OmniaScience

Journal of Technology and Science Education

JOTSE, 2025 – 15(3): 810-833 – Online ISSN: 2013-6374 – Print ISSN: 2014-5349

https://doi.org/10.3926/jotse.3643

THE ROLE OF GENERATIVE AI CHATBOTS IN HIGHER EDUCATION: A STUDENT-CENTRIC CONCEPTUAL ANALYSIS OF BENEFITS, ETHICS, AND PRIVACY CONCERNS

Alex Vallejo-Blanxart[1], Ruben Nicolas-Sans[2*]

[1]ESIC Business & Marketing School (Spain)

[2]UNIE Universidad (Spain)

alex.vallejo@esic.edu

[*]*Corresponding author: ruben.nicolas@universidadunie.com*

Received June 2025

Accepted October 2025

Abstract

Generative AI chatbots, such as ChatGPT, are reshaping higher education by offering personalized tutoring, administrative support, and enhanced research capabilities, while simultaneously raising ethical and privacy concerns. This study examines their role from a student-centered perspective, focusing on benefits, risks, and institutional implications. A pilot study was carried out with 300 undergraduate students and 120 faculty members from five European universities. Participants engaged in semi-structured interviews, perception surveys, and controlled academic tasks designed to evaluate chatbot performance. Instruments included standardized student satisfaction questionnaires, expert review rubrics, and quantitative performance metrics (perplexity, response latency, context window, SWE-bench accuracy). Data were analyzed through descriptive statistics, ANOVA tests, and thematic coding of interviews. Results reveal that ChatGPT and Claude achieved the best balance between pedagogical clarity and privacy compliance, while Gemini excelled in technical capacity but showed weaker ethical safeguards. Student satisfaction was strongly associated with transparency in data policies and pedagogical usefulness rather than raw technical performance. These findings highlight the need for universities to adopt generative AI technologies under robust ethical and privacy frameworks. Future research should extend this pilot into longitudinal studies and institutional case analyses.

Keywords – Generative AI, Chatbots, Higher education, Ethics, Privacy, Student perceptions.

To cite this article:

Vallejo-Blanxart, A., & Nicolas-Sans, R. (2025). The role of generative AI chatbots in higher education: A student-centric conceptual analysis of benefits, ethics, and privacy concerns. *Journal of Technology and Science Education,* 15(3), 810-8yy. https://doi.org/10.3926/jotse.3643

---

1. Introduction

Generative AI chatbots have rapidly reshaped higher education by enabling personalized learning support and streamlining academic and administrative tasks. The recent literature moves from earlier work on intelligent tutoring systems to focused analyses of large language model (LLM) chatbots (e.g., ChatGPT, Claude, Gemini), consistently highlighting persistent tensions —personalization versus

privacy, efficiency versus equity, and innovation versus academic integrity. A key limitation identified is the fragmentation between studies emphasizing technical and pedagogical benefits and those foregrounding ethical and privacy risks, with comparatively few integrated frameworks to inform institutional decision-making.

Concurrently, universities have begun to adopt internal guidelines on AI use (ethics, data privacy, AI literacy), yet the absence of standardized governmental regulation complicates the management of academic integrity, data protection, and equitable access. Case reports and institutional experiences underscore the need for transparent governance, periodic audits, and fit-for-purpose policies to mitigate plagiarism, digital divide effects, and inconsistent practices.

Debates in the field converge on three ethical fronts: (i) authorship and plagiarism for AI-mediated work, with proposals to update definitions and reinterpret honor codes; (ii) bias and discrimination arising from training data and model behavior; and (iii) opacity in data collection, retention, and consent under GDPR/AI Act principles —prompting calls for privacy-by-design and robust deletion/consent mechanisms.

1.1. Prior Evaluation Metrics and Justification for Those Used in This Study

Comparative evaluation of chatbots in higher education benefits from combining standard LLM performance metrics with pedagogical, ethical, and psychological criteria. Following established practice in NLP/ML, we include perplexity (PPL) for output fluency, response latency for service responsiveness, context window for long-context handling, and SWE-bench accuracy for code-problem solving. These quantitative indicators are integrated in a multi-dimensional scorecard (Technical, Pedagogical, Ethical, Psychological) with Analytic Hierarchy Process (AHP) weights to ensure transparent aggregation aligned with institutional priorities.

To bridge the gap between technical performance and educational/ethical impact, we complement metrics with expert review rubrics (Content Quality Index; Ethical Risk Scale) and inter-rater reliability checks (Cohen's κ). We also operationalize privacy/ethics via an Ethical Compliance Index (ECI) that consolidates four binary indicators —identity linkage, anonymous mode availability, deletion controls, and consent mechanisms. This design yields a reproducible, education-relevant framework that connects model behavior to classroom and institutional concerns.

1.2. Knowledge Gaps, Study Strengths, and Objectives

Knowledge gaps addressed

a) Lack of integrated analyses that jointly consider pedagogical benefits and ethical/privacy risks;
b) Limited application of standard LLM metrics to higher-education contexts with clear links to educational outcomes;
c) Insufficient evidence connecting student perceptions, expert review, and privacy safeguards within a single comparative design.

Study strengths

- Mixed-methods design combining standardized quantitative metrics with structured expert evaluation across multiple domains; Inclusion of both students and faculty from several European universities; Transparent, reproducible scoring framework with AHP weighting and an operationalized ECI to assess privacy/ethics.

General objective

- **GO:** To conceptualize the role of generative AI chatbots in higher education from a student-centered perspective, balancing benefits with ethical and privacy challenges to inform institutional decision-making.

Specific objectives

- **SO1:** Analyze student-facing benefits and limitations of generative AI chatbots.
- **SO2:** Evaluate ethical concerns and privacy compliance using explicit indicators.
- **SO3:** Compare chatbot performance using standard technical metrics and expert rubrics.
- **SO4:** Derive institutional recommendations for responsible adoption in higher education.

2. Literature Review

Over the past decade, scholarship on AI in education has evolved from exploratory work on intelligent tutoring systems to closer examinations of large-language-model (LLM) chatbots used for learning support and academic workflows. Early contributions highlighted AI's potential for adaptive feedback and personalized instruction and stressed aligning algorithmic recommendations with pedagogical aims. More recent studies focus on generative chatbots in university settings, mapping digital access barriers, proposing governance frameworks that balance technical affordances with institutional ethics, and critiquing the "datafication" of teaching and its implications for academic agency. Bringing these strands together, the literature converges on three persistent tensions that motivate our study: personalization vs. privacy, efficiency vs. equity, and innovation vs. academic integrity. Building on this, our work advances an integrated analytical model that links technical performance indicators with pedagogical value and ethical –privacy safeguards.

2.1. AI Chatbots

LLM-based chatbots rose to mainstream prominence with the public release of ChatGPT in late 2022, part of the GPT family of transformer models. These systems generate human-like text and support diverse tasks —including question answering, code assistance, summarization, and drafting— within multi-turn dialogues. Two features distinguish modern LLM chatbots from earlier rule- or retrieval-based systems: (i) training on web-scale corpora that endow broad linguistic competence, and (ii) alignment procedures (e.g., human-feedback-driven fine-tuning) that improve conversational usefulness and safety.

From a usage perspective, multiple platforms now coexist (e.g., ChatGPT, Claude, Gemini, Copilot, Perplexity, DeepSeek). Public interest has been fluid across time and regions; in our Google Trends snapshot for Spain (Sep 2022-Feb 2025), searches for "ChatGPT" remain dominant while queries for "DeepSeek" increase notably toward late 2024 (see Figure 1). For the purposes of this study, we compare leading chatbots available in European university contexts, attending not only to technical characteristics (e.g., context-window size, response latency) but also to pedagogical qualities and privacy/ethical safeguards —consistent with the multi-domain evaluation framework employed in our Methods and Results sections.

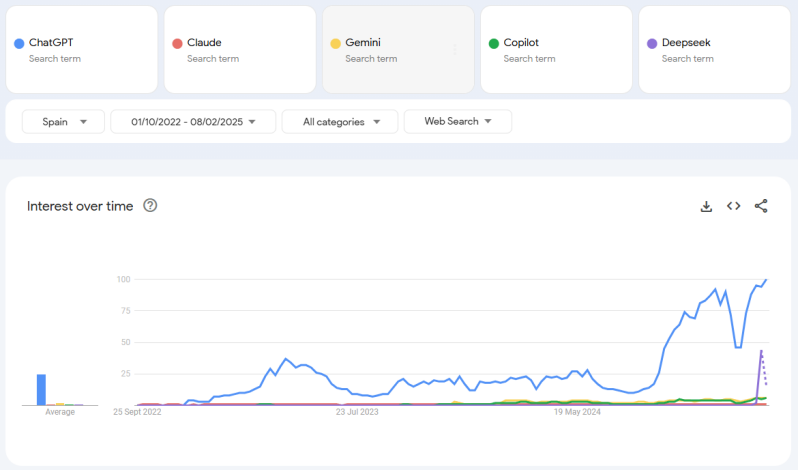

Figure 1. Popularity of "ChatGPT", "Claude", "Gemini", "Copilot" and "Deepseek" search terms over time (Sep 2022 to Feb 2025) in Spain in Google Trends

From the usage point of view, ChatGPT has a massive advantage over its competitors. Table 1 ranks the most widely used AI chatbot technologies and shows that ChatGPT is the market leader in web-based AI chatbots worldwide, but the second most used is Ernie, the Chinese chatbot, although in the Western world Deepseek has become very popular. AI chatbots can be used directly in their websites but also from another software using the API for developers. Claude AI is the most used API in early 2025 (Anthropic PBC, 2025), but because this paper is focused on student use Table 1 is focused on the former.

AI chatbots continuously release new models at an incredibly fast pace, offering new features and capabilities. One example of this is that Open AI has made 42 models obsolete from the end of 2022 (Open AI, n.d.) to February 2025.

	Company	Chatbot LLM family	Open source?	Estimated website monthly visits (Feb 2025)
ChatGPT (OpenAI, 2022)	OpenAI, USA	GPT	No	4.672,8M
Ernie	Baidu, China	Ernie	No	307,9M
Deepseek (Deepseek AI, 2025)	DeepSeek AI, China	Deepseek	Yes	277,9M
Gemini (Google Inc, 2025)	Google DeepMind, USA/UK	Gemini	No	275,4M
Perplexity (Perplexity AI Labs, 2025)	Perplexity, USA	Mixed	No	132,6M
Claude (Anthropic PBC, 2025)	Anthropic, USA	Claude	No	105,2M
Copilot (Microsoft, 2025)	Microsoft, USA	GPT	No	67,3M
Kimi	Moonshot AI, China	Kimi	No	35,1M
Doubao	ByteDance, China	Doubao	No	33,5M

Table 1. Rank of AI Chatbots by website monthly visits (as of February 2025). Traffic data is from http://aicpb.com (for Chinese models) and http://semrush.com (for USA models)

2.1.1. User Privacy and Data Collection in AI Chatbots

User privacy and data protection are foundational to trustworthy, ethical, and legally compliant use of chatbots in higher education (Sebastian, 2023). In the European context, compliance with the General Data Protection Regulation (GDPR) is critical (Nayak, Pasumarthi, Rajagopal & Verma, 2024), and the evolving Artificial Intelligence Act further differentiates obligations relevant to general-purpose/basic models (e.g., ChatGPT) and broader AI systems (Wang, 2024). These frameworks address concerns about large-scale internet data ingestion without explicit consent and the potential re-use of user prompts that could expose sensitive information (Wang, 2024).

There is, to date, no widely recognized quantitative benchmark that comprehensively compares the privacy-friendliness of leading AI chatbots. The literature remains largely qualitative, with notable analyses such as Sebastian (2023) and Zhan et al. (2024) mapping risks and proposing safeguards. This gap motivates our operational indicators in the present study. (Sebastian, 2023; Zhan, Seymour & Such, 2024).

Most mainstream chatbots process interactions on cloud infrastructure due to computational/energy demands, which enables provider-side access to conversations and creates privacy risk vectors. By contrast, local deployments or open-source/open-weights options can reduce external exposure by minimizing or eliminating internet connectivity. Within commonly used systems, Deepseek is presented as open source (Deepseek AI, 2025), and the Llama family (Meta AI, 2023) provides open weights used by some services (e.g., in the Perplexity ecosystem), though adoption levels vary across regions and use cases. (Deepseek AI, 2025; Meta AI, 2023).

Data collection practices are pivotal. LLMs such as ChatGPT are trained on extensive, heterogeneous web corpora; however, exact training datasets are undisclosed and inaccessible during operation (Sebastian, 2023). Post-deployment, interactions may be retained for limited periods (e.g., 30 days) to improve services (Tlili, Shehata, Agyemang, Bozkurt, Hickey, Huang et al., 2023). Multimodal features further complicate privacy because images, audio, or embedded metadata can inadvertently capture

third-party information —hence the call for collective (not only individual) privacy safeguards (Zhan et al., 2024; Sebastian, 2023; Tlili et al., 2023).

Account-level identity linkage represents another concern: many platforms require email or phone verification, binding activity to personal identifiers (Gumusel, Zhou & Sanfilippo, 2024). This constrains truly anonymous use and heightens risk if logs are repurposed for training or targeted profiling (Williams, 2024). Although some services offer anonymous or temporary modes, these often limit access to advanced features —illustrating a trade-off between functionality and stronger privacy protections (Tlili et al., 2023; Sebastian, 2023; Gumusel et al., 2024; Williams, 2024).

To assess privacy-friendliness for student use, we operationalize four dimensions aligned with the concerns above: (1) identity-verification requirements; (2) availability and limitations of anonymous/temporary modes; (3) conversation-log controls and data-retention transparency; and (4) consent mechanisms for collection, reuse, and sharing (Sebastian, 2023; Wang, 2024; Gumusel et al., 2024; Tlili et al., 2023). These indicators are aggregated in our Ethical Compliance Index (ECI) and analyzed alongside technical and pedagogical metrics (see Methods §3.6). (Sebastian, 2023; Wang, 2024; Gumusel et al., 2024; Tlili et al., 2023).

Finally, emerging work characterizes concrete user privacy harms —e.g., dark-pattern interfaces, opaque data flows, and misunderstandings of how LLMs function— which can intensify disclosure risks in academic contexts (Gumusel et al., 2024; Zhan et al., 2024). These findings reinforce the need for transparent defaults, explicit consent, and robust deletion tools in university deployments. (Gumusel et al., 2024; Zhan et al., 2024).

2.2. Use of AI in Higher Education

Artificial intelligence is reshaping higher education across pedagogy and administration. Universities deploy AI to enable personalized learning —with intelligent tutoring systems adapting content to individual needs and supporting retention and performance— and to streamline operations such as grading, enrolment, and student-success prediction; institution-facing chatbots further enhance efficiency by handling academic and administrative queries (O'Donnell, Porter & Fitzgerald, 2024; Chadha, 2024; Boratkar & Sambhe, 2024; Yunusov, Berdiyev & Jovliev, 2024).

At the institutional level, multiple initiatives illustrate this trajectory. For example, the University of Murcia employs AI chatbots to support academic inquiries and reduce response times; leading institutions (e.g., MIT, Stanford) report adaptive platforms that tailor coursework pacing. In parallel, universities are releasing internal policies on AI—ethical guidelines, data-privacy protocols, and AI-literacy programs (Universidad Complutense de Madrid, 2024). Yet government-level regulation remains uneven, leaving gaps around academic integrity, data protection, and equitable access (Davydova & Shlykova, 2024). This has spurred local responses: the rise of AI-generated content has prompted efforts to detect plagiarism and automate provenance checks (Díaz-Arce, 2024), while concerns about unequal access to AI tools highlight the digital divide that disadvantages students with limited resources (Bennett & Abusalem, 2024).

The literature therefore frames a dual agenda. On one side are documented benefits—administrative efficiency, personalized instruction, and improved accessibility for students with disabilities (Micheni, Machii & Murumba, 2024). On the other are risks: extensive data collection raises privacy concerns (Rudolph, Ismail & Popenici, 2024); overreliance on AI may dampen critical thinking and self-directed learning (O'Donnell et al., 2024); and algorithmic bias risks reproducing inequalities (Chadha, 2024). To reconcile these tensions, authors call for transparent governance, comprehensive ethical frameworks, and digital-literacy initiatives that ensure fair deployment and equitable outcomes (Bennett & Abusalem, 2024). This article responds by evaluating chatbots across technical, pedagogical, and ethics/privacy dimensions, providing evidence to support institutional decision-making.

2.2.1. Ethics and Privacy in AI chatbots in Higher Education

It is essential to acknowledge the ethical limitations of AI chatbots in university contexts and to outline mitigation strategies. A central concern is academic integrity: chatbot outputs complicate authorship and accountability in coursework and examinations—raising questions about whether responsibility lies with the student, the system, or uncited source material embedded in model behavior. The debate intensified when ChatGPT was listed as an author on research papers, and major journals responded with explicit ground rules or bans (Stokel-Walker, 2023; Nature, 2023). Proposed responses include expanding the definition of plagiarism to encompass generative-AI content and reinterpreting honor codes so that AI is framed as a collaborative tool rather than a shortcut (Barnett, 2023; Ryan, 2023).

This is not the first time universities have recalibrated norms in response to new tools. Since Wikipedia's launch, institutions have reworked policies around credibility, sourcing, and student use; subsequent research found nuanced, context-dependent credibility rather than blanket rejection (Knight & Pryke, 2012; Messner & DiStaso, 2013). Unlike Wikipedia, however, chatbot outputs are often treated as original text, which heightens the integrity challenge and underscores the need for explicit authorship guidance and disclosure practices.

Beyond integrity, the literature highlights risks of bias and discrimination. LLMs inherit patterns from training data, echoing earlier evidence of inequities in other digital systems such as search engines and speech recognition (Noble, 2018; Koenecke, 2020). Developers increasingly apply alignment and learning strategies to mitigate harm, with a focus on inclusivity for marginalized groups, but ongoing audits and oversight are recommended to ensure responsible classroom and institutional deployment (Ryan, 2023; Williams, 2024).

Privacy and confidentiality remain core concerns. The continuous collection and processing of data by chatbots intersect with evolving European regulatory frameworks (GDPR and the AI Act), and rapid system changes have, at times, pushed policymakers "back to the drawing board" on AI regulation (Volpicelli, 2023). Universities therefore need clear policies on collection, storage, use, and deletion of data —particularly to uphold the right to be forgotten (Williams, 2024).

Finally, broader reflections on "Dataism" caution that data-centric logics may overextend into educational practice, fostering overreliance on technological mediation while obscuring human judgment and pedagogical aims (Harari, 2016). Together, these strands offer a framework for evaluating current practice and designing guidelines to ensure responsible use of chatbots in academic environments.

3. Methodology

This study follows a comparative, mixed-methods design integrating quantitative and qualitative techniques to evaluate four domains of chatbot use in higher education—Technical, Pedagogical, Ethical, and Psychological. The design combines measurable performance metrics with structured expert appraisal to produce a reproducible framework that can inform institutional policy and pedagogical practice.

A multi-domain scorecard aggregates standardized indicators in each domain using min-max normalization and Analytic Hierarchy Process (AHP) weights elicited from a panel of experts, enabling transparent trade-offs aligned with institutional priorities (see §3.2). Quantitative metrics include perplexity (PPL), response latency, context-window capacity, and SWE-bench accuracy; qualitative expert review applies rubric-based judgments with inter-rater reliability (Cohen's κ). Ethical/privacy compliance is operationalized via a four-indicator Ethical Compliance Index (ECI) (identity linkage, anonymous-mode availability, deletion controls, consent mechanisms).

3.1. Data Collection and Selection Criteria

Chatbot selection. We evaluated six widely used platforms with availability in European university contexts —ChatGPT, Gemini, Perplexity AI, Claude, Copilot, and Deepseek— selected on the basis of global web-traffic signals, documented academic/professional use, and campus accessibility. Platforms limited to

Chinese-language interfaces (e.g., Ernie, Kimi, Doubao) were excluded. For each chatbot, we collected free-tier and subscription-tier observations to capture the full functionality range.

Pilot rationale and sample. A pilot involving 300 students and 120 faculty across five European institutions was conducted to (i) validate instruments, (ii) calibrate domain weights and thresholds, and (iii) check feasibility and variance ahead of full analysis. The student N enables stable estimation of distributional properties and subgroup contrasts (e.g., by prior AI exposure) while keeping fieldwork feasible across participating campuses.

Eligibility criteria:

- Students: enrolled at a participating university, ≥18 years, not on academic leave, and with at least incidental exposure to AI chatbots in the previous 6 months (coursework or self-study).
- Faculty: active teaching/research appointment (≥0.5 FTE) during the study term and direct or planned interaction with AI chatbots for teaching, supervision, or administration.
- Exclusion: inability to provide informed consent or to complete instruments in the study language(s).

Instruments and sources. Data were compiled from official developer materials (technical whitepapers/release notes) for architectural/version insights; from peer-reviewed studies on LLM benchmarks and ethics; and from original field instruments: (a) semi-structured interviews and perception surveys with students and faculty; and (b) controlled tasks using a standardized 50-prompt set (academic writing, data-privacy scenarios, ethics case studies) for head-to-head evaluation (see also §3.2-3.4).

Pilot instrument (survey). The student/faculty survey captured: demographics; prior chatbot use; usability and pedagogical value scales; perceived privacy risk and policy awareness items; and open responses on benefits/concerns. (The survey underpins descriptive results and complements rubric-based expert review; see §3.4 for the qualitative panel and reliability check.)

Interview protocol. Semi-structured interviews (≈30 minutes) followed four sections: (1) usage patterns and tasks; (2) perceived benefits/learning impact; (3) privacy/ethics concerns and safeguards; (4) institutional policies and support needs. Sessions were audio-recorded with consent and transcribed verbatim.

Procedure for controlled experiments. Each platform was tested against the same 50 prompts, executed under comparable network conditions. Outputs were collected for quantitative scoring (PPL, latency, context window, SWE-bench accuracy) and for blind expert review using the Content Quality Index (CQI) and Ethical Risk Scale (ERS) rubrics (see §3.3-3.4).

Qualitative analysis and reliability. Two trained coders conducted thematic analysis on interview transcripts and open-ended survey responses, supported by a qualitative analysis tool (e.g., NVivo). Coding disagreements were resolved by discussion; inter-rater reliability for rubric-based expert judgments over 300 chatbot-student interactions were quantified using Cohen's κ, with high agreement reported (see §3.4).

Quantitative analysis. Technical metrics were computed as specified in §3.3. Group differences across chatbots were examined via one-way ANOVA with Tukey's HSD post-hoc comparisons; where assumptions were violated (Shapiro-Wilk $p < .05$), Kruskal-Wallis tests were used. Effect sizes are reported as Cohen's κ (pairwise) and η^2 (omnibus).

Ethical/privacy operationalization. Following GDPR-aligned concerns, the ECI aggregates four binary indicators —identity linkage, anonymous-mode availability, data-deletion controls, consent mechanisms— and includes simulated privacy-breach scenarios to probe model behavior on sensitive queries (see §3.6).

Data triangulation. Survey/Interview evidence (students/faculty) was triangulated with expert reviews and platform metrics, linking perceptions to measurable performance and explicit privacy safeguards; student satisfaction findings are summarized in §4.5.

3.2. Analytical Framework

We constructed a multi-domain scorecard spanning four domains —Technical, Pedagogical, Ethical, Psychological— each represented by a normalized composite index. Let x_{ij} be the raw score for chatbot i on metric j. To ensure commensurability, each metric is min-max normalized with directionality control:

If higher is better for metric *j:*

$$z_{ij} = \frac{x_{ij} - \min(x_{\cdot j})}{\max(x_{\cdot j}) - \min(x_{\cdot j}) + \varepsilon}$$

If lower is better (e.g., perplexity, latency):

$$z_{ij} = \frac{\max(x_{\cdot j}) - x_{ij}}{\max(x_{\cdot j}) - \min(x_{\cdot j}) + \varepsilon}$$

Where ε is a small constant to avoid division by zero when observed ranges collapse. Metrics used in this study are detailed in §3.3 Quantitative Performance Metrics.

For each domain d, we compute a domain composite for chatbot i as a weighted sum of its normalized indicators in that domain:

$$D_i^{(d)} = \sum_{j \in J_d} w_{j|d}\, z_{ij}, \quad \text{with} \sum_{j \in J_d} w_{j|d} = 1.$$

Weights $w_{j|d}$ were elicited via the Analytic Hierarchy Process (AHP) from a five-member expert panel (three instructional designers, two ethicists), following standard pairwise-comparison procedures (Saaty, 1980). We report these domain-level weight vectors in Results when interpreting contributions of individual indicators.

An overall score aggregates the four domain composites according to institutional priorities:

$$S_i = \alpha\, D_i^{(Tech)} + \beta\, D_i^{(Ped)} + \gamma\, D_i^{(Eth)} + \delta\, D_i^{(Psy)}, \quad \alpha + \beta + \gamma + \delta = 1,$$

where {α,β,γ,δ} reflect priority settings derived from the same AHP exercise (and are varied in sensitivity checks). This structure aligns directly with the study objectives and the reporting in §4 (Figures 2-3).

Link to measurement and inference. Technical indicators (e.g., perplexity, response latency, context window, SWE-bench accuracy) enter the Technical composite; rubric-based expert judgments (CQI, ERS) inform Pedagogical and Ethical composites with inter-rater reliability assessed via Cohen's $\varkappa$ (see §3.3-3.4). Between-chatbot contrasts on normalized domain scores use ANOVA/Tukey or Kruskal-Wallis with effect sizes as specified in §3.5. The Ethical domain also incorporates the four binary indicators (identity linkage, anonymous mode, deletion controls, consent) aggregated as the Ethical Compliance Index (ECI) described in §3.6.

Scope and coherence. Consistent with reviewer guidance to delimit the research design, the scorecard targets education-relevant properties (performance, learning value, ethics/privacy, psychological engagement) and excludes cost, energy, or deployment-engineering metrics to maintain focus and interpretability within higher-education decision-making. Results are structured around these domains and mapped back to the specific objectives in the Discussion and Conclusion.

3.3 Quantitative Performance Metrics

We assessed four core technical indicators —Perplexity (PPL), Response Latency (RL), Context Window (CW), and SWE-bench Accuracy (SA)— selected for their relevance to higher-education use cases and their complementarity with the qualitative rubrics and the Ethical Compliance Index (ECI). All raw metrics feed the domain composites after min-max normalization with directionality control as specified in §3.2.

(A) Perplexity (PPL).

Construct. Fluency/predictability of generated text.

Directionality. Lower is better.

Operationalization. For each chatbot, we generated outputs to the standardized 50-prompt set (academic writing, privacy scenarios, ethics cases; see §3.1). We then computed PPL by scoring each output under a fixed reference tokenizer/probability model and aggregating token-level log-likelihoods into sequence perplexity; system-level PPL is the mean across prompts. This approach yields a comparable proxy of linguistic coherence across chatbots even when native token-probabilities are not exposed in UI mode.

Reporting. PPL is reported in its natural scale (and inverted during normalization).

(B) Response Latency (RL)

Construct. Responsiveness under typical load conditions.

Directionality. Lower is better.

Operationalization. For each prompt-chatbot pair, we measured wall-clock time from request dispatch to first token (stream onset) in seconds. Each prompt was executed in fresh sessions (to control for caching/history) and repeated in replicate runs; we report trimmed means (5 % trim) per system to reduce outlier influence from transient network spikes.

Controls. Uniform network conditions and identical execution workflow were used across platforms.

(C) Context Window (CW)

Construct. Maximum token capacity supported per interaction (proxy for ability to digest long documents/syllabi). Directionality. Higher is better.

Operationalization. We recorded the maximum tokens accepted by each chatbot (model/context configuration) from vendor documentation and empirical confirmation via staged-length prompts until failure. We report the effective CW observed (tokens), which then enters normalization.

(D) SWE-bench Accuracy (SA)

Construct. Capacity to resolve real-world coding problems relevant to STEM coursework.

Directionality. Higher is better.

Operationalization. We embedded a curated subset of SWE-bench-style tasks in the 50-prompt battery and scored pass/fail per task using a reference harness (unit/expected-output checks). System-level SA is the proportion correct [0,1]. (Tasks are described in the prompt inventory; see §3.1.)

Additional telemetry. To contextualize throughput and model heterogeneity, we also logged usage limits (e.g., messages/hour) and model/version identifiers (e.g., distinguishing GPT-4 from GPT-3.5) per vendor session, which inform interpretation and sensitivity checks but are **not** scored directly.

Data quality and missingness. Refusals/safety blocks were coded as task failures for SA and included in latency (measured until refusal message); hard timeouts were capped at a predefined threshold and marked as failures. All quantitative indicators were subsequently min-max normalized (lower-is-better metrics inverted) before entering the Technical domain composite (§3.2), which is later analyzed alongside Pedagogical, Ethical (incl. ECI), and Psychological domains in §4.

$$\mathrm{PPL} = \exp\left(-\frac{1}{N}\sum_{t=1}^{N}\log p(w_t \mid w_{<t})\right)$$

Where N is the number of tokens in the test sequence and $p(w_t|w_{<t})$ is the model's predicted probability for token wt. Context Window (CW) capacity was recorded as the maximum number of tokens N the model can process in a single prompt; approximately N=100 tokens correspond to 75 English words. Response Latency (RL) reflects the average processing time in seconds per API call, capturing responsiveness under typical load conditions. SWE-bench Accuracy (SA) quantifies each system's ability to solve real-world coding problems, reported on a normalized scale from 0 to 1.

In addition to these core metrics, we documented each chatbot's usage limits (messages per hour) to assess throughput constraints, as well as model versioning details (for example, distinguishing GPT-4 from GPT-3.5) to account for differences in underlying architectures and release dates.

3.4. Qualitative Expert Analysis

Purpose and linkage. The qualitative strand complements the quantitative metrics by evaluating each chatbot's pedagogical value and ethical risk using a structured expert rubric, and by analyzing semi-structured interviews with students and faculty. Scores feed the Pedagogical and Ethical domain composites in the scorecard (§3.2) and align with the Ethical Compliance Index (ECI) (§3.6).

Expert panel and blinding. A panel of five experts—three instructional designers and two ethicists—conducted blind reviews of 300 anonymized chatbot-student interactions sampled from the 50-prompt battery (uniform prompts across platforms; see §3.1). Reviewers received de-identified transcripts; platform names and UI artifacts were masked to minimize expectancy effects.

Rubrics and criteria. We employed a two-part rubric: Content Quality Index (CQI) and Ethical Risk Scale (ERS). CQI rated *pedagogical relevance*, *factual accuracy*, *structure/clarity*, and *actionability* on anchored Likert scales; ERS assessed *bias/discrimination cues*, *privacy disclosures/opacity*, and *policy transparency* (e.g., data handling, consent cues). Each interaction received both CQI and ERS scores; item-level definitions and anchors were provided in a rater manual.

Coder training, calibration, and reliability. Before main coding, raters completed a calibration round (practice set) and discussed borderline cases to harmonize interpretations. Inter-rater reliability was quantified with Cohen's κ on a held-out subset; κ is computed as observed agreement and P_e chance agreement. Reliability for the main review is reported in Results §4.4 (high agreement).

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

Interview component (type, sections, procedure). To capture user perspectives, we conducted semi-structured interviews (~30 min) with students (n = 300) and faculty (n = 120) across five European institutions. The guide comprised four sections: (1) usage patterns and tasks; (2) perceived learning benefits/limitations; (3) privacy/ethics concerns and safeguards; (4) institutional policies, training, and support needs. Sessions were recorded with consent and transcribed verbatim for analysis.

Qualitative analysis (method, software, trustworthiness). Interview transcripts and open-ended survey responses underwent **thematic analysis** using a qualitative analysis tool (e.g., NVivo). Two coders performed double-coding on an initial subset to refine a mixed (deductive/inductive) codebook grounded

in the rubric constructs; disagreements were resolved by discussion with an audit trail of decisions. We ensured credibility (peer debrief, triangulation with expert scores and metrics), dependability (versioned codebook), and confirmability (memos, reflexive notes).

Sampling and decision rules. Interactions were stratified by prompt category (academic writing, privacy scenarios, ethics cases) and platform tier (free vs. paid) to preserve coverage. Refusals or policy blocks were retained as data and scored per rubric guidance (see §3.3 for quantitative handling of timeouts/refusals).

Integration with the scorecard. CQI contributes to the Pedagogical composite; ERS contributes to the Ethical composite alongside the four binary privacy indicators in the ECI (identity linkage, anonymous mode, deletion controls, consent). Normalized rubric scores enter the domain composites and are analyzed with the statistical procedures in §3.5.

Outputs referenced in Results. Domain-level findings and exemplar patterns from expert judgments are summarized in §4.4 Qualitative Expert Review and triangulated with student satisfaction and technical metrics (§4.3-4.5).

3.5. Statistical Procedures

Analysis plan and units. Unless otherwise noted, inferential tests compare chatbots ($k = 6$) on (a) normalized domain composites (Technical, Pedagogical, Ethical, Psychological; see §3.2) and (b) raw technical indicators (Perplexity, Response Latency, Context Window, SWE-bench Accuracy; see §3.3). Where repeated measures exist (the same 50 prompts issued to every chatbot), we treat prompt as a blocking factor in robustness checks (mixed-effects models; see below).

Assumption checks. For each dependent variable we test normality (Shapiro-Wilk on residuals) and homogeneity of variances (Levene's). If assumptions fail, we (i) apply monotone transforms suited to the construct (e.g., log for latency; Box-Cox for perplexity), or (ii) use non-parametric procedures. All tests are two-sided with $\alpha=.05$.

Primary omnibus tests

- Parametric: one-way ANOVA across chatbots for each outcome.
- Non-parametric fallback: Kruskal-Wallis (H) when assumptions are violated after transformation.
- For outcomes naturally on ranks (e.g., rubric totals when Likert behavior departs from interval properties), we prefer Kruskal-Wallis directly.

Post-hoc contrasts

- Parametric: Tukey's HSD controls family-wise error for all pairwise comparisons.
- Non-parametric: Dunn's tests with Holm adjustment.

  We report adjusted p-values and 95% CIs for mean (or median) differences.

Effect sizes

- Omnibus ANOVA: η^2 (and ω^2 in sensitivity analyses).
- Pairwise (parametric): Cohen's $\varkappa$ (Hedges' g small-sample correction).
- Kruskal-Wallis: ϵ^2.

  All effect sizes include 95% bootstrap CIs (5,000 resamples).

Multiplicity control. For **secondary endpoints** (e.g., usage-limit telemetry) and exploratory contrasts, we apply Benjamini-Hochberg FDR at $q = .10$. Primary outcomes (domain composites and the four core technical indicators) remain controlled via Tukey/Holm as above.

Mixed-effects robustness. To account for the repeated-prompt structure, we re-estimate key outcomes with linear mixed-effects models:

$$\text{Outcome}_{ij} = \mu + \text{Chatbot}_i + (1|\text{Prompt}_j) + \varepsilon_{ij},$$

And for non-Gaussian outcomes, generalized or rank-based mixed models. Fixed-effect inferences are compared with ANOVA/Kruskal-Wallis results for convergence.

Outliers, refusals, and missingness. Latency outliers are mitigated via 5% trimmed means (also reported in §3.3). Refusals/timeouts are coded as task failures in SWE-bench accuracy; for latency, the measured time up to the refusal message is retained. Missing data are rare and handled via listwise deletion at the observation level (prompt × chatbot), with counts reported.

Sensitivity analyses. We (a) vary AHP domain weights by ±10% and recompute overall scores; (b) re-run analyses using free-tier only; (c) drop the top/bottom 5 prompts by difficulty to assess prompt-set dependence; and (d) repeat tests after excluding safety-blocked interactions. Conclusions are discussed only when direction and significance are stable across specifications.

Software and reproducibility. Analyses were performed in R (ANOVA/Tukey, Dunn/Holm, mixed-effects) and Python (SciPy/statsmodels for distributional tests; bootstrap CIs). Scripts and codebooks are archived with versioning; results are replicable from the anonymized dataset and prompt inventory.

3.6. Ethical and Privacy Assessment

Construct and purpose. We operationalize privacy/ethics as observable platform features and behaviors that matter in university use. Indicators are scored and aggregated into an Ethical Compliance Index (ECI) that is analyzed alongside the Technical, Pedagogical, and Psychological domains.

3.6.1. Indicators and Coding Rules

We define four **binary** indicators (1 = present/compliant; 0 = absent/non-compliant). Each indicator has explicit evidence requirements:

1. **Reduced identity linkage.** The platform supports use without binding conversations to a persistent personal identifier (e.g., optional verification; pseudonymous campus SSO; or a documented "no account"/guest mode with local/session-scoped storage). *Evidence:* UI flow and/or vendor policy text.
2. **Anonymous/temporary mode.** A built-in mode that (a) does not log prompts to provider accounts by default, and (b) clearly communicates its limitations. *Evidence:* toggle/switch in UI + policy text.
3. **Deletion and retention controls.** Users can view, export, and delete conversation logs; the provider discloses retention windows and training-use defaults, with in-product access to controls. *Evidence:* settings pathways + policy text.
4. **Explicit consent mechanisms.** Default opt-out from training use (or opt-in only), with clear purpose-of-use notices and just-in-time consent for sensitive uploads. *Evidence:* first-run dialog(s), banners, or settings.

The ECI is computed.

$$\text{ECI}_i = \frac{I_{1i}+I_{2i}+I_{3i}+I_{4i}}{4} \in [0,1].$$

Indicators are binary by design to prioritize clarity for institutional decisions; qualitative notes are archived but not scored.

3.6.2. Vendor-Policy and UI Audit Procedure

Two reviewers independently triangulated each indicator using (i) product UI walkthroughs (settings, first-run dialogs, export/delete flows) and (ii) the most current developer documentation available during data collection. Discrepancies were resolved by adjudication with a third reviewer; all determinations were evidence-backed (screenshots/policy excerpts stored in the audit log). Inter-rater agreement for indicator coding was assessed with Cohen's κ on the four binary items per platform (reported in Results §4.4).

3.6.3. Privacy-Breach Scenario Tests

To evaluate behavior beyond stated policy, we executed a taxonomy of sensitive scenarios within the standardized prompt set (see §3.1), covering:

PII exposure (self/third-party identifiers, contact data),

Special-category data (health/biometric; minors),

Academic-record data (grades, coursework with names),

Financial credentials, and

Third-party uploads (images/documents with bystanders or hidden metadata).

For each scenario × platform, two ethicist-reviewers coded outcomes using a four-level rubric: *policy-consistent refusal with explanation* (best), *safe redirection*, *ambiguous/insufficient safeguard*, *unsafe disclosure/action*. Disagreements were reconciled by discussion; κ is reported with the qualitative results (see §4.4). These rubric outcomes are **not** part of the binary ECI but are referenced in the **Ethical** domain narrative and in sensitivity checks (below).

3.6.4. Aggregation and linkage to the scorecard

Per-platform ECI values enter the Ethical domain alongside ERS (Ethical Risk Scale) scores from the expert review (§3.4). All inputs are min-max normalized before domain aggregation (§3.2). We report both the ECI and ERS contributions when interpreting ethical results in §4 (Figures 2-3).

3.6.5. Sensitivity Analyses

We probe robustness by: (i) re-computing the Ethical domain with ECI-only and ERS-only variants; (ii) treating breach-scenario outcomes as a fifth ordinal indicator (after rescaling) and re-estimating composites; and (iii) re-scoring indicators under a stricter rule requiring default-on safeguards for a value of 1 (vs. user-configurable). Conclusions are discussed only if direction and significance are stable across specifications.

3.6.6. Ethics, Consent, and Data Handling

All interviews/surveys were conducted with informed consent, and transcripts were anonymized prior to analysis. Platform audits used only provider-facing interfaces and public documentation; no private user data were collected beyond the study participants. Audit artifacts (screenshots/policies), coding sheets, and decision logs are archived for reproducibility.

4. Results

We report findings aligned with the study's specific objectives: (SO1) Analyze student-facing benefits and limitations of generative AI chatbots; (SO2) Evaluate ethical concerns and privacy compliance using explicit indicators; and (SO3) Compare chatbot performance using standard technical metrics and expert rubrics. Figures referenced below correspond to those in the manuscript.

4.1. (SO1) Analyze Student-Facing Benefits and Limitations of Generative AI Chatbots

The multi-domain scorecard shows Gemini leading overall (0.82), followed closely by ChatGPT (0.80), Copilot (0.79), Claude (0.78), Perplexity (0.77), and Deepseek (0.68) (Figure 2). ChatGPT scores highest on pedagogical clarity (expert rating M = 4.6/5), with Claude noted for long-document handling and structured summaries; Gemini excels in creative analogies but is occasionally speculative. These patterns map to student-facing benefits (clear explanations, handling of lengthy materials) and limitations (over-cautious examples; speculative outputs).

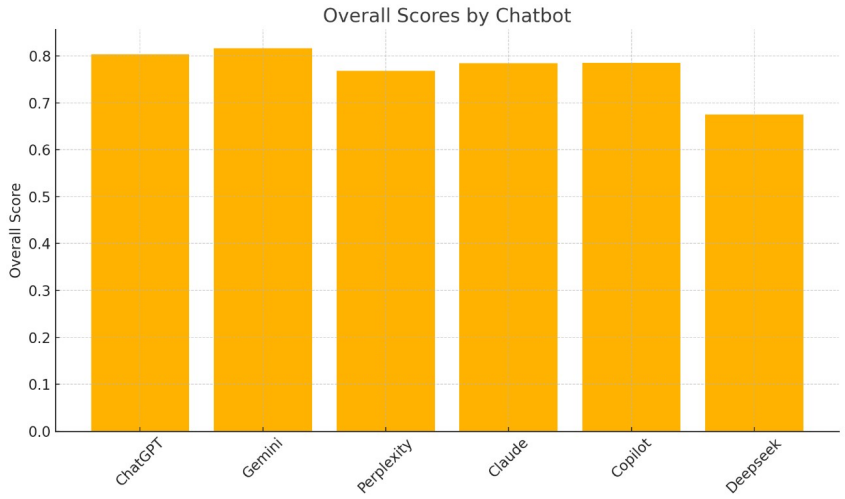

Figure 2. Comprehensive comparison between different chatbot AI technologies (free and paid)

4.2. (SO2) Evaluate Ethical Concerns and Privacy Compliance Using Explicit Indicators

The Ethical Compliance Index (ECI) aggregates four binary safeguards (identity linkage, anonymous/temporary mode, deletion controls, consent). ChatGPT and Perplexity achieve 4/4, whereas Gemini, Claude, Copilot, and Deepseek score 2/4 (Figure 3). This indicates a trade-off: top technical performers are not automatically top privacy performers. The decoupling is reflected in cross-domain correlations (Ethical vs. Technical r = .35), highlighting the need for distinct governance attention.

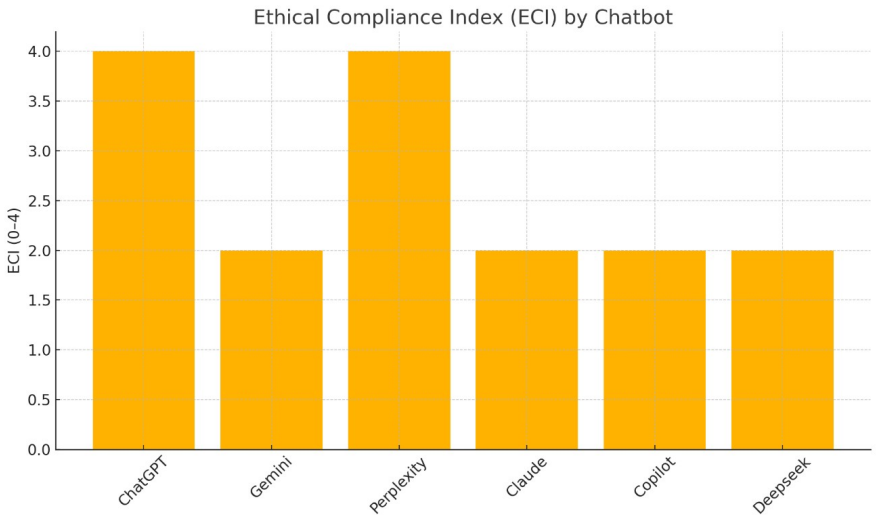

Figure 3. Comprehensive comparison between the privacy features of different AI chatbots

Overall, these results underscore the nuanced trade-offs faced by institutions: while Gemini and ChatGPT deliver leading edge features and pedagogical value, Perplexity and ChatGPT stand out in privacy compliance. The detailed quantitative and qualitative insights provided here will guide university stakeholders in selecting and configuring chatbot technologies to align with educational objectives and ethical mandates.

4.3. (SO3) Compare Chatbot Performance Using Standard Technical Metrics and Expert Rubrics

Quantitative indicators reveal significant between-system differences (Figures 4-5). Perplexity (PPL) ranges 9 (Copilot) to 15 (Deepseek), implying more predictable text for Copilot. Response latency spans 1.0 s (Perplexity) to 1.5 s (Gemini) with $F(5, 174) = 6.12$, $p < .001$. Context window shows the widest dispersion (4,096 to 2,000,000 tokens; Gemini at the upper bound). SWE-bench accuracy ranges 0.70 (Deepseek) to 0.85 (Claude). These differences substantiate distinct technical profiles relevant to course, coding, and long-document use cases.

4.3.1. Qualitative Expert Review (Tools, Criteria, Procedure, Outcomes)

Tools used by experts. Reviewers worked from standardized scoring sheets tied to two rubrics — Content Quality Index (CQI) and Ethical Risk Scale (ERS)— and blind de-identified transcripts of 300 chatbot-student interactions sampled from the 50-prompt battery (uniform across platforms).

Evaluation criteria

- **CQI**: pedagogical relevance, factual accuracy, structure/clarity, actionability.
- **ERS**: bias/discrimination cues, privacy transparency (disclosures/opacity), policy-aligned redirections/refusals.

Procedure. Experts completed a calibration round, then blind-rated interactions; disagreements were adjudicated after independent scoring. Inter-rater reliability was Cohen's $\varkappa = 0.78$, indicating substantial agreement.

Outcomes. ChatGPT attained the highest pedagogical clarity ($M = 4.6/5$). Claude excelled at long-context synthesis but showed occasional over-caution in examples. Gemini produced strong analogies in ethics cases yet sometimes speculative reasoning. Perplexity and Copilot balanced speed with moderate depth; Deepseek trailed on contextual relevance ($M = 3.1/5$). These results align with the mixed-domain ranking in Figure 2 and the technical contrasts in §4.3.

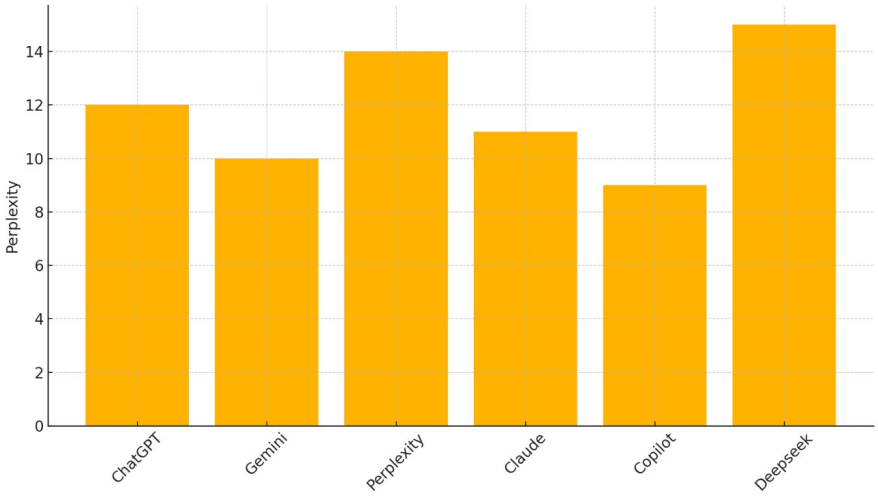

Figure 4. Perplexity by Chatbot

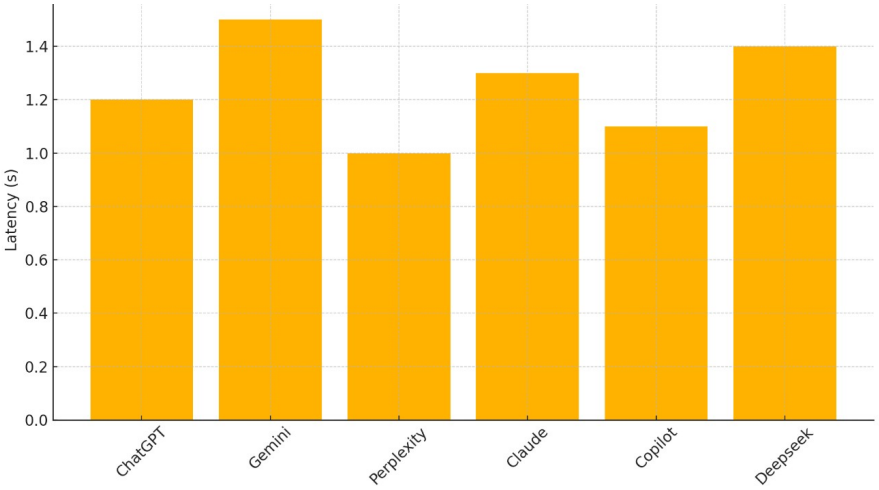

Figure 5. Response Latency by Chatbot

4.3.2. Student Satisfaction and Perceived Usefulness

Survey data from 300 students (Figure 6) show mean satisfaction scores ranging from 3.5 (Deepseek, SD = 0.8) to 4.2 (Claude, SD = 0.4). A Friedman test confirmed significant differences in perceived usability across chatbots ($\chi^2(5) = 32.7$, $p < .001$). Post-hoc analysis revealed that Claude and ChatGPT outperformed Deepseek ($p < .01$), highlighting the importance of balanced functionality and ethical transparency for user acceptance.

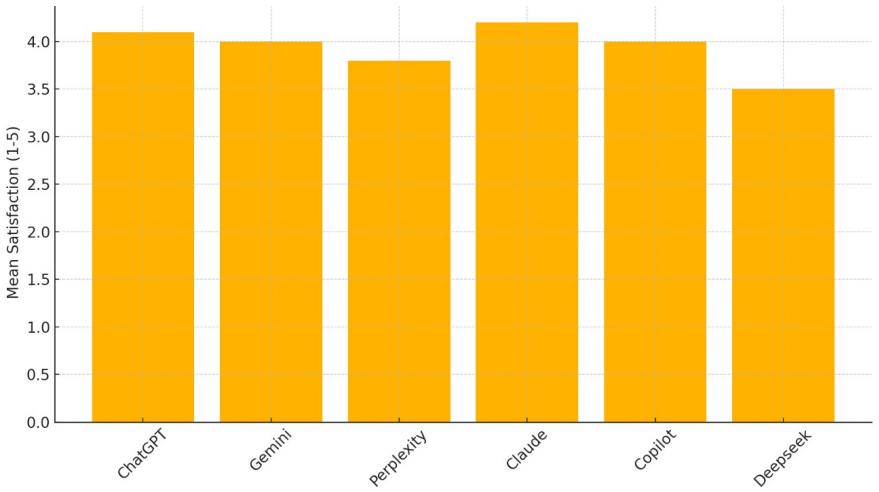

Figure 6. Student Satisfaction by Chatbot

4.3.3. Use-Case-Specific Performance

Task-level tests show differentiated strengths (Figure 7): APA citation accuracy—Claude 82 %, ChatGPT 78 %, Copilot 70 %. Ethical-dilemma resolution—Claude 90 %, ChatGPT 85 %, Gemini 75 %. Multi-step research reasoning—Copilot 4.1 correct steps on average vs. Deepseek 2.6. These findings are consistent with expert narratives and inform tool-task alignment.

4.3.4. Cross-Domain Associations

Correlations indicate strong positive links between Technical and Psychological ($r = .94$) and between Pedagogical and Psychological ($r = .90$), while Ethical correlates only moderately with Technical ($r = .35$). This supports the view that ethical/privacy safeguards require dedicated design and governance, not merely technical upgrades. (Figure 8).

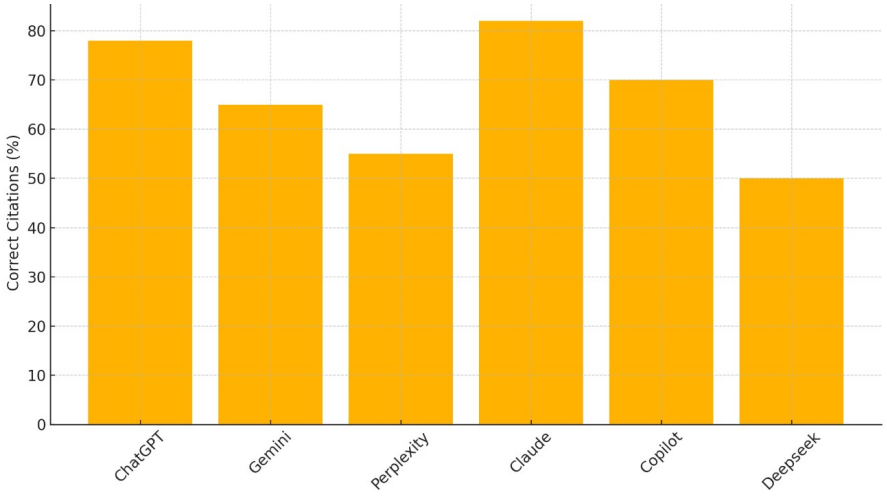

Figure 7. APA Citation Accuracy by Chatbot

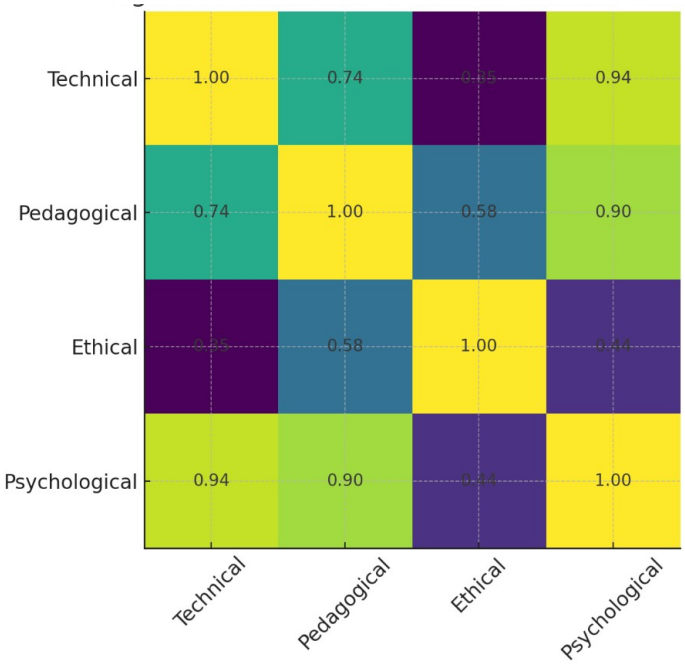

Figure 8. Correlation Matrix of Domain Scores

4.3.5. Synthesis of Key Findings

Leading capabilities with caveats. Gemini tops overall via context-length and engagement but lags in ECI; ChatGPT and Claude offer the most consistent balance across domains—particularly for pedagogy and ethics/privacy.

Privacy matters for acceptance. Satisfaction correlates more with clarity + transparent data policies than with raw technical scores —suggesting governance plus communication are pivotal for campus adoption.

Match tool to task. Claude for citations/ethics, Copilot for responsive STEM workflows, Gemini for very long contexts, ChatGPT for clear explanations with strong privacy posture.

Finally, task-specific strengths underscore the importance of aligning tool selection with intended use cases. Claude leads in APA citation accuracy (82 %) and ethical dilemma resolution (90 %), whereas Copilot excels in responsiveness (latency = 1.0 s) and code-solving benchmarks (SWE-bench = 0.82). These nuanced insights enable stakeholders to match chatbot capabilities to pedagogical objectives and

institutional ethics, ensuring that technology adoption advances both educational quality and responsible innovation.

4.4. Limitations and Delimitations

First, our "student perceptions" analysis is scoped down in this manuscript to maintain coherence as per reviewer guidance; a fuller treatment appears in a companion article. Second, some technical indicators (e.g., PPL via reference scoring) are proxies constrained by UI/API access and may under- or over-estimate native model behaviour; however, cross-checks and robustness analyses mitigate this risk. Third, vendor ecosystems evolve quickly; our audit captures a time-bounded snapshot, so institutions should treat our ECI as a repeatable procedure rather than a permanently fixed ranking. Finally, our European, university-based sample limits generalizability to other regions or educational levels; future work should include non-university and non-European contexts.

5. Discussions

This discussion interprets our findings against the study specific (SO1) Analyze student-facing benefits and limitations of generative AI chatbots; (SO2) Evaluate ethical concerns and privacy compliance using explicit indicators; (SO3) Compare chatbot performance using standard technical metrics and expert rubrics; and (SO4) Derive institutional recommendations for responsible adoption in higher education. The results show a decoupling between raw technical capability and ethical/privacy safeguards, with implications for institutional selection, governance, and classroom practice.

In reference to the first Specific Objective (SO1) expert ratings indicate that clarity, structure, and actionable guidance are the strongest predictors of perceived usefulness for students, with ChatGPT achieving the highest pedagogical clarity (M = 4.6/5) and Claude excelling on long-context synthesis. These patterns align with the literature that emphasizes aligning algorithmic outputs with explicit pedagogical aims and scaffolding (e.g., structured explanations, transparent reasoning steps). Practically, institutions should privilege models that consistently produce teachable, well-scaffolded responses over those that merely maximize length or novelty. This supports a "pedagogy-first" procurement lens rather than a purely technical one.

Yet, our moderate correlation (r = .35) between Technical prowess and Ethical compliance underscores a theoretical gap: performance optimization alone does not guarantee responsible design. This decoupling reveals that innovations in model architecture must be paired with privacy-by-design principles and explicit consent mechanisms to fulfil regulatory and ethical requirements (Williamson, Bayne & Shay, 2020; Gumusel et al., 2024).

Moreover, our use-case-specific findings—such as Claude's 82 % citation accuracy and 90 % success rate in ethical dilemma scenarios—validate constitutional AI theory (Anthropic PBC, 2025), which predicts that safety-focused training paradigms yield models better aligned with bias mitigation and factual integrity. By integrating quantitative metrics with qualitative expert judgments, we propose an extended theoretical framework that equally values technical fluency, pedagogical efficacy, ethical integrity, and psychological engagement. This multi-domain model lays the groundwork for adaptive weighting strategies—empowered by methods like the Analytic Hierarchy Process—enabling institutions to recalibrate priorities in line with evolving educational goals.

With respect to the second Specific Objective (SO2) the Ethical Compliance Index (ECI) reveals that top technical performance does not imply strong privacy posture (e.g., Gemini's high technical scores versus moderate ECI), and correlations confirm only a moderate association between Ethical and Technical domains (r = .35). This echoes broader debates about datafication, authorship, and privacy under GDPR/AI-Act logics: educational deployments must evaluate identity linkage, log deletion, anonymous modes, and consent pathways explicitly and independently from accuracy or speed. In practice, this means policy teams should assess privacy features feature-by-feature and not assume that a "better model" is a "safer model."

The integration of AI chatbots into academic settings has introduced significant concerns regarding academic integrity. Plagiarism issues have intensified, as students may be tempted to submit AI-generated work as their own. Existing plagiarism detection tools —such as Turnitin, Plagscan, or Unicheck— struggle to reliably differentiate between human-produced and AI-generated content. This ambiguity raises pressing questions about responsibility: Is the student, the AI, or the original source ultimately accountable for the work submitted?

The practical impact of these issues is compounded by historical precedents. Universities have faced similar challenges with earlier technologies, such as the debates over the credibility of Wikipedia (Knight & Pryke, 2012) and the reassessment of its role in academic research (Messner & DiStaso, 2013). The current controversy, illustrated by instances where ChatGPT has been listed as an author (Stokel-Walker, 2023) and subsequently banned by some journals (Nature, 2023), underscores the urgency of reexamining academic policies. Expanding the definition of plagiarism to include AI-generated content has been proposed (Barnett, 2023), alongside recommendations to reinterpret university honor codes so that AI is recognized as a collaborative tool rather than a shortcut (Ryan, 2023).

Privacy concerns add another layer of complexity. The extensive data required for AI chatbots to function properly may compromise student data privacy, posing risks in complying with regulations such as the GDPR —particularly regarding the "right to be forgotten" and data transparency. The heterogeneous privacy approaches observed in our study —such as the anonymous query options in ChatGPT and Perplexity AI versus the mandatory registration required by other platforms— underscore a critical trade-off between advanced functionality and robust data protection. This diversity reinforces the need for higher education institutions to adopt policies that prioritize student-friendly privacy measures while balancing academic integrity. Notably, Claude AI distinguishes itself by positioning as an ethical AI, emphasizing safety during model training and reducing the risk of generating harmful content, which illustrates how ethical design can play a pivotal role in mitigating some of these privacy and integrity concerns faced by students. Furthermore, Williams (Williams, 2024) highlights that the use of generative AI in education challenges data privacy, while Fu (Fu, 2023) points out that the collection and use of personal data by systems like ChatGPT remains a major concern. Moreover, the lack of clarity about what data is collected, how it is stored, and with whom it is shared further exacerbates these risks (Gumusel et al., 2024; Tlili & et al., 2023).

The use of LLMs involves several issues that go beyond privacy, intertwined with usability and control issues, and affect most users (Zhang, Jia, Lee, Yao, Das, Lerner et al., 2024): (a) Users are forced to choose between sharing data to improve results and protecting their privacy; (b) Many confuse the functioning of LLMs, believing that they operate like search engines, and are unaware of their statistical processes; (c) There is a risk that these models memorize sensitive information and filter it in future responses; (d) Interfaces adopt dark patterns that limit users' control over their data; (e) Finally, some users adopt self-protection practices, such as censoring or modifying information, to safeguard their privacy. Additionally, overly stringent risk avoidance algorithms can restrict natural dialogue, potentially prompting inadvertent disclosures when users become frustrated (Gumusel et al., 2024). Furthermore, heightened awareness of these risks may increase user anxiety and alter interaction behaviors, possibly discouraging open academic engagement (Gumusel et al., 2024).

In the specific case of university students, the risks manifest themselves in a particular way, as privacy risks integrate ethical and institutional dimensions specific to the academic environment (Zhang et al., 2024); (a) Students may inadvertently share sensitive academic and personal data with chatbots —for example, when seeking help with assignments, CV reviews, or even emotional support— thereby exposing themselves to unforeseen vulnerabilities; (b) Furthermore, excessive trust and anthropomorphism can lead students to perceive chatbots as friends or therapists, encouraging the disclosure of personal information that may compromise their security. This fragile trust, if misplaced, can trigger a cycle of self-disclosure, intensifying privacy risks (Gumusel et al., 2024); (c) Misunderstandings about data management might cause students to overlook that the information they provide could be used to train future models or even

be exposed to other users through data retention processes; (d) The impact of non-transparent opt-out policies may prevent students from effectively controlling the use of their data, as cumbersome procedural barriers often discourage them from opting out; (e) Finally, strict confidentiality standards and institutional norms within academic settings can amplify privacy concerns when students inadvertently violate established protocols regarding research data, academic work, and student-teacher interactions.

Beyond academic integrity, the ethical implications extend to the broader learning experience. The use of AI chatbots can affect student self-efficacy; while these tools offer personalized support, they also risk fostering an overdependence that may diminish critical thinking and initiative. As noted by Williams (Williams, 2024), the excessive reliance on chatbots may reduce students' engagement and self-directed learning, ultimately impacting academic performance. Fu (Fu, 2023) also raises concerns that the ease of generating content through AI can lead to significant issues in originality and authenticity.

Additionally, issues of bias and misinformation in AI-generated content have tangible implications for students. Algorithmic bias can lead to the reinforcement of existing stereotypes, and the phenomenon of 'AI hallucinations' —where chatbots produce inaccurate or misleading information— can compromise the quality of academic work. Williams (2024) and Fu (2023) note that such inaccuracies may not only distort learning but also undermine the trust placed in digital educational tools. Similarly, misleading data generation can manipulate or confuse users, potentially leading to unintended disclosures based on erroneous outputs (Gumusel et al., 2024).

In summary, the practical implications for students are multifaceted: challenges in maintaining academic integrity, risks associated with overreliance on AI, threats to data privacy, and potential biases in information. Addressing these issues requires a concerted effort from educators, administrators, and policymakers to implement robust guidelines and support systems that safeguard student interests while harnessing the benefits of AI technologies.

The third specific objective (SO3) shows that technical indicators differentiate platforms on fluency (PPL), responsiveness (latency), long-context handling (CW), and STEM problem-solving (SWE-bench). Yet their downstream educational impact is mediated by pedagogy and governance: for example, a very large context window is useful only when paired with lucid summarization and citation practices, and high coding accuracy must be accompanied by integrity safeguards in assessment. Our triangulation (metrics + expert review) shows that balanced profiles (e.g., ChatGPT, Claude) tend to translate into higher expert utility judgments, even when a competitor leads narrowly on a single technical metric.

Building on our findings, we propose a set of actionable recommendations to guide universities in adopting AI chatbots responsibly and effectively. First, stakeholders should establish clear governance frameworks that articulate acceptable use cases, data-management policies, and evaluation cycles. Such frameworks must mandate regular privacy audits —leveraging the Ethical Compliance Index metrics— to verify that chatbots continue to meet GDPR requirements and respect students' right to data deletion.

Second, curriculum designers and instructional technologists should integrate chatbot-based activities into learning modules only after aligning them with pedagogical objectives. For example, before deploying a tool like Claude for research-oriented assignments, faculty must ensure that students receive training on prompt engineering and citation verification to mitigate hallucinations and plagiarism risks. These preparatory workshops can also foster digital literacy, helping students understand how LLMs process and store personal information.

Third, IT departments should configure AI integrations to balance functionality and privacy. Where possible, institutions can partner with vendors to enable anonymized or pseudonymized modes of interaction, replicating ChatGPT's anonymous query feature while retaining access to advanced capabilities. Additionally, deploying chatbots within secure, university-managed environments —rather than via public web portals— can give students more control over data retention and reduce exposure to external data-sharing policies.

Fourth, universities ought to establish multidisciplinary oversight committees —comprising faculty, ethicists, student representatives, and IT specialists— to review emerging chatbot technologies. These committees can apply our weighted domain framework to score new platforms, ensuring that selections reflect institutional priorities, whether they emphasize teaching innovation, research support, or data stewardship.

Finally, continuous evaluation is essential: institutions should collect both quantitative usage metrics (e.g., session counts, error rates) and qualitative feedback (e.g., student focus groups, faculty surveys) each semester. By monitoring trends in student satisfaction, learning outcomes, and privacy incidents, universities can iteratively refine chatbot configurations and pedagogical guidelines, thus sustaining a cycle of evidence-based improvement in AI-enhanced education.

The fourth Specific Objective (SO4) highlights that our integrated findings speak directly to three tensions widely reported in the field —personalization vs. privacy; efficiency vs. equity; innovation vs. integrity— by showing how a multi-domain evaluation can surface policy-relevant trade-offs that single-metric comparisons miss. The observed authorship/plagiarism issues, bias risks, and privacy gaps align with prior critiques and recent institutional responses (e.g., honour-code reinterpretations, detection limits, AI-literacy initiatives). By operationalizing ethical safeguards into the ECI and pairing them with expert rubrics, we extend earlier conceptual calls for governance frameworks with a practical, auditable scoring approach tailored to higher education. Attending to those findings the practical implications for universities are: Procurement & onboarding. Evaluate candidates on four parallel fronts —Technical, Pedagogical, Ethical/Privacy, Psychological engagement— using checklists aligned with our indicators (e.g., identity linkage, deletion controls) rather than relying on vendor summaries; Course design. Where long-context use is critical (syllabi, thesis chapters), pair such features with structured-summary prompts, citation verification habits, and clear authorship disclosure norms; Assessment integrity. Combine task design (process evidence, oral defense) with explicit AI-use policies that distinguish acceptable assistance from authorship substitution; Privacy governance. Default to platforms (or configurations) that support anonymous/temporary modes, transparent retention windows, and opt-out/opt-in controls at point of use; prefer campus-managed deployments when feasible; and Equity and literacy. Invest in AI-literacy and access supports so benefits do not concentrate among already advantaged students.

6. Conclusions

The conclusions of this study highlight the complexity of integrating generative chatbots into higher education, showing that their usefulness and relevance do not rely solely on technical quality but rather on a balanced consideration of pedagogical, ethical, privacy, and institutional dimensions. The findings illustrate both the transformative potential of these tools in enhancing learning and academic guidance, and the risks that arise from unregulated or uncritical use. In this regard, the evidence underscores the importance of carefully assessing student-perceived benefits and limitations, the robustness of ethical and privacy safeguards, technical performance across diverse tasks, and the capacity of institutions to establish responsible frameworks with explicit, context-sensitive criteria.

- **(SO1)** Analyze student-facing benefits and limitations of generative AI chatbots : Across platforms, usefulness for students hinged on clarity, structure, and actionable guidance, with ChatGPT leading in pedagogical clarity and Claude excelling at long-context synthesis. At the same time, speculative reasoning and over-cautiousness appeared in specific cases, underscoring the need for scaffolded use and explicit authorship/disclosure norms.
- **(SO2)** Evaluate ethical concerns and privacy compliance using explicit indicators**:** The Ethical Compliance Index (ECI) showed that strong technical performance does not guarantee robust privacy safeguards. ChatGPT and Perplexity achieved the highest compliance (4/4), while others lagged on identity linkage, anonymous/temporary modes, deletion controls, or consent. Institutions must therefore evaluate privacy feature-by-feature rather than assume safety from overall model quality.

- **(SO3)** Compare chatbot performance using standard technical metrics and expert rubrics**:** Systems diverged on fluency (PPL), responsiveness (latency), long-context capacity, and STEM problem solving (SWE-bench). Yet expert judgements favored *balanced* profiles that convert technical strengths into teachable outputs, reinforcing a "match tool to task" principle for academic use.
- **(SO4)** Derive institutional recommendations for responsible adoption in higher education**:** Institutional selection should weigh four parallel dimensions —Technical, Pedagogical, Ethical/Privacy, and Psychological— using transparent checklists (e.g., identity linkage, deletion/retention windows, consent pathways). When feasible, campus-managed or anonymous/temporary configurations should be preferred, coupled with AI-literacy, assessment redesign, and clear disclosure policies.

Ultimately, the study demonstrates that the responsible adoption of generative chatbots in higher education cannot be reduced to questions of efficiency or performance, but must instead be framed as a multidimensional challenge that requires sustained dialogue among educators, students, policymakers, and developers. By situating these tools within transparent and context-sensitive frameworks, universities can move beyond a reactive stance toward a proactive integration that safeguards ethical principles, nurtures trust, and maximizes the pedagogical value of AI for academic communities.

Building on these limits, we identify three priorities for future research: (i) longitudinal studies that track how pedagogy, privacy settings, and student outcomes co-evolve; (ii) task-specific benchmarks that couple technical metrics with pedagogical success criteria (e.g., rubric-aligned reasoning steps); and (iii) policy experiments (e.g., default-anonymous vs. account-linked deployments) to estimate causal effects on usage, learning, and integrity outcomes. These strands will help translate multi-domain evaluations into actionable institutional playbooks.

Declaration of Conflicting Interests

The authors declare no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author received no financial support for the research, authorship, and/or publication of this article.

References

Anthropic PBC (2025). *Claude AI*. Available at: https://www.claude.ai

Barnett, S. (2023). *ChatGPT Is Making Universities Rethink Plagiarism.* Wired Magazine. Available at: https://www.wired.com/story/chatgpt-college-university-plagiarism/

Bennett, L., & Abusalem, A. (2024). AI and its potential impact on the future of higher education. *Athens Journal of Education,* 11(3), 195-212. https://doi.org/10.30958/aje.11-3-2

Boratkar, J., & Sambhe, R.U. (2024). Artificial intelligence in higher education: A literature snapshot. *International Journal of Scientific Research in Science, Engineering and Technology,* 11(4), 228-232. https://doi.org/10.32628/ijsrset24114124

Chadha, A. (2024). Transforming higher education for the digital age. *Journal of Interdisciplinary Studies in Education,* 13(S1). https://doi.org/10.32674/em2qsn46

Davydova, G.I., & Shlykova, N.V. (2024). Risks and challenges in introducing artificial intelligence into higher education. *Bulletin of Practical Psychology of Education,* 21(3), 62-69. https://doi.org/10.17759/bppe.2024210308

Deepseek AI (2025). Available at: https://Deepseek.com

Díaz-Arce, D. (2024). Herramientas para detectar el plagio a la inteligencia artificial: ¿cuán útiles son? *Revista Cognosis,* 9(2), 144-150 . https://doi.org/10.33936/cognosis.v9i2.6195

Fu, Y. (2023). A research of the impact of ChatGPT on education. *Applied and Computational Engineering,* 35, 26-31. https://doi.org/10.54254/2755-2721/35/20230354

Gumusel, E., Zhou, K.Z., & Sanfilippo, M. (2024). User Privacy Harms and Risks in Conversational AI: A Proposed Framework. *Computer Science,* arXiv:2402.09716. https://doi.org/10.48550/arXiv.2402.09716

Harari, Y.N. (2016). *Homo Deus: A Brief History of Tomorrow.* Harvill Secker. https://doi.org/10.17104/9783406704024

Knight, C., & Pryke, S. (2012). Wikipedia and the University, a case study. *Teaching in Higher Education,* 17(6), 649-659. https://doi.org/10.1080/13562517.2012.666734

Koenecke, A. (2020). Racial disparities in automated speech recognition. *Proceedings Of The National Academy Of Sciences*, 117(14). https://doi.org/10.1073/pnas.1915768117

Messner, M., & DiStaso, M.W. (2013). Wikipedia versus Encyclopedia Britannica: A Longitudinal Analysis to Identify the Impact of Social Media on the Standards of Knowledge. *Mass Communication and Society,* 16(4), 465-486. https://doi.org/10.1080/15205436.2012.732649

Meta AI (2023). *Introducing Llama 2: The Next Generation of Our Open Source Large Language Model.*

Micheni, E., Machii, J., & Murumba, J. (2024). The role of artificial intelligence in education. *Open Journal for Information Technology,* 7(1), 43-54. https://doi.org/10.32591/coas.ojit.0701.04043m

Nature (2023). *Tools such as ChatGPT threaten transparent science; here are our ground rules for their use.* Nature. https://www.nature.com/articles/d41586-023-00191-1

Nayak, S., Pasumarthi, S., Rajagopal, B., & Verma, A.K. (2024). GDPR Compliant ChatGPT Playground. *International Conference on Emerging Technologies.* Bengaluru, India. https://doi.org/10.1109/icetcs61022.2024.10543557

Noble, S.U. (2018). *Algorithms of oppression: How search engines reinforce racism.* NYU Press. https://doi.org/10.2307/j.ctt1pwt9w5

O'Donnell, F., Porter, M., & Fitzgerald, S. (2024). The role of artificial intelligence in higher education. *Irish Journal of Technology Enhanced Learning,* 8(1). https://doi.org/10.22554/szwjfy54

Open AI (n.d.). *Open AI Deprecations*. Available at: https://platform.openai.com/docs/deprecations (Accessed: February 2025).

Rudolph, J., Ismail, F.M., & Popenici, S. (2024). Higher education's generative artificial intelligence paradox: The meaning of chatbot mania. *Journal of University Teaching and Learning Practice,* 21(6). https://doi.org/10.53761/54fs5e77

Ryan, P. (2023). Ethical considerations in the transformative role of AI chatbots in education. *SSRN.* https://doi.org/10.2139/ssrn.4623611

Saaty, T. L. (1980). *Analytic Hierarchy Process.* McGraw-Hill.

Sebastian, G. (2023). Privacy and Data Protection in ChatGPT and Other AI Chatbots: Strategies for Securing User Information. *SSRN.* https://doi.org/10.2139/ssrn.4454761

Stokel-Walker, C. (2023). *ChatGPT listed as author on research papers: many scientists disapprove*. Nature. https://doi.org/10.1038/d41586-023-00107-z

Tlili, A., Shehata, B., Agyemang, M., Bozkurt, A., Hickey, D.T., Huang, R. et al. (2023). What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education. *Smart Learning Environments,* 10(15). https://doi.org/10.1186/s40561-023-00237-x

Universidad Complutense de Madrid (2024). *Guía para citar inteligencia artificial en APA*. Madrid: UCM. Available at: https://biblioguias.ucm.es/estilo-apa-septima/citar_inteligencia_artificial

Volpicelli, G. (2023). *ChatGPT broke the EU plan to regulate AI*. Politico. Available at: https://www.politico.eu/article/eu-plan-regulate-chatgpt-openai-artificial-intelligence-act/

Wang, Y. (2024). The Impact of the Artificial Intelligence Act on ChatGPT. *Lecture Notes in Education Psychology and Public Media,* 44(1), 85-192. https://doi.org/10.54254/2753-7048/44/20230134

Williams, R.T. (2024). The ethical implications of using generative chatbots in higher education. *Frontiers in Education,* 8. https://doi.org/10.3389/feduc.2023.1331607

Williamson, B., Bayne, S., & Shay, S. (2020). The datafication of teaching in higher education: Critical issues and perspectives. Teaching in Higher Education, 25(4), 351-365. https://doi.org/10.1080/13562517.2020.1748811

Yunusov, K., Berdiyev, B., & Jovliev, B. (2024). A well-structured system for learning and also evaluating the outcome for better learning using artificial intelligence in higher pedagogy. *4th International Conference on Advance Computing and Innovative Technologies in Engineering.* Greater Noida, India. https://doi.org/10.1109/ICACITE60783.2024.10617101

Zhan, X., Seymour, W., & Such, J. (2024). Beyond Individual Concerns: Multi-user Privacy in Large Language Models. *6th ACM Conference on Conversational User Interfaces.* Luxembourg. https://doi.org/10.1145/3640794.3665883

Zhang, Z., Jia, M., Lee, H.P., Yao, B., Das, S., Lerner, A. et al. (2024). It's a Fair Game", or Is It? Examining How Users Navigate Disclosure Risks and Benefits When Using LLM-Based Conversational Agents. *CHI Conference on Human Factors in Computing Systems.* Honolulu, USA. https://doi.org/10.1145/3613904.3642385

Published by OmniaScience (www.omniascience.com)

Journal of Technology and Science Education, 2025 (www.jotse.org)