acknowledge AI as part of the learning ecosystem, albeit with uneven strategies and support. This reflects a broader cultural turn toward “constructive realism” rather than reactionary control.

Equity remains an area of continuity and concern. D2.1 warned that marginalised learners may be disproportionately affected by AI adoption, and this review confirms that such risks persist. Only 10% of studies in the current corpus (n=22 of 168 screened titles, 12 in the final set of 147) addressed inclusion and equity in depth, highlighting a major evidence gap. At the same time, there is growing interest in AI’s inclusive potential — particularly for multilingual learners and students with disabilities — though robust empirical evidence is still lacking. The literature continues to echo calls for more inclusive design and evaluation.

Methodologically, the field shows signs of maturation. While D2.1 observed an over-reliance on anecdote and opinion pieces (around 70% conceptual vs. 30% empirical), this review identifies a stronger empirical base: surveys (n=67), qualitative fieldwork (n=64), technology evaluations (n=67), and a modest rise in experiments (n=23). Nevertheless, most remain small-scale, cross-sectional, or exploratory, with rigorous long-term evaluations still rare. The imbalance across disciplines also persists, with STEM fields dominant (notably computer science, engineering, and medicine) and humanities and social sciences underrepresented.

Finally, D2.1 predicted the need for institutional policy frameworks; this review confirms their emergence. The number of policy- or governance-focused studies grew from 20 in D2.1 to 63 in this review. Yet many policies remain reactive, vague, or inconsistently enforced. Some universities now articulate “responsible use” clauses and AI literacy initiatives, but systematic evaluations of their impact are still absent. The urgency has shifted from speculative discussion to the practical challenge of developing and implementing policies that are co-created with students and staff, grounded in evidence, and attentive to both risks and opportunities.

Taken together, these shifts suggest that higher education has entered a second phase of engagement with LLMs: moving beyond initial alarm and speculation towards constructive adaptation. To consolidate this progress, the next phase will require inclusive, cross-disciplinary, and evidence-based frameworks that ensure AI integration enhances learning, supports equity, and upholds academic integrity.

Conclusion

This second review (D2.2) showed that research on generative AI in higher education expanded rapidly between October 2024 and July 2025. We identified 168 peer-reviewed titles and analysed 147 full texts. Compared with the first review (D2.1; Bektik, 2024, Jan 2022–Oct 2024), the evidence base matured: empirical designs became more common, policy analyses proliferated, and the literature coalesced around a set of recurring concerns—adoption and perception, assessment and academic integrity, teaching and learning, and ethics—while widening to include institutional governance and student support.

Across the corpus, coverage was broad and methods were more varied than in D2.1. Adoption/perception, assessment/integrity, and teaching/learning remained the most frequently studied themes, with a marked rise in policy and governance work (63 studies, up from 20 in D2.1). Methodologically, surveys (n=67), qualitative fieldwork (n=64), and technology development/evaluation (n=67) dominated; experimental or quasi-experimental designs (n=23) and a small number of meta-analyses (n=2) also appeared, signalling a shift away from the largely conceptual/descriptive balance reported in D2.1. Early, discipline-focused evaluations of learning and assessment accumulated, and institutions began documenting policy shifts from prohibition towards “responsible use,” alongside emerging AI-literacy initiatives.

At the same time, important limitations persisted. Many empirical studies were still small, cross-sectional, or single-site, constraining causal inference and generalisability. Diversity, inclusion, and accessibility remained under-represented (10% of studies), and the geographic spread continued to skew towards a handful of countries, with large regions of the Global South scarcely visible. Policy papers were largely descriptive; few evaluated the real-world impact of disclosure requirements, responsible-use clauses, or training programmes. Student support and wellbeing constituted a newer thread (n=53), but robust outcome evidence—especially around psychological effects and over-reliance—was limited. Technology evaluations multiplied (n=75), yet most benchmarked capabilities or detection accuracy rather than documenting longitudinal pedagogical impact.

These patterns suggested a clear agenda for the next phase of research. Methodologically, the field would benefit from larger, multi-institutional, and longitudinal designs, as well as stronger quasi-experimental or experimental approaches that can support causal claims. Substantively, priorities include systematic measurement of learning outcomes (performance, critical thinking, transfer), rigorous evaluations of institutional policies and AI-literacy programmes, and a decisive broadening of equity-focused work to include disability, language diversity, and Global South contexts. Evidence on student support and wellbeing needs scaling and standardisation, and cross-disciplinary/cross-cultural comparisons should extend beyond STEM to humanities and social sciences. Finally, scholarship needs to catch up with the rapid deployment of multimodal and agentic systems, which remain under-examined relative to text-only LLMs.

Overall, higher education research on generative AI moved from initial commentary to applied experimentation and institutional response. The corpus analysed here was larger and more diverse than in D2.1 (Bektik et al, 2024), but still uneven in quality and scope. Realising the promise of GenAI responsibly will require methodological rigour, inclusive sampling, and systematic evaluation of real-world practice. With those shifts, the sector will be better placed to determine which uses of AI genuinely enhance learning and assessment, how to uphold academic integrity and equity, and how to support educators and students through a durable, evidence-led transition.

Gaps and Areas for Further Research

Drawing on both the thematic review and the methodological observations, several important gaps remain in the evidence base. These highlight where research is thin or absent, and where future work could make the greatest contribution.

A first and most prominent gap is the absence of large-scale, controlled studies. The current corpus is dominated by small-sample surveys, descriptive case studies, or single-course pilots. No study involved thousands of students across multiple institutions in a randomised design, yet such scale is likely necessary to detect effects with confidence and to provide evidence that policy makers will trust. Without large, multi-site, experimental or quasi-experimental research, questions such as whether AI tutoring systems improve retention or long-term performance remain unanswered. Scaling up research through cross-university collaboration and targeted funding should therefore be a priority.

Second, most studies are cross-sectional snapshots, leaving a major gap in longitudinal evidence. We do not yet know how student or staff use of AI evolves over multiple semesters, how initial novelty effects wear off, or whether reliance on AI increases or diminishes over time. The literature is also weighted towards STEM, computer science, and business education, while disciplines such as the arts, humanities, and some social sciences are scarcely represented. Given that the affordances and challenges of AI may vary significantly across disciplines, future research should not only expand coverage to under-studied areas but also compare disciplinary contexts directly.

A third gap lies in equity, diversity, and accessibility. Only around 11% of studies explicitly addressed these questions, leaving critical uncertainties about whether generative AI tools exacerbate or mitigate inequality. Very little is known about how AI performs for students with disabilities, whether outputs are reliably accessible (e.g., screen reader compatibility), or whether cognitive demands disadvantage neurodivergent learners. Similarly, cultural and linguistic bias in AI systems remains underexplored in higher education settings, raising concerns for non-native speakers and students from non-Western contexts. Access gaps—driven by differences in infrastructure, connectivity, or device availability—are also seldom considered. To ensure inclusive adoption, future research needs to integrate equity analyses systematically and design targeted studies around accessibility and bias.

Policy and governance also remain underdeveloped. While more institutions are issuing responsible-use guidelines, most studies in this review analysed documents rather than evaluating outcomes. We still lack evidence on whether explicit AI policies reduce misconduct, whether disclosure requirements increase transparency, or whether faculty training changes practice in classrooms. The same is true of detection tools and honour code clauses: many are being introduced, but their actual effectiveness remains unknown. Without evaluation, much of current governance rests on assumptions. Future research should therefore focus on testing policy approaches, measuring their outcomes, and identifying unintended effects.

Finally, comparative and cross-cultural studies are rare. Most research implicitly assumes Western higher education models, but cultural attitudes to collaboration, plagiarism, or innovation may alter how AI is perceived and used. Early cross-country analyses—for example, comparing policy responses across Asian universities or contrasting perspectives from Latin America with those from Europe—suggest that context matters significantly. Yet these examples are exceptions. More comparative work, involving researchers from underrepresented regions as lead investigators, is essential for building a globally relevant evidence base.

Taken together, these gaps suggest that the literature is still in an exploratory phase. Addressing them will require larger and more rigorous designs, broader disciplinary and geographical coverage, systematic attention to equity and accessibility, and evaluation of policies in practice. Filling these gaps is crucial if higher education is to move from descriptive insights towards robust, evidence-based recommendations for the responsible and effective integration of generative AI.

References

Acosta-Enriquez, B. G., Agustín, M., Ballesteros, A., Arbulu, C. G., Vargas, P., Naara, M., ... Ulloa, O. (2024). Knowledge, attitudes, and perceived Ethics regarding the use of ChatGPT among generation Z university students.

Agostini, D. & Picasso, F. (2024). Large language models for sustainable assessment and feedback in higher education Towards a Pedagogical and Technological Framework.

Ahmed, Z., Shan, S. S., Humayra, M., Rime, K., Kishor, M. D., Morol, ... Fahad, N. (2024). The Generative AI Landscape in Education: Mapping the Terrain of Opportunities, Challenges, and Student Perception.

Albuquerque, J., Rienties, B., Holmes, W., & Hlosta, M. (2024). From hype to evidence: exploring large language models for inter-group bias classification in higher education.

Alghazo, R., Fatima, G., Malik, M., Abdelhamid, S. E., Jahanzaib, M., Nayab, D. E., ... Raza, A. (2025). Exploring ChatGPT's Role in Higher Education: Perspectives from Pakistani University Students on Academic Integrity and Ethical Challenges.

Allen, M., Naeem, U., & Gill, S. S. (2024). Q-Module-Bot: A Generative AI-Based Question and Answer Bot for Module Teaching Support.

Amr, A., Elsayed, Caeiro-Rodríguez, M., Mikic-Fon, F. A., Lugilde-López, A., & Llamas-Nistal, M. (2024). Measuring and Promoting Self-Regulated Learning Using Spaced Questionnaires.

Ardito, C. G., Cesare, P. H. Y., & Ardito, G. (2024). Generative AI detection in higher education assessments.

Arum, R., Calderon, M., Xunfei, L., Li, J., & Lopes (2025). ChatGPT Early Adoption in Higher Education: Variation in Student Usage, Instructional Support, and Educational Equity.

Bai, S., Gonda, D. E., & Foon, K. (2024). Write-Curate-Verify: A Case Study of Leveraging Generative AI for Scenario Writing in Scenario-Based Learning.

Banerjee, P., Srivastava, A. K., Adjeroh, D. A., Reddy, R., & Karimian (2025). Understanding ChatGPT: Impact Analysis and Path Forward for Teaching Computer Science and Engineering.

Bannister, P., Peñalver, E. A., & Urbieta, A. S. (2025). Transnational higher education cultures and generative AI: a nominal group study for policy development in English medium instruction.

Barea, M. A. P., Boeren, D., & Goncalves, J. F. F. (2023). At the intersection of humanity and technology: a technofeminist intersectional critical discourse analysis of gender and race biases in the natural language processing model GPT-3.

Barus, O. P., Hidayanto, A. N., Handri, E. Y., Sensuse, D. I., & Yaiprasert, C. (2025). Shaping generative AI governance in higher education: Insights from student perception.

Bektik, Duygu; Edwards, Chris; Whitelock, Denise and Antonaci, Alessandra (2024). *Use of LLM tools within higher education: Report 1*. Quality Enhancement and Innovation, The Institute of Educational Technology, The Open University, Milton Keynes, UK.

Boadu, G., Tenakwah, E. J., Brady, M., Radavoi, C., Senior, E., Parzakonis, M., ... Kansiime, P. (2025). Assessment in the age of artificial intelligence: Interdisciplinary analysis of ChatGPT response to higher education assessment tasks.

Campo, L., Delgado, N., Etxab-Urbieta, J. M., Kanso, H., Gueye, M. L., & Etxabe-Urbieta, J. M. (2025). Relationship Between the Use of ChatGPT and Plagiarism in Higher Education: The Influence of Gender, Age and Previous Academic Results.

Carlos, J., Batista, L., Apostolou, D., Ilieva, G., Yankova, T., Ruseva, M., & Kabaivanov, S. (2025). A Framework for Generative AI-Driven Assessment in Higher Education.

Chedrawi, C., Kazoun, N., & Kokkinaki, A. (2025). The role of AI agents in fostering inclusivity for HEIs' students with special needs against backdrops of the accreditation trend.

Chen, Z., Wang, J., Xia, M., Shigyo, K., Liu, D., Zhang, R., & Qu, H. (2024). StuGPTViz: A Visual Analytics Approach to Understand Student-ChatGPT Interactions.

Chen, X., Zhang, J., Zhou, T., & Zhang, F. (2025a). LLM-CDM: A Large Language Model Enhanced Cognitive Diagnosis for Intelligent Education.

Chen, K., Tallant, A. C., & Selig, I. (2025c). Exploring generative AI literacy in higher education: student adoption, interaction, evaluation and ethical perceptions.

Ching, L., Tiong, O. W., Lee, Y., Lim, K. L., & Lee, J. H. (2024). Advancing Online Assessment Integrity: Integrated Misconduct Detection via Internet Protocol Analysis and Behavioral Classification.

Christidis, N., Christidis, M., Erhardt, C., Kullenberg, H., Grigoriadis, A., & Kumar, A. (2025). From policy to practice: the regulation and implementation of generative AI in Swedish higher education institutes.

Chung, J. & Kwon, H. (2025). Privacy fatigue and its effects on ChatGPT acceptance among undergraduate students: is privacy dead?.

Cong, D., Duong, T., Nghia, V., Viet, T., Ngo, N., Do, N. D., ... Tran, N. M. (2024). Reduced Student Life Satisfaction and Academic Performance: Unraveling the Dark Side of ChatGPT in the Higher Education Context.

Cubillos, C., Mellado, R., Cabrera-Paniagua, D., & Urra, E. (2025). Generative Artificial Intelligence in Computer Programming: Does It Enhance Learning, Motivation, and the Learning Environment?.

Dabis, A. & Csáki, C. (2024). AI and ethics: Investigating the first policy responses of higher education institutions to the challenge of generative AI.

Dai, Y., Lai, S., Lim, C. P., & Liu, A. (2024). University policies on generative AI in Asia: promising practices, gaps, and future directions.

Dang, A. & Wang, H. (2024). Ethical use of generative AI for writing practices: Addressing linguistically diverse students in U.S. Universities' AI statements.

De, E. & Evangelista, L. (2025). Ensuring academic integrity in the age of ChatGPT: Rethinking exam design, assessment strategies, and ethical AI policies in higher education.

Duah, J. E. & Mcgivern, P. (2025). How generative artificial intelligence has blurred notions of authorial identity and academic norms in higher education, necessitating clear university usage policies.

Florentin, G., Mvondo, N., Niu, B., & Eivazinezhad, S. (2025). Exploring the ethical use of LLM chatbots in higher education.

Gadekallu, T. R., Yenduri, G., Kaluri, R., Rajput, D. S., Lakshmanna, K., Fang, K., ... Chen, J. (2025). The role of GPT in promoting inclusive higher education for people with various learning disabilities: a review.

Gao, B., Yan, J., & Zhong, R. (2025). How Digital Teacher Appearance Anthropomorphism Impacts Digital Learning Satisfaction and Intention to Use: Interaction With Knowledge Type.

Gonzo, F., Orcid, Geller, L., Mukesi, M., & Gonzo, M. (2025). AI Meets Academia: the Impact of ChatGPT on Transforming Assessment Design in UK Universities.

Heil, J., Ifenthaler, D., Cooper, M., Mascia, M. L., Conti, R., & Penna, M. P. (2025). Students' perceived impact of GenAI tools on learning and assessment in higher education: the role of individual AI competence.

Hsiao, C. & Tang, K. (2024). Beyond acceptance: an empirical investigation of technological, ethical, social, and individual determinants of GenAI-supported learning in higher education.

Hussain, F., Anwar, M. A., & Hussain, F. (2024). Towards informed policy decisions: assessing student perceptions and intentions to use ChatGPT for academic performance in higher education.

Isiaku, L., Kwala, A. F., Sambo, K. U., Chidozie, F., Ukaegbu, Isaku, H. H., & Crawford, J. (2024). Academic Evolution in the Age of ChatGPT: An In-depth Qualitative Exploration of its Influence on Research, Learning, and Ethics in Higher Education.

Ji, Y., Zhan, Z., Li, T., Zou, X., & Lyu, S. (2025). Human-Machine Cocreation: The Effects of ChatGPT on Students' Learning Performance, AI Awareness, Critical Thinking, and Cognitive Load in a STEM Course Toward Entrepreneurship.

Jin, Y., Yan, L., Echeverria, V., Gašević, D., & Martinez-Maldonado, R. (2024). Generative AI in higher education: A global perspective of institutional adoption policies and guidelines.

Karkoulian, S., Sayegh, N., & Sayegh, N. (2024). ChatGPT Unveiled: Understanding Perceptions of Academic Integrity in Higher Education -A Qualitative Approach.

Kirwan, A. (2023). ChatGPT and university teaching, learning and assessment: some initial reflections on teaching academic integrity in the age of Large Language Models.

Klidas & Karakitsou, C. (2025). Between a Bot and a Hard Place: Student Engagement.

Klyshbekova, M. & Abbott, P. (2024). ChatGPT and Assessment in Higher Education: A Magic Wand or a Disruptor?.

Ku, M., Li, X., Obiomon, P., Qian, L., & Dong, X. (2025). Improving Student Learning Outcome Tracing at HBCUs Using Tabular Generative AI and Deep Knowledge Tracing.

Li, M., Cheng, F., & Yamamoto, B. A. (2023). Ethical Implications of ChatGPT in Higher Education: A Scoping Review.

Li, K., Yang, Q., & Yang, X. (2024). Can Autograding of Student-Generated Questions Quality by ChatGPT Match Human Experts?.

Li, X., Zappato, M., Li, T., Zhang, W., Tao, S., Wei, X., ... Zhou, X. (2025a). GAI Versus Teacher Scoring: Which is Better for Assessing Student Performance?.

Li, M., Enkhtur, A., Yamamoto, B. A., Cheng, F., & Chen, L. (2025b). Potential Societal Biases of ChatGPT in Higher Education: A Scoping Review.

Licht, K. D. F. (2025). Rethinking the Ethics of GenAI in Higher Education: A Critique of Moral Arguments and Policy Implications.

Liu, Y., Awang, H., & Mansor, S. (2025). Exploring the Potential Barrier Factors of AI Chatbot Usage Among Teacher Trainees: From the Perspective of Innovation Resistance Theory.

López-Pernas, S., Misiejuk, K., Kaliisa, R., & Saqr, M. (2025). Capturing the Process of Students' AI Interactions When Creating and Learning Complex Network Structures.

Ma, W. W. K. & Sánchez-Vera, F. (2024). Subject-Specialized Chatbot in Higher Education as a Tutor for Autonomous Exam Preparation: Analysis of the Impact on Academic Performance and Students' Perception of Its Usefulness.

Mahrishi, M., Gramothan, S., Abbas, A., & Siddiqui, M. K. (2024). Global Initiatives Towards Regulatory Frameworks for Artificial Intelligence (AI) in Higher Education.

Mariyono, D., Nur, A., & Hidayatullah, A. (2025). Navigating the Moral Maze: Ethical Challenges and Opportunities of Generative Chatbots in Global Higher Education.

Martin, A. F., Tubaltseva, S., Harrison, A., & Rubin, G. J. (2025). Participatory Co-Design and Evaluation of a Novel Approach to Generative AI-Integrated Coursework Assessment in Higher Education.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Medicine, 6*(7), e1000097. https://doi.org/10.1371/journal.pmed.1000097

Moisan, S. & Alotaibi, F. (2025). Determinants of Generative Artificial Intelligence (GenAI) adoption among university students and its impact on academic performance: the mediating role of trust in technology.

Morjaria, L., Burns, L., Bracken, K., Levinson, A. J., Ngo, Q. N., Lee, M., & Sibbald, M. (2024). Examining the Efficacy of ChatGPT in Marking Short-Answer Assessments in an Undergraduate Medical Program.

Newton, P. M. (2025). pragmatic risk assessment.

Nikolic, S., Sandison, C., Haque, R., Daniel, S., Grundy, S., Belkina, M., ... Lyden, S. (2024). ChatGPT, Copilot, Gemini, SciSpace and Wolfram versus higher education assessments: an updated multi-institutional study of the academic integrity impacts of Generative Artificial Intelligence (GenAI) on assessment, teaching and learning in engineering.

Obed, K. & Anangisye, W. A. L. (2025). Academic integrity considerations of using ChatGPT in assessment activities among university student teachers.

Ogunleye, B., Ibilola, K., Zakariyyah, Ajao, O., Olayinka, O., & Sharma, H. (2024). A Systematic Review of Generative AI for Teaching and Learning Practice.

Oh, S. (2025). Integration of MATH41 and Generative AI in Pre-Service Mathematics Teacher Education: An Empirical Study on Lesson Design Competency.

Ouyang, F., Guo, M., Zhang, N., Bai, X., & Jiao, P. (2024). Comparing the Effects of Instructor Manual Feedback and ChatGPT Intelligent Feedback on Collaborative Programming in China's Higher Education.

Peters, M. (2025). Leveraging Generative Artificial Intelligence with Transparency: Enhancing Academic Integrity in Higher Education.

Pimbblet, K. A. & Morrell, L. J. (2024). Can ChatGPT pass a physics.

Polyportis, A. & Pahos, N. (2024). Understanding students' adoption of the ChatGPT chatbot in higher education: the role of anthropomorphism, trust, design novelty and institutional policy.

Pwanedo, J., Amodu, O. A., Azlina, R., Mahmoo, R., Bolakale, A., Faziehan, A., ... Rhoda, A. (2025). A Bibliometric Exposition and Review on Leveraging LLMs for Programming Education.

Qiu, Y. (2024). The Impact of LLM Hallucinations on Motor Skill Learning: A Case Study in Badminton.

Qu, Y., Xin, M., Tan, Y., & Wang, J. (2024). Disciplinary differences in undergraduate students' engagement with generative artificial intelligence.

Quezada-Sarmiento, P. A., Marisol, P., Cañaveral, C., Buele, J., Sabando-García, Á. R., Bosco, J., ... Sabando-García (2025). Ethical Use of Generative Artificial Intelligence Among Ecuadorian University Students.

Rafifah, N., Syallya, P., Apriliyanti, A., & Helen, A. (2025). NLP-Based Intent Classification Model for Academic Curriculum Chatbots in Universities Study Programs.

Rana, N. K. (2024). Generative AI and Academic Research: A Review of the Policies from Selected HEIs.

Riaz, S., Jain, S. P., & Mushtaq, A. (2024). Optimizing Generative AI Integration in Higher Education: A Framework for Enhanced Student Engagement and Learning Outcomes.

Rivas-Echeverr, F., Ramos, T., Ibarra, J. L., Zerpa-Bonillo, S., Arciniegas, S., Asprino-Salas, M., & Ramos, L. T. (2025). LegalBot-EC: An LLM-Based Chatbot for Legal Assistance in Ecuadorian Law.

Santhos, J., Dengel, A., Ishimaru, S., & Santhosh, J. (2024). Gaze-Driven Adaptive Learning System With ChatGPT-Generated Summaries.

Shahzad, M. F., Xu, S., & Zahid, H. (2024). Exploring the impact of generative AI-based technologies on learning performance through self-efficacy, fairness & ethics, creativity, and trust in higher education.

Skalka, J., Przyby-Kasperek, M., Smyrnovatrybulsk, E., Klimeš, C., Farana, R., Dagien, V., ... Dolgopolovas, V. (2025). Artificial Intelligence Literacy Structure and the Factors Influencing Student Attitudes and Readiness in Central Europe Universities.

Slotnick, R. C. & Boeing, J. Z. (2024). Enhancing qualitative research in higher education assessment through generative AI integration: A path toward meaningful insights and a cautionary tale.

Soledad, C., González, G., Aleksieva, L., Miranda, P., Meissner, R., Pögelt, A., ... Ihsberner, K. (2006). EDITED BY.

Song, Y., Zhu, Q., Wang, H., & Zheng, Q. (2024). Automated Essay Scoring and Revising Based on Open-Source Large Language Models.

Subhani, F., Khan, S. A., Sandhu, M. A., & Shahzad, M. F. (2025). What Factors Affect the Adoption Intention and Actual Use of ChatGPT in Higher Education? The Moderating Role of Academic Integrity.

Tesi, A., Baroncelli, A., Carlo, M. E. D., Shi, J., Liu, W., & Hu, K. (2025). Exploring How AI Literacy and Self-Regulated Learning Relate to Student Writing Performance and Well-Being in Generative AI-Supported Higher Education.

Thüs, D., Malone, S., Brünken, R., Neumann, K., & Kovács, K. E. (2024). Exploring generative AI in higher education: a RAG system to enhance student engagement with scientific literature.

Torenvliet, N. & Zelek, J. S. (2024). Diffusion Partition Consensus: Diffusion-Aided Time-of-Flight Estimates, Anomaly Detection, and Localization for Ultrasonic Nondestructive Evaluation Data.

Tossell, C. C., Tenhundfeld, N. L., Momen, A., Cooley, K., & Visser, E. J. D. (2024). Student Perceptions of ChatGPT Use in a College Essay Assignment: Implications for Learning, Grading, and Trust in Artificial Intelligence.

Tsz, S., Chan, S., Po, N., Lo, K., Man, A., & Wong, H. (2024). Enhancing university level English proficiency with generative AI: Empirical insights into automated feedback and learning outcomes.

Ulla, M. B., Advincula, J. C., Dawn, C., Mombay, S., Mae, H., Mercullo, A., ... Nacionales, J. P. (2024). How can GenAI foster an inclusive language classroom? A critical language pedagogy perspective from Philippine university teachers.

Usher, M. (2025). Assessment & evAluAtion in HigHer educAtion Generative AI vs. instructor vs. peer assessments: a comparison of grading and feedback in higher education.

Valdivieso, T. & González, O. (2025). Generative AI Tools in Salvadoran Higher Education: Balancing Equity, Ethics, and Knowledge Management in the Global South.

Villagrán, I., Hernández, R., Schuit, G., Neyem, A., Fuentes-Cim, J., Miranda, C., ... Hilliger, I. (2024). Implementing Artificial Intelligence in Physiotherapy Education: A Case Study on the Use of Large Language Models (LLM) to Enhance Feedback.

Wa, H., Wang, C., Chen, Z., Liu, F., Bao, C., & Xu, X. (2025). Impact of AI-agent-supported collaborative learning on the learning outcomes of University programming courses.

Wang, M., Wang, M., Xu, X., Yang, L., Cai, D., & Yin, M. (2024). Unleashing ChatGPT's Power: A Case Study on Optimizing Information Retrieval in Flipped Classrooms via Prompt Engineering.

Wang, R., Zhao, D., Yuan, Z., Obi, I., Figueroa, N., & Min, B. (2025). PrefCLM: Enhancing Preference-Based Reinforcement Learning With Crowdsourced Large Language Models.

Williams, A. (2024). Comparison of generative AI performance on undergraduate and postgraduate written assessments in the biomedical sciences.

Wilson, T. D. (2025). The development of policies on generative artificial intelligence in UK universities.

Woodthorpe, J., Richards, M., Waugh, K., Slaymaker, M., Petre, M., Gooch, D., ... Richards, K. (2023). Bob or Bot: Exploring ChatGPT's Answers to University Computer Science Assessment.

Wu, T., Lee, H., Chen, P., Lin, C., & Huang, Y. (2024). Integrating peer assessment cycle into ChatGPT for STEM education: A randomised controlled trial on knowledge, skills, and attitudes enhancement.

Wu, S. & Luo, M. (2025). Selection and Resource Allocation Strategies for Chatbot Technologies in Higher Education: An Optimization Model Approach.

Xing, Y. U. (2024). The Influence of Responsible Innovation on Ideological Education in Universities Under Generative Artificial Intelligence.

Yeung, S. (2025). University students' engagement with generative AI-supported automated writing evaluation.

Yusuf, A., Pervin, N., Román-González, M., Cremer, D., & Narayanan, D. (2024). International Journal of Educational Technology in Higher Education Generative AI and the future of higher education: a threat to academic integrity or reformation? Evidence from multicultural perspectives.

Zeb, A., Ullah, R., & Karim, R. (2025). Exploring the role of ChatGPT in higher education: opportunities, challenges and ethical considerations.

Zhang, Y., Tian, Z., Grieve, A., & Agyemang, M. (2025). Digital competencies in student learning with generative artificial intelligence: Policy implications from world-class universities.

Zhou, W., Zhu, X., Han, Q., Li, L., Chen, X., Wen, S., ... Xiang, Y. (2025a). The Security of Using Large Language Models: A Survey With Emphasis on ChatGPT.

Zhou, Y., Tang, C., & Shimada, A. (2025b). Extracting Learning Data From Handwritten Notes: A New Approach to Educational Data Analysis Based on Image Segmentation and Generative AI.

Zlotnikova, I., Boseja, P., & Bag, P. (2024). GenAI in the Context of African Universities: A Crisis of Tertiary Education or Its New Dawn?.

Zlotnikova, I., Hlomani, H., Mokgetse, T., & Bagai, K. (2025). Establishing ethical standards for GenAI in university education: a roadmap for academic integrity and fairness.

Disclaimer: Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them

Co-funded by the European Union