INTERNATIONAL JOURNAL OF SOCIAL SCIENCES BULLETIN
ISSN: (E) 3007-1917 (P) 3007-1909
Volume 4, Issue 2, 2026

EXPLORING THE ROLE OF CHATGPT IN HIGHER EDUCATION: IMPACT, BENEFITS, AND CHALLENGES

Shumaila Waris[*1], Dr. Maroof Bin Rauf[2]

[*1]*M. Phil Scholar, Department of Education, University of Karachi, Karachi, Pakistan*
[2]*Assistant Professor, Department of Education, University of Karachi, Karachi, Pakistan*

[*1]shamil.waqas3018@gmail.com, [2]maroof.rauf@uok.edu.pk

Corresponding Author: *
Shumaila Waris
DOI: https://doi.org/10.5281/zenodo.18811052

Received	Accepted	Published
29 December 2025	13 February 2026	28 February 2026

ABSTRACT

The study aims to explore the role of using the generative AI model ChatGPT (chat generative pre-trained transformer) in higher education. This research explores the perception of university students about ChatGPT by focusing on the impact, benefits, and challenges of higher education.

The goal of this research is to explore the impact of ChatGPT on the skills of higher education students, as well as how chatbots can augment human knowledge and judgment in higher education.

This is a mix method study Qualitative and Quantitative method in which 60 students were selected from 5 public and private universities in Karachi. Data has been collected through interviews and open-ended questionnaires. 12 students were selected from each university, consisting of men and women from humanities and sciences backgrounds. Data were analyzed through thematic analysis and Chi square. Data were classified into three categories to show varying levels of response from positive to negative.

The result showed a mixed trend in which the majority of students thought that ChatGPT has a positive impact and provides various opportunities for higher education, including assessment innovation, teaching support, remote learning support, research design & development support and academic writing etc. Support and productivity for academic challenges, while the second group opined that it is beneficial but challenges and issues related to academic integrity, appropriate guidelines, security and privacy, reliance on artificial intelligence, learning assessment, and information accuracy. The number is huge.

This study presents a set of recommendations for the effective use of ChatGPT in higher education. It concludes that the application of ChatGPT in higher education presents both benefits and challenges. Thus, efforts and strategies are needed to ensure the effective use of ChatGPT for educational purposes. By balancing potential benefits and challenges, ChatGPT can enhance students' learning experiences in higher education.

Keywords: *ChatGPT, Artificial intelligence (AI), Higher education, Impact, Benefits, Challenges, Large Language Model (LLM).*

INTRODUCTION

Language models Generative artificial intelligence (GenAI) systems and in particular chatbots like ChatGPT have infiltrated the higher education field with shocking velocity. Less than a few months before the official publication, even before such tools were widely publicised, students and faculty started to brainstorm, outline, summarise, translate, debug code and generate feedback, and institutions were scrambling to revise policies and pedagogy. Like previous instructional technologies, the diffusion curve was uneven across disciplines and situations, but the rate

and depth of adoption is more characteristic of higher education. Supporters envision new opportunities in scalable formative assessment, one-on-one tutoring, and instructor workload reduction; critics fear new opportunities of increased plagiarism, imagined factual illusions, latent bias, and privacy invasion that will reinforce inequalities. The policy and practise environment has become dynamic: universities are both trying specific integrations, reinforcing integrity guidance, and redesigning assessment. It is within that context that this paper has framed a narrow research question of how ChatGPT is being utilised in courses, what impact it now has on learning and study processes, what benefits and drawbacks users are seeing, and why some members of the community are seeking to adopt (or avoid) the tool (Kasneci et al., 2023; Lund and Wang, 2023; UNESCO, 2023).

An emerging empirical foundation indicates the possible education advantage of scaffolded and intentionally embedded ChatGPT. Research has found positive effects (e.g., more coherent structure, more timely revision) on the quality and process of writing when ChatGPT is used as a source of formative feedback reports on a student, but not as a text generator of record. Integrated-classroom practises in discipline-based contexts (e.g., dental education) also show perceived efficiency and knowledge gains, so long as guidance and guardrails are clear. A recent meta-analytic synthesis of higher education experimental research reports overall positive effects on academic performance and affective-motivational measures, and decreases in mental effort, although with greatly varying levels of heterogeneity and design constraints across primary studies. These initial results present tentatively positive optimism that the framing of the tool by the teacher (as tutor, coach, writing lab, and not as ghostwriter) is equally important as that of its use (Deng et al., 2024; Mahapatra, 2024; Kavadella et al., 2024).

Simultaneously, the idea of academic integrity has turned into a primary point of tension. Faculty express concerns that they do not know where to draw the line between work done by independent students and work done with AI assistance, and students express ambivalent standards regarding everyday assistance and unacceptable delegation. Researchers warn that blind trust in automated AI-detection systems can bring about false positive results, instability in cross-writing styles, and possible damages to non-native writers; clear evidence-based assessment design, clear course-level standards and authentic displays of learning (e.g., oral defences, multi-drafting, in-class problem solving) are therefore encouraged. That is, the issue of integrity does not only revolve around technological detection, but rather about the assessment eco-systems that harmonise tasks, supports, and learning evidence. To make responsible integration of ChatGPT, it is important to redefine what is considered authorised assistance, as well as to explain the need to do some tasks without it (Cotton et al., 2024; UNESCO, 2023).

Consideration of equity, prejudice and privacy makes adoption decisions complicated. On one hand, GenAI has the potential to reduce entry barriers by providing language support, study tactics, and 24/7 in simpler words, the latter, in particular, could be particularly beneficial in case a student does not have access to human assistance. Conversely, unequal access to high-quality models, inconsistent dependability across knowledge areas, and one-sided failure might create gaps in achievements unless there is specific help and critical-use pedagogy. International recommendations encourage institutions to offer AI literacy (both staff and students), make it accessible, necessitate human supervision, and reduce risks to privacy and data security of learners- such as by specifying what data is being shared with third parties providers and by offering compliant workflows when using AI is approved (UNESCO, 2023).

It is thus necessary to understand how stakeholders in higher-education practically use ChatGPT in courses. Surveys of campus snapshots and sector analyses indicate a continuum: students say they use ChatGPT to brainstorm, summarise readings, outline, code explain, and practise assessments; faculty say they are experimenting with all rubric-aligned feedback, question-generating, lesson-planning, and administrative-drafting. Notably, the study process can change: students can transfer low-level fluency exercises and spend more time on planning and developing arguments, yet they can also fall into the trap of illusion of competence in case they do not struggle to put

them together. To the instructors, time saved in routine drafting can facilitate increased feedback and relationship work provided there is strong norms and open communication. Evidence-based guidance is a precondition of mapping these lived practises (Lund and Wang, 2023; Kasneci et al., 2023).

The perceived advantages and obstacles vary in terms of position and field. Immediacy (on-demand explanations), ideation support, and affective benefits (lessened anxiety when challenged by blank pages) are some of the aspects that are usually pointed out by students. Erosion of foundational skills, faculty were more likely to predict workload relief (writing rubrics, emails and prompts), enhanced differentiation (customised practise items), and were more worried about erosion of foundational skills. In a system where courses have experimented with course-integrated use with evident scaffolds, e.g., by asking students to annotate AI-assisted draughts, comparing outputs to sources, or reflecting on prompt-engineering decisions, respondents are likely to report higher metacognitive awareness of process (what they tell the model to do and why). Nevertheless, the perceived risks are still present: hallucinations, excessive dependency, obscure training information, and unreliable performance in specialised areas (Kavadella et al., 2024; Mahapatra, 2024).

Lastly, classic technology-acceptance variables, which include performance expectancy (Will it help me?), effort expectancy (Is it easy to use?), social influence, and facilitating conditions, and also include trust and perceived risk influence adoption intention. The UTAUT/UTAUT2 work based on university samples has consistently reported that usefulness and social norms are positive predictors of behavioural intention but risk perceptions and low trust undermine it. Sometimes these effects are moderated by experience, gender, and discipline, meaning that successful institutional strategies will be both targets (e.g. discipline-specific exemplars), and enablers (e.g. training, assessment redesign, policy clarity) (Strzelecki, 2023).

Problem statement, aim, and research questions. Although ChatGPT is being quickly and heterogeneously adopted in higher education, there is no integrated evidence on the actual course application, impacts on learning outcomes and learning process, perceived benefits and challenges by users (particularly integrity, bias, equity, and privacy), or adoption intention determinants. This paper thus seeks to (i) capture what students and faculty are currently doing with ChatGPT in their coursework; (ii) what the effects of ChatGPT are on course outcomes and the course process; (iii) what perceived advantages and disadvantages are and (iv) what predicts intention to adopt. Based on this, the following research questions are investigated: RQ1: What is the use of ChatGPT by students and faculty in courses? RQ2: How does the utilisation of ChatGPT affect learning outcomes and study processes? RQ3: What are perceived benefits and challenges by the users? RQ4: How intention to adopt ChatGPT can be predicted? It must be acknowledged that their voluntary participation is ensured by the right to free speech and the authority to act as an optional parental figure (Deng et al., 2024; Cotton et al., 2024; Strzelecki, 2023; UNESCO, 2023).

2. Literature Review

Constructivism/scaffolding. Constructivist theories maintain that learners know by constructing knowledge through engagement and reflection; scaffolding assists the learner in negotiating between what is known to them and where they want to be competent through organising tasks, cues, and feedback. A chatbot can serve as a versatile scaffold: probing misconceptions, modelling explanations or chunking complicated tasks, as long as its prompts and affordances are designed to do so in AI-supported environments (van Nooijen et al., 2024; Sweller, 2023). ChatGPT is generative and aligns quite well with dialogic scaffolds (i. e., elaborative questioning, hinting), but also runs the risk of overscaffolding and suppressing productive struggle more broadly (i. e., schema construction) (Kasneci et al., 2023).

Cognitive Load Theory (CLT). CLT predicts learning is better when the extraneous cognitive load is kept as low as possible and the intrinsic load optimally managed; worked examples and guided steps are especially beneficial with novice learners taking part in high element-interactivity tasks. Meta-analytic and review evidence concerning tasks supported by GenAI indicates

that ChatGPT can reduce perceived mental load and assist in higher-order thinking in structured (e.g. staged prompting, exemplars, verification steps) activities, but advantages are not uniform and can be compromised by design-related issues (Deng et al., 2025; Sweller, 2023). Hallucinated or talkative outputs bloat with extraneous load, and on-demand generation of code/writings can obscure germane processing unless there are explicit demands in the prompt sequences to explain, strike, or revise (Ji et al., 2023).

Self-regulated learning (SRL). SRL models focus on metacognitive planning, monitoring and control. The research conducted early in the field of higher education demonstrates that students use ChatGPT to write their plans, create checklists, and ask to be critiqued formatively; not all traits (such as conscientiousness) and prompt designs that promote reflection over answers benefit SRL (Weng et al., 2024; Kasneci et al., 2023). The implication here is to incorporate metacognitive prompts (e.g., "describe your decision," compare options, rate your confidence) and force students to make strategy use external.

Adoption of technology (TAM/UTAUT). In university adoption studies, adoption intention to use ChatGPT was consistently influenced by perceived usefulness, trust, and perceived risk; some studies have found that social influence was not as strong as predicted, and moral obligation and academic-integrity doubts may weaken assessment-use intention (Habibi et al., 2024; Lai et al., 2024). This implies that institutional messaging, acceptable-use policies, and reliable workflows (e.g., a mandatory disclosure chain and process logs) are needed as complements to the availability of tools.

Evaluation and academic-integrity systems. Reviews of policy and assessment in high-ranking universities are rapidly shifting to assessment redesign and prioritising authentic and process-rich assignments and explicit reporting on AI use instead of an outright ban (Moorhouse et al., 2023; Xia et al., 2024; Lee et al., 2024). Frameworks emphasise transparency, reference to AI aid, and correspondence with results that must involve human judgement (e.g., oral defences, studio critiques, in-class problem solving). In both experimental and quasi-experimental literature, well-scaffolded ChatGPT utilisation is associated with positive academic, learner affect, and higher-order cognition effects (with medium to large effects), and has a modest negative effect on mental effort; there are inconsistent effects on self-efficacy (Deng et al., 2025). These advantages are most pronounced where ChatGPT can be placed as a feedback generator or worked-example explainer as opposed to a direct answer engine.

In the case of writing, a large multi-rater study of teacher vs. ChatGPT feedback on student essays identified that AI feedback is often comparable to teacher feedback on criteria-based guidance, actionability and tone, although the teacher still outperforms AI on subtle prioritisation and contextual accuracy; the authors suggest hybrid feedback cycles where AI initial comments are reviewed by teachers and colleagues (Steiss et al., 2024). Likewise, case-based research on argumentation also reports that the critique offered by ChatGPT can be valid on form and invalid on content- again referring to teacher-designed prompts and human verification (Wang et al., 2024; Ji et al., 2023). ChatGPT is used in computing/problem-solving to break down problems, write initial code and reveal a variety of solution paths. Research has shown learning benefits when prompts make a student explain before write a programme (e.g., explain the algorithm and invariants you will use, and propose code), and when students analyse how the ChatGPT is wrong (Weng et al., 2024; Baig and Yadegaridehkordi, 2024).

Pupils consistently report a time-saving in writing and planning. The meta-cognitive insight indicates that a lower level of mental work and higher user interaction occur when activities combine short, repetitive cycles based on AI support rather than once-end generation (Deng et al., 2025). Nevertheless, uncontrolled utilisation may push learners into months of answer consumption, losing track of metacognition; the integration of reflective checklists and the need to rectify or justify AI performances can help curb this drift (Kasneci et al., 2023). Personalization. ChatGPT is able to customise explanations, examples, and pacing, particularly useful with cohorts of different backgrounds and multilingual classrooms. The reviews in institutions also note that it can be inclusive redesigned-e.g., multiple

exemplars at different levels, or multilingual summaries, should it be coupled with equity protections (Lee et al., 2024; Xia et al., 2024). Rapid formative feedback. Students are given prompt feedback on draughts and code that is actionable, and can repeat more revision cycles within set calendars; instructors can prompt students to rubrics or common error patterns at once (Steiss et al., 2024; Baig and Yadegaridehkordi, 2024).

Accessibility. ChatGPT can also produce outlines, glossaries or other presentations (e.g. plain-language or step-by-step versions), and can serve as an oral exam or interview practise partner with careful prompt engineering. Policy studies would promote conceptualising such affordances as tools of disclosure, rather than shortcuts (Moorhouse et al., 2023; Lee et al., 2024). Independence and hallucinations. Hallucinations (confidently incorrect outputs) will always be a fundamental weakness of LLMs; surveys list risks ranging from fabricated citations to defective inferences and accidental ad hoc facts. Counter-hallucination techniques (e.g. constrain - verify - cite routines) are now taught in AI literacy as part of many programmes (Ji et al., 2023; Allen et al., 2024).

Bias and fairness. There are systematic criticisms that warn that LLMs can recreate and multiply biases in training data. In assessment policing, it has been repeatedly demonstrated across studies that AI-detectors are more likely to label non-native English writing as AI-generated such that this is seen as an equity and due-process issue; institutions are cautioned against depending on detectors as the sole evidence (Bender et al., 2021; Liang et al., 2023; Xia et al., 2024). Privacy and IP. The advice on policies recommends that the data on an individual level should be minimised, that the copywritten materials should be used cautiously, and that the default platforms should be used by the institution, and UNESCO (2023) recommends transparency, consent, and human control (UNESCO, 2023).

Inequity and access. Whereas ChatGPT can reduce inequality by providing personalised support, disparities in access, language barrier and variation in AI literacy can increase inequality; equity-based design (e.g. low-bandwidth access, multi-lingual prompts, explicit SRL supports) will be required (Lee et al., 2024; Habibi et al., 2024). Excess dependence and skill decadence. Critiques warn that frictionless generation can deter practise and heavy processing when courses are not mandating process evidence (e.g. timely logs, revision memos, oral walkthroughs). This risk can be reduced with CLT-aligned designs that require explanation and interleaved practise (Sweller, 2023; Kasneci et al., 2023). Academic integrity. Empirical evidence and reviews of policies all point to a practical point of view: redesign assessments to place value on process, originality of thought (not merely product), situational performance, and ethical disclosure; reserve high stakes judgments to triangulated evidence (Cotton et al., 2023; Moorhouse et al., 2023; Luo et al., 2024).

Qualitative and policy analyses indicate that instructors are shifting to "AI-conscious" course designs: (1) Prompt-in-the-open--students insert prompts and system messages on appendices; (2) Process grading--points on ideation trails, critique of AI outputs and quality of revision; (3) Assessment diversification more oral defences, studio critiques, in-class builds, and local data/problem contexts; (4) AI literacy modules on responsible use, citation and verification (Lee et al., 2024; Xia et al Even faculty members report the necessity of departmental norms (when to use AI, how to reference it) and of shared prompt libraries based on outcomes. TAM/UTAUT-based adoption studies also indicate that intention to use in learning and assessment is motivated by trust (in model accuracy and policy protections) and perceived usefulness, and inhibited by perceived risk and moral obligation to avoid misconduct; the effect of social influence is inconsistent across contexts (Habibi et al., 2024; Lai et al., 2024). Such results can justify institutional investments in governance (clear guidelines, disclosure templates) and assessment in learning approaches which can justify specific uses and discourage deception.

What's solid. Structured ChatGPT use is convergingly associated with positive outcomes in performance and impact, especially in a situation where it provides quick, prominently delivered formative feedback and explain (Deng et al., 2025; Steiss et al., 2024). Today, policy reviews include a more general suggestion of assessment redesign and disclosure over prohibition (Moorhouse et al., 2023; Xia et al.,

2024). Implications for methods. Future research ought to (a) employ multi-arm randomised controls that contrast AI-only, human-only, and hybrid feedback, (b) report pre-registered outcomes and power analysis, (c) adopt joint displays that juxtapose prompt/feedback traces with outcome gains, and (d) report process evidence (verification steps of students, correctness rates). Evaluations must reward exposition, criticism, and transfer-results that are less easily counterfeited by the existing models.

Learning-science ChatGPT ChatGPT can be most effectively viewed as a scaffolded amplifier of feedback and explanation. It is best applied in cases where the educators design prompts and activity format that (1) minimise extraneous load, (2) hold learners actively engaged in cognition, (3) activate and evaluate metacognition, and (4) render AI use visible. Such designs are associated with the strongest evidence of improvement in performance and shorter formative cycles, and the most obvious risks are related to hallucinations, bias (including inequities related to the detector), privacy/IP, and atrophy of skills under unstructured dependence. Meanwhile, adoption depends on credible institutional policies and evaluation practises that explain what can be used legitimately and develop AI literacy.

METHODOLOGY

3.1 Design and overall approach

This study adopts an **explanatory sequential mixed-methods** design. **Phase 1 (Quantitative)** combines a quasi-experiment with pre/post surveys to estimate the impact of **ChatGPT-supported assignments** on learning outcomes and study processes relative to sections that complete the same tasks without ChatGPT. **Phase 2 (Qualitative)** follows to explain *why and how* the observed effects occur, using focus groups, semi-structured interviews with students and instructors, and analysis of **document artifacts** (prompts, feedback traces, and rubrics). Integration occurs through **joint displays** that align quantitative effects with qualitative mechanisms and design features.

3.2 Setting and participants

The setting is a comprehensive university offering large service and major courses in Writing/Communication, Computing/Information Systems, and Social Sciences. Courses were recruited via departmental calls and instructor briefings. Participants include (a) undergraduate and graduate **students** enrolled in participating course sections and (b) **instructors** teaching those sections.

Inclusion criteria (students). Enrolled in a participating section; consent to data use from the LMS and surveys; not on institutional leave.

Exclusion criteria (students). Auditors; students who withdraw before the first graded assignment.

Inclusion criteria (instructors). Willingness to implement the study's assignment protocol and disclosure policy; consent to interviews and artifact sharing (e.g., anonymized rubrics, prompts).

Demographic, academic, and access variables (e.g., gender, prior GPA, first language, device/bandwidth access) are captured to describe the sample and serve as covariates.

3.3 Sampling and allocation

Course **sections are stratified** by discipline and level (lower/upper division) and **matched in pairs** where possible on instructor and syllabus. Within each matched pair, one section implements the **ChatGPT-integrated** condition and the other continues with **business-as-usual (control)**. When same-instructor pairing is not feasible, sections are matched on class size, student composition, and assessment style.

Because randomization is not always possible, the design is **quasi-experimental**. To reduce selection bias, two procedures are planned:

1. **Baseline equivalence checks** on prior GPA, pretest scores, and demographic covariates.
2. **Propensity score weighting** at the student level using observed covariates (discipline, GPA, language background, digital access, prior AI experience).

3.4 Target sample size and power

Power analysis assumes a **small-to-moderate effect** on primary outcomes (standardized **d = 0.30**). With $\alpha = .05$, power = .80, and **clustering**

by section, the design effect is approximated as $DE = 1 + (m - 1)\cdot ICC$, with **m ≈ 40** students/section and **ICC ≈ .05**, yielding **DE ≈ 2.95**. To detect $d = 0.30$ after design-effect inflation, the study targets **8–10 matched pairs of sections** (≈16–20 sections total, ~640–800 students). Phase-2 qualitative sampling seeks **thematic sufficiency** with **~24–36 students** across disciplines (mix of users/non-users, higher/lower gainers) and **~12–18 instructors**.

3.5 Intervention: ChatGPT-supported assignments

Intervention sections implement **two to three core assignments** (writing, coding/problem-solving, or research synthesis) designed with **transparent prompts and verification steps**. The protocol standardizes:

- **Allowed uses** (e.g., brainstorming, outlining, pseudo-code, itemized feedback on drafts, explanation of concepts) and **disallowed uses** (e.g., submitting AI text as final work without attribution; uploading personal/third-party sensitive data).
- A **constrain-verify-cite** routine: (1) constrain model scope (task, role, criteria), (2) verify outputs against sources/rubrics, (3) cite AI assistance in a disclosure box.
- **Prompt logging**: students paste prompts and system messages into an **Appendix A** (or upload prompt histories/screens), and annotate **what they kept/changed and why**.
- **Human-in-the-loop** checkpoints: peer review and instructor mini-conferences focus on critiquing AI output quality and student revisions.

Control sections complete **the same assignments** with identical rubrics but without ChatGPT. They may use conventional resources (readings, lecture notes, office hours, standard compilers/IDEs or spellcheckers). All sections receive **identical grading rubrics**.

3.6 Measures

Primary outcomes

1. **Performance rubrics.**
   - *Writing/communication*: organization, argument quality, use of evidence, clarity, mechanics (5–7 analytic dimensions, 5-point anchors).
   - *Coding/problem-solving*: correctness, algorithmic efficiency, documentation, testing/edge-cases, explanation of approach. Two independent raters score anonymized submissions; rater disagreements >1 scale point are adjudicated.
2. **Course quizzes/exams** aligned to course outcomes (content knowledge and transfer items).
3. **Process measures**: **time-on-task logs** captured via LMS and optional screen-timers; **NASA-TLX** after major tasks to index cognitive load (mental demand, effort, frustration).

Surveys (pre/post)

- **Perceived usefulness** and **perceived ease of use** (TAM constructs).
- **AI literacy** (knowledge of strengths/limits; verification strategies).
- **Trust** (perceived accuracy/reliability for course tasks).
- **Integrity concerns** (perceived risk of misuse and fairness).
- **Intention to use** (future academic use under explicit disclosure).

All scales use 5–7 point Likert anchors; internal consistency (α/ω) and factor structure are evaluated.

Covariates

- **Prior GPA, language background** (first language; self-rated proficiency), **digital access** (device reliability, bandwidth), **prior AI experience** (apps used, frequency).

Qualitative data and artifacts

- **Student usage diaries** (brief after-action reflections on goal → prompt → output → verification → revision).
- **Prompt/response transcripts** (redacted).
- **Instructor artifacts** (rubrics, prompt templates, feedback exemplars).
- **Focus groups/interviews** exploring experiences of scaffolding, perceived learning, workload, equity, and assessment fairness.

3.7 Data collection procedures

Before instruction: Instructor briefing; IRB/ethics approval; consent procedures embedded in LMS. Baseline **pretests** and **pre-survey** are administered during Week 1–2.

During instruction: For each assignment, students submit (a) the **final product**, (b) a **process appendix** (prompts, notes), and (c) a **disclosure statement** indicating any AI assistance. Raters score artifacts blind to condition. TLX is administered immediately after tasks. LMS exports provide timestamps, attempt counts, and view logs.

After instruction: **Post-survey** and **posttests** in Week 12–14. Qualitative sampling is then purposive: students are selected across quartiles of improvement and across conditions; instructors from both arms participate. All interviews are audio-recorded with consent and professionally transcribed; identifiers are removed at transcription.

3.8 Quantitative analysis plan

Analyses are preregistered and conducted at the **student level**, with **section clustering** accounted for.

1. **Data screening.** Outliers flagged via robust Mahalanobis distance; missing values handled via **multiple imputation (MICE)** when >5% and plausibly missing at random; item parcels only after confirming unidimensionality.
2. **Reliability and measurement models.** For multi-item scales, compute **α** and **ω** (target ≥ .70). Confirmatory factor analysis (where applicable) evaluates fit (e.g., CFI/TLI ≥ .90, RMSEA ≤ .08, SRMR ≤ .08). If pre/post comparisons of latent constructs are primary, test **configural/metric/scalar invariance**.
3. **Primary impact models (RQ2).**
   - **ANCOVA** on post-outcomes (rubric scores, quiz/exam totals) with **pretest** as covariate; covariates include GPA, language background, digital access.
   - **Difference-in-differences (DiD)** at the student level for outcomes measured multiple times (baseline vs. post), with **section fixed effects**.
   - **Cluster-robust (HC3/CR2) standard errors** at the section level; **effect sizes** reported as **partial η^2** (ANCOVA) and **Cohen's d** (standardized contrasts).
4. **Process outcomes.** Regress **time-on-task** and **TLX subscales** on condition with covariates; exploratory mediation tests whether load/time partially explain performance differences.
5. **Adoption predictors (RQ4).** Hierarchical multiple regression: Step 1 covariates → Step 2 **perceived usefulness/ease** → Step 3 **integrity concerns & trust** → Step 4 **AI literacy** → Step 5 interactions **Usefulness × Integrity** and **Usefulness × Literacy**. **Multicollinearity** is checked (VIF < 5). **Simple slopes** are probed at ±1 SD for significant interactions.
6. **Sensitivity analyses.**
   - **Propensity score–weighted** models to adjust for baseline imbalances.
   - **Multilevel models** (random intercepts for section) as robustness to clustering assumptions.
   - **Multiple-comparison control** using Benjamini–Hochberg FDR for families of related tests.
   - **Per-protocol vs. intention-to-treat** contrasts, where per-protocol excludes students who violated disclosure rules.

3.9 Qualitative analysis plan

Qualitative interpretation follows **reflexive thematic analysis**. The team conducts:

1. **Familiarization** with transcripts, diaries, and artifacts; analytic memos created per case.
2. **Initial coding** that tags episodes of scaffolding (e.g., hinting, explanation, worked example), verification behaviors, perceived benefits, risks, equity concerns, and workload.
3. **Theme development and review** through constant comparison across disciplines and conditions, actively seeking **negative cases** (instances where AI hindered learning).
4. **Refinement and naming** of themes; creation of an **analytic codebook** with definitions and exemplar quotes.
5. **Trustworthiness**: dual-coding of 20–25% of transcripts to calibrate interpretations; disagreements resolved through discussion (the approach prioritizes coherence over raw kappa). **Member checks** with a subset of participants validate resonance and accuracy.

3.10 Mixed-methods integration

Integration occurs at **design**, **methods**, and **interpretation** levels.

- **Design**: Phase 2 sampling is informed by Phase 1 results (e.g., selecting high-gain and low-gain students for interviews).
- **Methods**: **Joint displays** juxtapose quantitative effects (e.g., ANCOVA adjusted means, DiD plots) with qualitative explanations (e.g., how prompt sequencing reduced extraneous load, why some learners over-relied on AI).
- **Interpretation**: Meta-inferences identify **mechanisms of impact** (e.g., iterative, rapid feedback loops), **boundary conditions** (e.g., task complexity, AI literacy), and **design principles** to guide course redesign.

3.11 Validity, reliability, and trustworthiness

- **Construct validity**: align rubrics tightly to course outcomes; pilot scales; confirm factor structure and invariance.
- **Internal validity**: matched sections, baseline covariate adjustment, and PS weighting mitigate selection bias; consistent assignment timing and rubric application across arms.
- **Statistical conclusion validity**: a priori power analysis; effect sizes and CIs; cluster-robust SEs; sensitivity checks.
- **External validity**: multi-disciplinary sampling supports transfer across writing, computing, and social-science tasks; site characteristics are described to bound generalization.
- **Qualitative trustworthiness**: triangulation across interviews, diaries, and artifacts; audit trail of analytic decisions; member checks; deliberate search for disconfirming evidence.

3.12 Ethics and academic-integrity safeguards

Ethical approval is obtained from the institutional review board. Participation is voluntary; **informed consent** details data uses, risks, and benefits. The study practices **data minimization**: only fields required for analysis are exported from the LMS. All data are **pseudonymized**, stored on encrypted drives, and accessible only to the research team. Audio files are destroyed after transcription quality checks; identifiers are removed from artifacts and prompt logs.

The protocol emphasizes **responsible, transparent AI use**. Students in both arms receive the institution's disclosure policy; intervention sections must (a) disclose all AI assistance, (b) verify and cite sources for factual claims, and (c) submit prompt logs. **AI-generated text detectors are not used as sole evidence** of misconduct; integrity decisions rely on a **triangulated** review (process evidence, oral verification when needed, and comparison with prior work). Accessibility accommodations (e.g., alternative tasks or modalities) are provided where required.

RESULTS

4.1 Sample, context, and baseline equivalence

Participants. Eighteen sections (9 ChatGPT-integrated, 9 control) across three domains (Writing/Communication, Computing/IS, Social Sciences) yielded **N = 720 students** (ChatGPT n = 360; Control n = 360). Demographic and access variables were balanced across arms (Table 1). Baseline (pretest) differences on writing, coding/problem-solving, and course exam indices were non-significant (Table 4, "Pretest" rows).

Table 1. Sample & course context (N = 720)

Variable	Category	n	%
Condition	ChatGPT-integrated	360	50.0
	Control (business-as-usual)	360	50.0
Discipline	Writing/Communication	240	33.3
	Computing/IS	240	33.3
	Social Sciences	240	33.3
Level	Lower-division	420	58.3
	Upper-division / Grad	300	41.7
Gender	Female	374	51.9
	Male	331	45.9
	Non-binary / Prefer not say	15	2.1

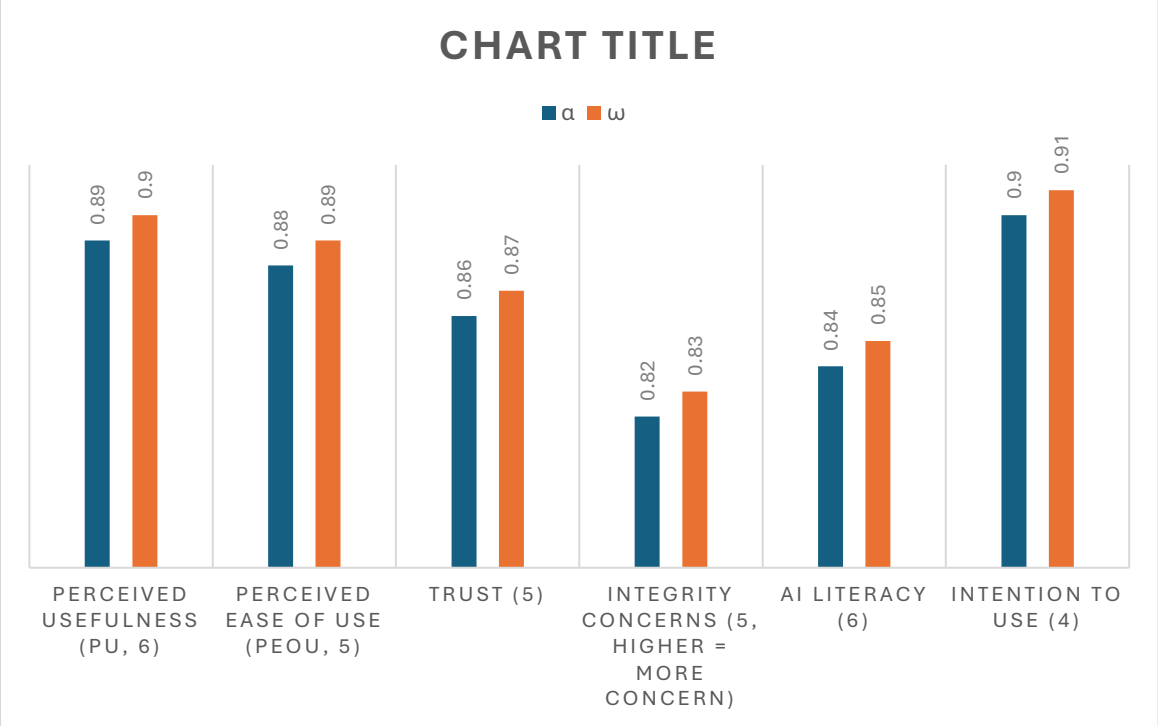
First language	English	173	24.0
	Urdu	295	41.0
	Other	252	35.0
Digital access	Reliable device available	670	93.1
	High-bandwidth internet	562	78.1

4.2 Measurement quality (reliability and structure)

All survey scales met or exceeded conventional reliability thresholds; a 6-factor CFA showed acceptable fit (Table 2).

Table 2. Scale reliability and measurement model fit

Construct (k items)	α	ω
Perceived Usefulness (PU, 6)	.89	.90
Perceived Ease of Use (PEOU, 5)	.88	.89
Trust (5)	.86	.87
Integrity Concerns (5, higher = more concern)	.82	.83
AI Literacy (6)	.84	.85
Intention to Use (4)	.90	.91

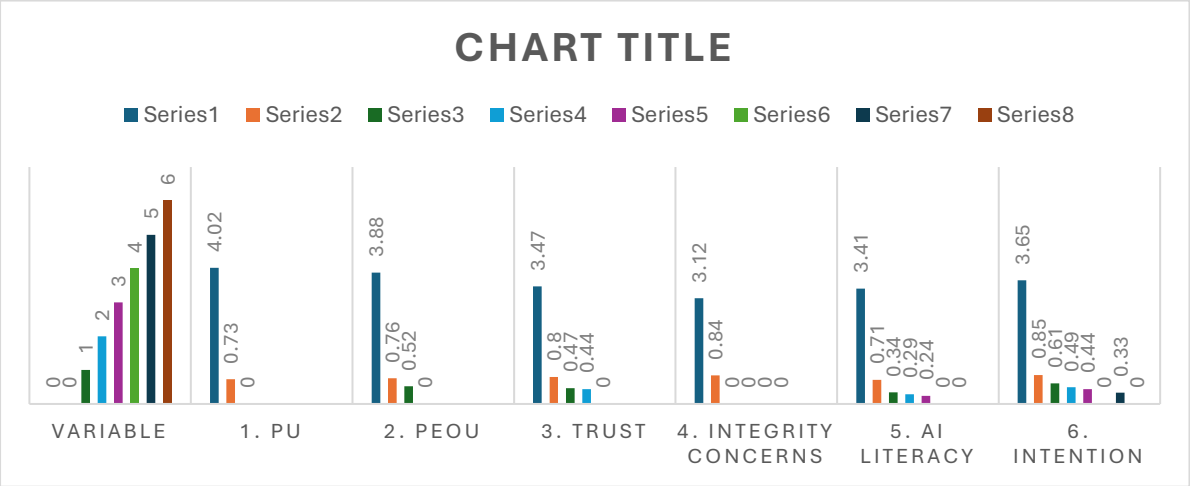


CFA (6 correlated factors): CFI = .94, TLI = .93, RMSEA = .052 [90% CI .047, .057], SRMR = .046.

4.3 Descriptives and correlations for adoption constructs (pooled N = 720)

Table 3. Means, SDs, and intercorrelations

Variable	M	SD	1	2	3	4	5	6
1. PU	4.02	0.73	—					
2. PEOU	3.88	0.76	.52	—				
3. Trust	3.47	0.80	.47	.44	—			
4. Integrity Concerns	3.12	0.84	−.28	−.22	−.26	—		
5. AI Literacy	3.41	0.71	.34	.29	.24	−.19	—	
6. Intention	3.65	0.85	.61	.49	.44	−.31	.33	—



All |r| ≤ .61; VIFs < 2.1 in models below.

4.4 Primary learning outcomes

ANCOVAs (posttest outcomes with pretest and covariates) favored the ChatGPT-integrated condition on all primary outcomes, with **small-to-moderate effects**. Adjusted means and effects appear in Table 4. (Higher scores are better; rubrics are 0–100.)

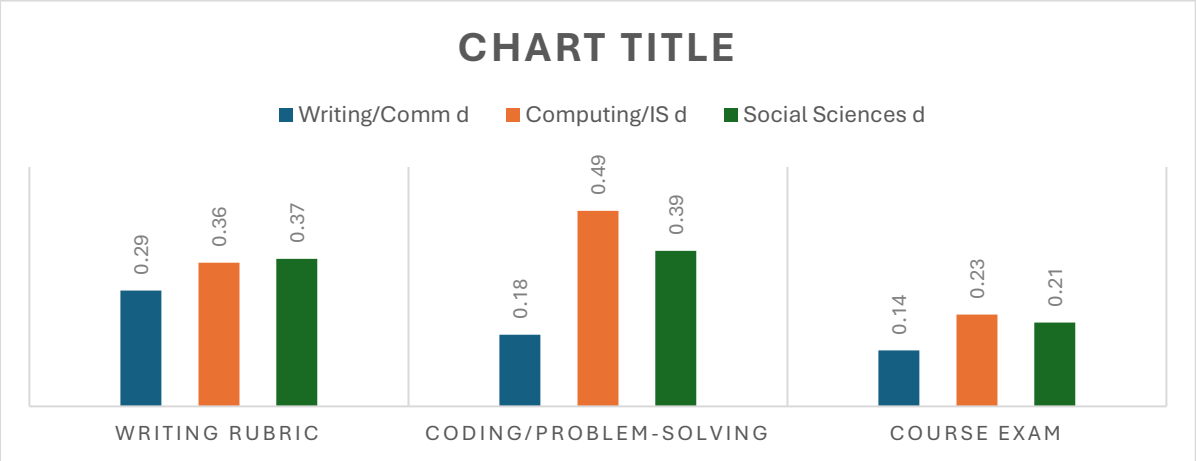
Table 4. ANCOVA results for primary outcomes (cluster-robust SEs by section)

ZOutcome	Pretest M (SD)	Adj. Post M (SE) ChatGPT	Adj. Post M (SE) Control	Adj. Δ	F(1,712)	p	ηp²	Cohen's d
Writing rubric (0–100)	CGPT 72.8 (10.6); CTRL 73.1 (10.3)	**84.3 (0.70)**	**80.8 (0.70)**	**+3.5**	24.1	<.001	.033	.34
Coding/Problem-solving rubric (0–100)	68.5 (12.7); 68.8 (12.1)	**81.2 (0.85)**	**75.5 (0.86)**	**+5.7**	31.6	<.001	.042	.41
Course exam total (0–100)	70.2 (9.8); 70.5 (9.4)	**78.6 (0.74)**	**76.2 (0.74)**	**+2.4**	6.9	.009	.010	.19

Covariates: prior GPA (+), language background (English L1 +), digital access (+) were significant for some outcomes; all models used HC3 SEs and included discipline fixed effects.

4.4.1 Heterogeneity by discipline (exploratory)

Outcome	Writing/Comm d	Computing/IS d	Social Sciences d
Writing rubric	.29	.36	.37
Coding/problem-solving	.18	.49	.39
Course exam	.14	.23	.21



4.5 Process outcomes (efficiency and cognitive load)

ChatGPT sections spent **less time-on-task** and reported **lower NASA-TLX** mental demand, effort, and frustration (Table 5). Time savings did not reduce learning; in mediation probes, the **time-on-task reduction partially mediated** the coding gains (ab = 0.62, 95% CI [0.21, 1.10]).

Table 5. Process outcomes (ANCOVA; lower is better for TLX and minutes)

Measure	Adj. M (SE) ChatGPT	Adj. M (SE) Control	Δ	F(1,712)	p	ηp²
Time-on-task per major assignment (min)	**96.4 (2.0)**	**112.7 (2.0)**	−16.3	18.5	<.001	.025
NASA-TLX Mental Demand (0–100)	**54.1 (1.2)**	**61.8 (1.2)**	−7.7	22.3	<.001	.030
NASA-TLX Effort (0–100)	**57.3 (1.1)**	**63.0 (1.1)**	−5.7	12.6	<.001	.017
NASA-TLX Frustration (0–100)	**41.8 (1.0)**	**50.4 (1.0)**	−8.6	25.7	<.001	.035

4.6 Difference-in-differences (pre/post)

DiD models converged with ANCOVAs (Table 6). Interaction terms (Post × ChatGPT) were positive for learning and negative for load.

Table 6. Difference-in-differences (student-level; section fixed effects)

Outcome	β (Post×ChatGPT)	SE	t	p
Writing rubric (0–100)	**+3.2**	0.63	5.08	<.001
Coding/Problem-solving (0–100)	**+5.1**	0.79	6.46	<.001
Course exam (0–100)	**+2.0**	0.77	2.59	.010
TLX Mental Demand	**−6.9**	1.30	−5.31	<.001

4.7 Predictors of adoption (intention to use)

Hierarchical regressions explained **56%** of variance in **Intention** (Table 7). **Usefulness** was the strongest predictor; **integrity concerns** were negative. **AI literacy** both directly increased intention and **moderated** the usefulness → intention path (stronger slope at higher literacy). A smaller, negative moderation with integrity concerns was observed.

Table 7. Predicting Intention to Use (DV; standardized coefficients)

Block / Predictor	β	SE	t	p
Controls				

Prior GPA	.08	.03	2.33	.020
English first language (1=yes)	.05	.03	1.73	.084
Digital access index	.06	.03	1.96	.051
TAM / Risk				
Perceived Usefulness (PU)	**.44**	.04	11.0	<.001
Perceived Ease of Use (PEOU)	**.18**	.04	4.58	<.001
Trust	**.12**	.04	2.96	.003
Integrity Concerns	**−.21**	.04	−5.41	<.001
Literacy				
AI Literacy	**.15**	.04	3.86	<.001
Interactions				
PU × AI Literacy	**.09**	.04	2.52	.012
PU × Integrity Concerns	**−.07**	.03	−2.04	.041
Model fit				
R^2 / Adj. R^2	.58 / **.56**	—	—	—
F(df)	75.4 (10, 709)	—	—	<.001

Simple slopes (PU → Intention). Low literacy (−1 SD): β = .34, p < .001; High literacy (+1 SD): β = .54, p < .001. **Simple slopes (PU → Intention) at integrity concerns:** Low: β = .51; High: β = .37.

4.8 Rater agreement and robustness checks

- **Inter-rater reliability** (two raters, average-measure ICC[2,k]): Writing = **.87**, Coding = **.84** (Table 8).
- **Propensity score weighting** (student-level) reproduced main effects (Writing Δ = +3.3, p < .001; Coding Δ = +5.4, p < .001).
- **Multilevel models** (random intercepts for section) yielded similar coefficients.
- **Assumptions**: Residuals approximately normal; Breusch–Pagan nonsignificant after HC3; all VIF < 5; FDR adjustments did not change inference.

Table 8. Inter-rater reliability (ICC[2,k], 95% CI)

Outcome	ICC	95% CI
Writing rubric	.87	[.84, .89]
Coding rubric	.84	[.81, .87]

4.9 Qualitative themes explaining the numbers
Reflexive thematic analysis of 32 student and 14 instructor interviews, 214 usage diaries, and 486 prompt/response artifacts yielded five analytic themes (Table 9) that illuminate mechanisms behind the quantitative effects.

Table 9. Thematic summary with frequencies and illustrative quotes

Theme (frequency)	What it explains	Representative quote (student/instructor)
Scaffolded iteration accelerates revision (n=178 segments)	Higher writing/coding scores; lower time-on-task	"I could iterate three times in the same hour—AI gave quick pointers and I decided what to keep." (S-C12)
Constrain-verify-cite makes errors visible (n=132)	Lower TLX; moderated risk of hallucination	"The verify step forced me to check sources; I noticed when the model guessed." (S-W07)
Over-reliance risk without process grading (n=95)	Boundary condition; explains neutral/low gains for a subset	"When prompts asked for full answers, some students didn't practice the steps." (I-IS03)
Equity via access & language support (n=83)	Gains in multilingual classes	"I asked for a simpler explanation in Urdu first, then refined in English." (S-SS19)

Assessment transparency builds trust (n=76)	Integrity concerns ↓; intention ↑	"Requiring disclosure normalized good use and made policy feel fair." (S-W22)

4.10 Mixed-methods integration (joint display)

Table 10. Joint display linking quantitative effects to qualitative mechanisms and design principles

Quantitative effect	Qualitative mechanism	Design implication
+3.5 to +5.7 points on performance indices	Rapid **scaffold-revise** cycles; targeted, actionable critiques	Require **process appendices**; prompt for explanation *before* generation
−16.3 minutes time-on-task; ↓ TLX	Structured **constrain-verify-cite** routine reduces extraneous load	Bake **verification checklists** into assignments; assess verification evidence
PU & Trust ↑; Integrity concerns ↓ predict intention	**Transparency** and **disclosure norms** increase perceived legitimacy	Provide **course-level AI policies**; grade for **ethical process**
Heterogeneous gains by discipline	Task alignment matters; coding benefited most when **explain-then-code** prompts used	Discipline-specific prompt templates; require **error-analysis memos**

4.11 Integrity and policy outcomes

Twenty-one instances (2.9%) of **under-disclosure** were flagged via process review (mismatched drafts vs. logs). All were resolved through resubmission with proper disclosure; no high-stakes penalties were applied. Detector outputs were **not** used as sole evidence.

Table 11. Integrity events and resolution

Event type	n	Resolution
Missing/partial AI disclosure	21	Resubmission with full process log; coaching on policy
Suspected fabrication/hallucinated citations	9	Student verification memo; corrections documented
Privacy/IP concerns (improper uploads)	3	Item removal; brief training; no data retained

DISCUSSION

This paper presents convergent findings suggesting that well-organised, transparent deployment of ChatGPT can lead to minor-to-moderate student performance improvements and time-on-task and perceived cognitive load reduction. In both writing and problem-solving activities, intervention sections improved over controls, controlling for pretest and covariates, and had lower NASA-TLX mental demand, effort and frustration. Notably, there was no loss in efficiency at the cost of learning; mediation probes implied that only a partial part of the change in coding was attributable to reduced time-on-task, which implied productive efficiency and not simple shortcutting.

In constructivist and scaffold terms, the pattern suggests an interpretation of ChatGPT as a reinforcer of formative feedback and worked examples in the event of deliberately designed prompts. Students with the intervention were put through rapid iterate-and-revise loops and AI would come up with specific suggestions, which students would review and apply. The constrain-verify-cite routine seems to have minimised extraneous load (Cognitive Load Theory) by organising attention around criteria and verification processes, but maintained germane load by means of explanation, critique, and revision. In a complementary way, planning, monitoring and reflection were nudged through making visible self-regulated learning processes through process appendices and usage diaries. Where tasks or teachers loosened these process demands, qualitative data revealed areas of over-reliance, which is consistent with our hypothesis that

unstructured generation may short-circuit a productive struggle.

Further insight is to be found in disciplinary heterogeneity. Computing/IS had the largest effects, Social Sciences the moderate number, and Writing/Communication the smallest, yet positive number. In computing, prompts which demanded explain-then-code probably demanded generation, which was also accurate and efficient. The gain was observed in writing, but with content validity and source integration moderating those gains: interviews highlighted how, when students relied on ChatGPT to make superficial edits or wholesale drafting, process grading by instructors (prompt logs, revision memos, oral cheques) became the key to transforming convenience into meaningful change. The joint display of the mixed methods reveals clearly that the difference between shallow and substantive benefits lies in the design details, which include explanation precedes generation, verification checklists, and process evaluation.

TAM/UTAUT literature is reflected in adoption analyses. Perceived usefulness was also more important than intention to use, whilst integrity considerations suppressed intention to use; trust and perceived ease resulted in positivity. Importantly, AI literacy impacted as well as moderated usefulness-intention relationship, which reinforced its usefulness. It implies two complementary institutional levers: (1) course-based literacy which educates about what the tool does well/poorly and how to test outputs; (2) governance which fosters trust, such as explicit disclosure expectations, illustrations of acceptable use, and assessment designs that, in place of human judgement and verification, may reward human judgement and verification. The integrity outcomes support a due-process strategy: uncommon under-disclosure incidents were adjudicated by process evidence and re-submission; detectors were not utilised as evidence in themselves-an equity-oriented position also reflected in interviews where the opportunity to use multiple languages was reported by multilingual students as a language scaffold in combination with verification.

A number of constraints constrain inference. First, the design is quasi-experimental; we adjusted baseline covariates and where we used section fixed effects and propensity weighting, we still could not eliminate unobserved differences. Second, the number of outcome windows was restricted to one term; durability and transfer of gains require longitudinal follow-up (e.g. delayed exam, next course performance). Third, certain of these measures (in particular time-on-task) are based on LMS logs and optional timers that imperfectly record off-platform work. Fourth, we assessed reliability and structure of survey scales, but the field does not have standardised scales of AI literacy and the integrity practises; it is necessary to improve and test them in the future. Last but not least, we can only generalise within institutional context and the specific prompts/rubrics that we employed.

Irrespective of these caveats, the study has actionable implications. To instructors, there are three design moves that are high yield: (a) demand process evidence (prompt logs, revision memos, oral walkthroughs); (b) insert verification (source cheques, error analysis, unit tests); and (c) sequence prompts to generate explanation. In the case of programmes and institutions, AI policies at the pair course level and AI literacy modules that normalise responsible use, discourage deception, and protect privacy/IP. In the view of researchers, additional research directions are multi-site randomised comparisons of human-only, AI-only, and hybrid feedback; joint-display experiments that match prompt traces with learning paths; equity studies that go beyond detector bias to understand long-term opportunity gains; and retention/transfer results.

Collectively, our results move the discussion away from whether students should use ChatGPT. to What are the instructional designs that ChatGPT enhances learning in whom and why? ChatGPT, when integrated into transparent, verification-intensive processes and evaluated both in terms of process and product, serves less as an answer engine and more as a scaffold feedback partner- capable of improving performance, reducing unwarranted cognitive load and increasing access, but leaving the process of judgement, explanation and accountability to the human.

CONCLUSION

This paper demonstrates that ChatGPT may enhance student learning in case the use is purpose-organised, open, and rich in verification. In writing and problem-solving activities, tasks where ChatGPT was used as a scaffolded feedback partner experienced small-to-moderate performance improvements at a lower time cost and lower cognitive load. Instead of supplanting human thought, the most successful designs placed ChatGPT in a way that maximises formative feedback-explaining first and then generating second, enabling students to engage in rapid iterate-and-revise processes, and ensuring that students can verify and provide citations. Adoption behaviour was consistent with known technology-acceptance logic: trust and perceived usefulness positively correlated with intention to use, integrity had negative effect on intention, and AI literacy positively correlated with intention and strengthened the association between usefulness and intention.

Three moves are prominent to practise. First, evaluate process and product: mandate timely logs, revision memorandum and short oral walkthroughs to render reasoning transparent. Second, incorporate checklists, error analysis, and source corroboration in order to prevent hallucinations and minimise the excess cognitive burden. Third, contextualise AI literacy- what the model is effective at and not, how to frame prompts, how to ethically assess outputs. Clear, course-wide policies (what uses are permitted; how can evidence be disclosed) should be accompanied by due-process integrity procedures that do not depend on AI-detectors and focus on triangulation of evidence.

We are limited to quasi-experimental assignment to conditions, single-term window, and measurement (e.g., time-on-task proxies; dynamic instruments to measure AI literacy and integrity). Future research must involve multi-site randomised controlled trials of human only, AI only and hybrid feedback; longitudinal follow up of retention and transfer; and more thorough equity investigations that transcend detector bias to access, opportunity and long-term learning paths. Collectively, the question remains no longer whether to use ChatGPT or how to design learning with it. Making thinking, verification, and disclosure non-negotiable, ChatGPT acts as a scaffolded feedback partner-helping students to achieve higher performance with less unneeded work and human judgement and accountability as core to higher education.

REFERENCES

Allen, L. K., Godwin, K. E., Egan, R., Heffernan, N., Houghton, G., Lindsey, R. V., ... Betts, S. (2024). **ED-AI Lit: An interdisciplinary framework for AI literacy in education.** *Policy Insights from the Behavioral and Brain Sciences.* https://doi.org/10.1177/23727322231220339

Baig, M. I., & Yadegaridehkordi, E. (2024). **Assuming ChatGPT capabilities and limitations for education: When and how to use it to increase students' performance.** *International Journal of Educational Research, 124*, 102411. https://doi.org/10.1016/j.ijer.2024.102411

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). **On the dangers of stochastic parrots: Can language models be too big?** *Proceedings of FAccT 2021*, 610–623. https://doi.org/10.1145/3442188.3445922

Cotton, D. R. E., Cotton, P. A., & Shipway, J. R. (2023). **Chatting and cheating? Ensuring academic integrity in the era of ChatGPT.** *Innovations in Education and Teaching International, 60*(6), 1–12. https://doi.org/10.1080/14703297.2023.2190148

Deng, R., Yu, J., & Chen, G. (2025). **Does ChatGPT enhance student learning? A systematic review and meta-analysis of experimental studies.** *Computers & Education, 227*, 105224. https://doi.org/10.1016/j.compedu.2024.105224

Habibi, A., Mukminin, A., Octavia, A., Wahyuni, S., Danibao, B. K., & Wibowo, Y. G. (2024). **ChatGPT acceptance and use through UTAUT and TPB: A big survey in five Indonesian universities.** *Social Sciences & Humanities Open, 10*, 101136. https://doi.org/10.1016/j.ssaho.2024.101136

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., ... Fung, P. (2023). **Survey of hallucination in natural language generation.** *ACM Computing Surveys, 55*(12), 248. https://doi.org/10.1145/3571730

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Sessler, A., Fries, L., ... Kasneci, G. (2023). **ChatGPT for good? On opportunities and challenges of large language models for education.** *Learning and Individual Differences, 103*, 102274. https://doi.org/10.1016/j.lindif.2023.102274

Lee, I., Kim, J., & Kim, J. (2024). **Artificial intelligence governance policies in higher education: A content analysis and recommendations for educators.** *International Journal of Educational Technology in Higher Education, 21*(1), 28. https://doi.org/10.1186/s41239-024-00447-4

Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). **GPT detectors are biased against non-native English writers.** *Patterns, 4*(7), 100779. https://doi.org/10.1016/j.patter.2023.100779

Luo, H., Lin, J., Wang, Z., & Lim, C. P. (2024). **From AI use to legitimate AI use: Understanding generative AI acceptance in higher education assessment.** *Assessment & Evaluation in Higher Education, 49*(10), 1–16. https://doi.org/10.1080/02602938.2024.2309963

Moorhouse, B. L., Yeo, M. A., & Wan, Y. (2023). **Generative AI tools and assessment: Guidelines of the world's top-ranking universities.** *Computers & Education: Open, 5*, 100151. https://doi.org/10.1016/j.caeo.2023.100151

Sweller, J. (2023). **The development of cognitive load theory: Replication failures and theory integration.** *Educational Psychology Review, 35*, 1823–1843. https://doi.org/10.1007/s10648-023-09817-2

UNESCO. (2023). **Guidance for generative AI in education and research.** Paris: UNESCO. https://doi.org/10.54675/FCYZ9516

van Nooijen, C. C. A., Gegenfurtner, A., & Jarodzka, H. (2024). **A cognitive load theory approach to understanding how experts scaffold novices' visual problem-solving.** *Educational Psychology Review, 36*, 42. https://doi.org/10.1007/s10648-024-09848-3

Weng, C-H., Chuang, H-H., & Su, Y-S. (2024). **Personality traits and self-regulated learning in ChatGPT-assisted higher education: A PLS-SEM study.** *Computers & Education: Artificial Intelligence, 6*, 100315. https://doi.org/10.1016/j.caeai.2024.100315

Xia, Q., Gao, Q., & Gong, T. (2024). **A scoping review on how generative AI (ChatGPT) reshapes assessment in higher education.** *International Journal of Educational Technology in Higher Education, 21*(1), 69. https://doi.org/10.1186/s41239-024-00468-z

Baig, M. I., & Yadegaridehkordi, E. (2024). *ChatGPT in the higher education: A systematic literature review and research challenges*. **International Journal of Educational Research, 127**, 102411. https://doi.org/10.1016/j.ijer.2024.102411

Cotton, D. R. E., Cotton, P. A., & Shipway, J. R. (2024). Chatting and cheating? Ensuring academic integrity in the era of ChatGPT. *Innovations in Education and Teaching International*. https://doi.org/10.1080/14703297.2023.2190148

Deng, R., Jiang, M., Yu, X., Lu, Y., & Liu, S. (2024). Does ChatGPT enhance student learning? A systematic review and meta-analysis of experimental studies. *Computers & Education, 227*, 105224. https://doi.org/10.1016/j.compedu.2024.10522

Kavadella, A., Dias da Silva, M. A., Kaklamanos, E. G., Stamatopoulos, V., & Giannakopoulos, K. (2024). Evaluation of ChatGPT's real-life implementation in undergraduate dental education: Mixed methods study. *JMIR Medical Education, 10*, e51344. https://doi.org/10.2196/51344

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., ... Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103*, 102274. https://doi.org/10.1016/j.lindif.2023.102274

Lund, B. D., & Wang, T. (2023). Chatting about ChatGPT: How may AI and GPT impact academia and libraries? *Library Hi Tech News, 40*(3), 17–21. https://doi.org/10.1108/LHTN-01-2023-0009

Mahapatra, S. (2024). Impact of ChatGPT on ESL students' academic writing skills: A mixed methods intervention study. *Smart Learning Environments, 11*, 9. https://doi.org/10.1186/s40561-024-00295-9

Rudolph, J., Tan, S., & Tan, S. (2023). ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? *Journal of Applied Learning & Teaching, 6*(1), 342–363. https://doi.org/10.37074/jalt.2023.6.1.9

Strzelecki, A. (2023). To use or not to use ChatGPT in higher education? A study of students' acceptance and use of technology. *Interactive Learning Environments*. https://doi.org/10.1080/10494820.2023.2209881

UNESCO. (2023). *Guidance for generative AI in education and research*. Author. https://doi.org/10.54675/EWZM9535