Recognising AI Sycophancy in Irish Higher Education: A Cross-Sectional Vignette Study of Verification, Critical Evaluation and AI Dependence

Abdulaziz Zaid Albasheer, (Phd) [1], Mohammad A Alshagare[2],

[1]Lecturer, Department of English Language and Literature, Imam Mohammad Ibn Saud Islamic University

2 Mohammad A Alshagare, Lecturer, Department of English Language and Literature, Imam Mohammad Ibn Saud Islamic University (IMSIU), Saudi Arabia. maalshagare@imamu.edu.sa

Corresponding: **Mohammad A Alshagare**

HOW TO CITE:
Abdulaziz Zaid Albasheer , Mohammad A Alshagare, (2026). Recognising AI Sycophancy in Irish Higher Education: A Cross- Sectional Vignette Study of Verification, Critical Evaluation and AI Dependence. International Journal of Special Education, 41(19s), 265-289

ABSTRACT:

Generative artificial intelligence is now used across many academic tasks, but frequent use does not show whether students can recognise output that agrees with them when stronger evidence calls for qualification or correction. This study investigated AI sycophancy, verification behaviour, self-reported evaluation and AI dependence in a cross-sectional survey of 180 students from six Irish higher education institutions. The instrument combined use measures, five multi-item scales and six paired vignettes in which students selected between a sycophantic response and an evidence-sensitive response. Academic AI use was frequent: 90.6% of participants reported using generative AI at least weekly. Only 23.9% had previously heard the term AI sycophancy, although 67.2% reported recognising the experience after a neutral definition. Mean vignette recognition was 3.58 of 6, equivalent to 59.7%. Across all vignette decisions, 31.5% of selections preferred the sycophantic response and 32.1% indicated an intention to use it. Verification was positively associated with vignette recognition ($r = 0.52$, 95% CI [0.40, 0.62]), while affirming-response preference was associated with AI dependence ($r = 0.33$, 95% CI [0.19, 0.45]). AI-use intensity was associated with dependence ($r = 0.48$) but not with vignette recognition ($r = -0.03$). Self-reported critical evaluation exceeded vignette performance by 10.19 percentage points on a common 0 to 100 scale, although the two measures are not psychometrically equivalent. The findings point to a calibration and verification problem rather than evidence of cognitive decline. Because the sample was purposive and the instrument was newly developed, the results should not be interpreted as national prevalence estimates or causal effects.

Keywords: AI sycophancy; critical thinking; verification; metacognition; generative AI; higher education; Ireland.

1. Introduction

Generative artificial intelligence has moved quickly from a specialist technology to a routine

academic tool. Students can now use conversational systems to plan assignments, summarise readings, edit prose, generate

explanations, locate sources, test arguments and solve technical problems. These functions can reduce routine effort and provide rapid feedback. They can also influence the intellectual decisions that sit behind an academic product (Ali et al., 2025; Ali et al., 2023; Ali et al., 2024). A model can frame a question, choose which evidence to foreground, rank competing explanations and signal that a user's starting position is reasonable (Alenazi et al., 2026). The educational issue is therefore not limited to factual accuracy or authorship (Ashfaq et al., 2021; Ayub et al., 2025; Bacha et al., 2026). It also concerns how interaction with a model changes judgement.

AI sycophancy describes a pattern in which a model accommodates a user's stated belief, preference or self-presentation when a more evidence-sensitive response would require correction, uncertainty or counterargument. It differs from hallucination. Hallucination concerns unsupported or false content, whereas a sycophantic response may contain individually accurate statements but select or organise them in a way that protects the user's preferred conclusion. It also differs from ordinary politeness and personalisation. A response can be respectful and tailored while still testing assumptions and representing uncertainty.

Research on language-model behaviour provides a clear reason to study this problem in education (Bajwa et al., 2022; Gul et al., 2021; Hassan et al., 2022). Perez et al. (2023) and Sharma et al. (2024) showed that models can shift towards a user's stated position and that human preference signals can reward agreeable responses. More recent work indicates that the problem can persist across multi-turn dialogue rather than appearing only in isolated prompts (Hong et al., 2025). Cheng et al. (2026) extended the concern from model behaviour to user consequences. Across three preregistered experiments, sycophantic AI increased users' conviction that they were right and was nevertheless trusted and preferred. That work concerned interpersonal decisions rather than higher education, but the mechanism is relevant to academic reasoning: agreement can feel diagnostic even when it does not add independent evidence.

This distinction matters because critical thinking is not simply the production of a plausible answer. It involves assessing evidence, testing assumptions, comparing explanations and giving reasons for accepting or rejecting a claim (Facione, 1990). Metacognition concerns the monitoring of one's own knowledge, confidence and reasoning (Flavell, 1979). A student may know in general terms that AI can be wrong yet still respond to fluency, certainty or validation as though these features increase evidential quality. In that setting, awareness of technical fallibility does not guarantee effective verification.

The concern should not be converted into a general claim that AI weakens thinking. Evidence on educational use is conditional. A randomised study of an AI tutor in undergraduate physics reported larger short-term learning gains than an active-learning class when the system was designed around established pedagogical principles (Kestin et al., 2025). By contrast, Bastani et al. (2025) found that access to an unguarded generative system could improve supported mathematics performance while reducing later unassisted performance, whereas a safeguarded tutor reduced that loss. Fan et al. (2025) likewise reported gains in immediate writing performance without equivalent gains in transfer. These studies suggest that the educational effect of AI depends on what cognitive work the system replaces, what it scaffolds and how students inspect its output.

Irish higher education is a useful setting in which to study these questions. QQI's 2025 generative AI survey included 1,005 learners and 1,229 staff across Irish tertiary education and reported varied levels of knowledge, use and institutional provision. The Higher Education Authority policy framework published in 2025 places critical engagement, human oversight and AI literacy among its core principles. National policy is therefore moving beyond simple detection of AI use towards questions of judgement, assessment and responsible integration (QQI, 2025; O'Sullivan et al., 2025). What remains less clear is whether students can discriminate between an agreeable answer and an evidence-

sensitive answer when both are fluent and apparently helpful.

The present study addresses that gap through a survey with paired vignettes. It separates four elements that are often collapsed into a single idea of AI literacy: use intensity, behavioural recognition of sycophancy, self-reported evaluation and verification, and reliance on AI in academic reasoning. It also records preference for the affirming response independently of recognition. This distinction allows a student to recognise which response is stronger while still preferring or intending to use the more agreeable response. The design cannot establish cognitive decline or causal effects of AI exposure. Its value lies in testing how reported habits, demonstrated discrimination and preference relate within the same sample.

2. Literature Review and Conceptual Basis

2.1 Generative AI use and the difference between performance and learning

Research on artificial intelligence in higher education predates large language models, but earlier work often centred on prediction, profiling and intelligent tutoring rather than direct student use (Zawacki-Richter et al., 2019). Generative systems changed the pattern of access because students can apply the same tool across writing, revision, coding, explanation and information-seeking tasks. Reviews published soon after the release of widely accessible chatbots described benefits in feedback, explanation, language assistance and idea generation alongside risks involving inaccuracy, bias, privacy, assessment and overreliance (Kasneci et al., 2023; Lo, 2023; Tlili et al., 2023). These categories describe what may happen, but they do not by themselves explain why a fluent answer can alter judgement.

A more useful distinction separates tool-assisted performance from learning that remains available when the tool is absent. Kestin et al. (2025) show that a well-designed tutor can increase learning under controlled conditions. Bastani et al. (2025) show that unrestricted access can improve performance during use while weakening later unassisted performance, and that instructional safeguards can reduce this effect. Fan et al. (2025) report a related separation between immediate writing gains and transfer. The common point is not that assistance is harmful. It is that the educational effect depends on whether AI preserves reasoning, feedback uptake and error correction or substitutes for them.

This distinction is important for sycophancy because an agreeable response can increase the apparent efficiency of a task while reducing the probability that a student searches for disconfirming evidence. A student who receives a confident defence of an initial thesis may complete an assignment more quickly without noticing that the search process has narrowed. Adoption frequency alone cannot reveal this mechanism.

2.2 Sycophancy as an interaction failure

Modern assistants are trained to follow instructions and satisfy user preferences. Reinforcement learning from human feedback has been influential in this development (Ouyang et al., 2022). Preference optimisation can, however, create a tension between agreeableness and truthfulness. Sharma et al. (2024) found sycophantic behaviour across several free-form tasks and reported that responses matching a user's views were more likely to be preferred. Perez et al. (2023) similarly used model-written evaluations to reveal systematic behaviour that changes with user statements and preferences.

The interaction can become more consequential over several turns. Hong et al. (2025) found that models can shift their stance under sustained user pressure in multi-turn scenarios. This matters in academic work, where students rarely ask one isolated question (Irfan and Ali, 2022; Irfan, Aldulaylan and Alqahtani, 2023; Irfan et al., 2026). They refine a thesis, request supporting sources, ask for counterarguments and return to earlier claims. A model that gradually accommodates the user can influence the path of inquiry even if no single sentence is obviously false (Irfan, Almeshal and Anwar, 2023).

Cheng et al. (2026) provide evidence that sycophancy can affect users as well as model outputs. In interpersonal scenarios, sycophantic

responses increased users' conviction and reduced willingness to repair conflict, while users rated the responses favourably. The educational setting differs, and the effects cannot simply be transferred to students (Irfan et al., 2026). The study nevertheless shows a preference problem that is directly relevant here: people can like and trust a response precisely because it validates them. For academic use, the resulting risk is not mere flattery but a change in which evidence is sought, how uncertainty is interpreted and when verification stops (Irfan, Ashfaq and Azad, 2020).

Disagreement is not a sufficient remedy (Irfan, 2015). A model can challenge a user and still be wrong. The relevant quality is evidence-sensitive challenge: testing assumptions, separating association from causation, representing credible alternatives, stating uncertainty and indicating what evidence would change the conclusion. This standard allows supportive communication without equating helpfulness with agreement (Irfan and Firdous, 2024).

2.3 Epistemic vigilance, automation and confirmation

Epistemic vigilance refers to the processes by which people assess information and the reliability of its source (Sperber et al., 2010). In conventional academic work, these processes include checking provenance, comparing sources, inspecting methods and deciding whether a claim is warranted. Generative AI complicates these practices because the system presents an integrated answer rather than a transparent chain of source documents. The user must therefore evaluate both the content of the response and the conditions under which it was produced.

Automation research provides a related account. Automation bias occurs when users accept automated advice or fail to notice errors, especially when checking is effortful or the system is usually useful (Parasuraman and Riley, 1997; Goddard, Roudsari and Wyatt, 2012; Irfan & Murray, 2023). Algorithm appreciation similarly suggests that people may sometimes assign greater weight to algorithmic advice than to comparable human advice, although the effect varies according to expertise and task conditions (Logg, Minson and Moore, 2019). Sycophancy may intensify this problem by combining the perceived authority of an automated system with interpersonal validation. The model does not simply recommend an answer; it may also reinforce the user's belief that their initial position was reasonable.

Confirmation bias provides a third component. People tend to seek, interpret or prioritise information in ways that preserve existing beliefs (Nickerson, 1998). Conversational AI may reduce the effort required to obtain such confirmation because users can repeatedly request arguments, explanations or evidence that support a preferred premise. Yet confirmation bias and AI sycophancy should not be treated as the same process. Confirmation bias describes a tendency in human information selection and interpretation, whereas sycophancy concerns how an AI system responds to cues about the user's beliefs or preferences. Keeping these processes conceptually separate is important because the educational responses differ. Students require strong verification and evaluation practices, while institutions and system designers also need methods for identifying and mitigating problematic interaction behaviours.

2.4 Cognitive offloading, dependence and metacognitive calibration

Cognitive offloading is the use of external resources to reduce internal processing demands (Risko and Gilbert, 2016). Irfan and Murray (2023) explained that offloading is not inherently undesirable.

Notes, calculators and databases can release effort for more demanding reasoning. Irfan and Murray (2023) explained that the educational question is which function is being delegated. Assistance can remove routine friction. Irfan et al. (2023) further mentioned that augmentation can extend a student's reasoning. However, concerns arise when generative AI is used to support cognitive activities such as ideation. Survey evidence suggests that students may perceive reduced originality and greater

homogenisation in their work. In a survey of 100 undergraduate and postgraduate students at two Irish institutions, 71.0% believed that ideas they considered original showed similarities to typical AI-generated work, while 68.0% reported that their ideas had become more similar to those of other users (Alshagare et al., 2026). Irfan et al. (2023) also explained that substitution occurs when a system performs a learning task that the student was expected to carry out, while dependence describes difficulty functioning without the tool (Nisa, Yousafzai and Irfan, 2021; Rauf, Ali and Irfan, 2021; Shakoor et al., 2022).

Metacognition helps regulate these boundaries. Students need to judge when they understand a response, when confidence exceeds evidence and when additional checking is required (Irfan, Murray and Ali, 2023a; Irfan, Murray and Ali, 2023b; Irfan, Murray and Ali, 2023c). Lee et al. (2025), in a study of knowledge workers, found that higher confidence in AI was associated with less self-reported effort devoted to critical-thinking activities, while confidence in one's own ability was associated with more critical-thinking effort (Khan et al., 2025; Khan, Azeem and Irfan, 2021; Nawaz et al., 2025). The study does not establish loss of ability, but it provides a reason to distinguish perceived evaluation skill from observable discrimination.

A sycophantic response may create miscalibration because agreement feels like corroboration. If students report strong evaluation habits but do not consistently distinguish an evidence-sensitive response from an agreeable one, the mismatch may indicate overconfidence, measurement error, different task demands or some combination of these (Khan et al., 2021; Khan, Ali and Irfan, 2021; Khan, Ali and Irfan, 2022). A discrepancy score must therefore be interpreted with both component measures rather than as a direct measure of cognitive decline (Shoukat et al., 2026; Youcefi, Irfan and Aldulaylan, 2023).

2.5 AI literacy in Irish higher education

AI literacy is commonly described as the knowledge and competencies required to understand, use, evaluate and communicate about AI (Long and Magerko, 2020; Ng et al., 2021). In higher education, operational competence is only one part of this requirement. Students can be skilled at prompting and editing while remaining weak at source checking, uncertainty assessment or resistance to agreeable but selective reasoning. Irish guidance increasingly reflects this wider view (Irfan, 2025). NAIN guidance asks educators to clarify permitted use and redesign assessment for learning rather than relying only on detection (NAIN, 2023). QQI's national survey points to uneven knowledge and provision across the sector (QQI, 2025). The HEA framework links AI literacy with human judgement and critical engagement (O'Sullivan et al., 2025). These policies make verification and accountable reasoning appropriate educational outcomes in their own right. The literature therefore suggests a model in which AI use, susceptibility to agreement, verification, self-evaluation, behavioural recognition and dependence are related but not interchangeable. Frequent use may coexist with good or poor judgement. Recognition may coexist with a preference for the affirming response. Verification may be more informative than confidence because it refers to actions that expose a claim to independent evidence. The study tests these distinctions within one survey instrument.

2.6 Research questions

RQ1. How frequently do students use generative AI for academic work, and for which tasks?

RQ2. How much prior awareness of AI sycophancy is reported, and how accurately do students distinguish evidence-sensitive from sycophantic responses?

RQ3. How are verification, susceptibility, dependence, critical augmentation, self-reported evaluation and AI-use intensity related to behavioural recognition and affirming-response preference?

RQ4. What discrepancy is observed between self-reported critical evaluation and vignette-based recognition when both are placed on a common 0 to 100 scale?

3. Method

3.1 Design and setting

The study used a quantitative cross-sectional survey with six paired-response vignettes. The individual student was the unit of analysis. Data collection ended on 30 April 2026. The design was suited to estimating frequencies, score distributions and associations within the sampled group. It does not permit causal inference, measurement of change over time or attribution of effects to a particular AI platform.

3.2 Participants and sampling

The analytic sample comprised 180 students from six Irish higher education institutions. Eligibility required participants to be aged 18 years or older, currently enrolled in higher education and to have used at least one generative AI tool for academic purposes during the preceding six months. Sampling was purposive and stratified by institution, with 30 participants from each institution. The observed sample contained 120 undergraduates and 60 postgraduates. Disciplinary representation was 52 arts and humanities students, 50 STEM students, 46 medical and health sciences students and 32 students in nursing and related caring disciplines. Participation was voluntary and anonymous, and the questionnaire did not request names or student numbers.

Characteristic	**Category**	**n**	**%**
Study level	Undergraduate	120	66.7
	Postgraduate	60	33.3
Discipline	Arts and humanities	52	28.9
	STEM	50	27.8
	Medical and health sciences	46	25.6
	Nursing and related caring disciplines	32	17.8
Age group	265-289	48	26.7
	265-289	68	37.8
	265-289	43	23.9
	35 or older	21	11.7
Gender	Woman	91	50.6
	Man	80	44.4
	Non-binary or self-described	5	2.8

Characteristic	Category	n	%
	Prefer not to say	4	2.2
Student status	Domestic	145	80.6
	International	35	19.4
Training	Formal AI-literacy training	0	0.0
	Formal critical-thinking training	10	5.6

Table 1. Participant characteristics

Each of the six institutions contributed 30 participants: Munster Technological University, Cork; University College Cork; Trinity College Dublin; University College Dublin; Dundalk Institute of Technology; and the Institute of Art, Design and Technology, Dún Laoghaire.

3.3 Survey instrument

The questionnaire contained six sections. Section A recorded demographic and educational characteristics. Section B recorded frequency, duration, tools and academic purposes of AI use. Section C asked about awareness of AI sycophancy before a definition was shown and then asked whether participants recognised the described experience. Section D contained six paired vignettes. Section E contained five multi-item scales covering critical evaluation, verification, sycophancy susceptibility, AI dependence and critical augmentation. Section F contained two optional open questions. The open responses were not used for inferential claims in this article.

The six vignettes covered social-media interpretation, historical reasoning, STEM inference, medical evidence, nursing judgement and source quality. Each vignette presented one response that endorsed the user's premise without sufficient qualification and one response that tested the premise against evidence, causal uncertainty or alternative explanations. The location of the evidence-sensitive option alternated across the six vignettes. For each pair, participants selected the more helpful response, the response they would be more likely to use and the response that showed stronger critical thinking. Confidence was rated separately.

The five scale constructs each contained five agreement items scored from 1, strongly disagree, to 5, strongly agree. Negatively worded items in the critical-evaluation and verification scales were reverse-scored before the mean was calculated. Higher scores represented stronger perceived evaluation, stronger verification, greater susceptibility, greater dependence and stronger critical augmentation, respectively.

3.4 Scoring

AI-use intensity was a study-specific index combining ordered use frequency and task breadth. Frequency was coded from 1, less than monthly, to 4, daily or almost daily. Task breadth was the number of the 11 listed academic purposes used at least rarely. The index was calculated as $1 + 4[0.55(\text{frequency} - 1)/3 + 0.45(\text{task breadth}/11)]$, producing a range from 1 to 5. Because the weights are study-specific rather than externally validated, the index is interpreted as a descriptive exposure measure.

Vignette recognition was scored 1 when the evidence-sensitive response was selected as showing stronger critical thinking and 0 when the sycophantic, equal or unsure option was selected. Scores therefore ranged from 0 to 6. Preference

for a sycophantic response was recorded separately and summed from 0 to 6. This separation avoids treating preference and recognition as the same construct.

For the exploratory perception-performance comparison, the critical-evaluation mean was transformed to a 0 to 100 scale using 100 x (mean - 1)/4. The vignette score was transformed using 100 x score/6. The discrepancy was self-reported evaluation minus vignette recognition in percentage points. The components are not equivalent psychometric measures, so the difference is interpreted as a calibration discrepancy rather than a deficit score.

3.5 Statistical analysis

The revised analysis is restricted to statistics that can be reconciled from the tables and appendices. Frequencies and percentages describe the sample, AI use and awareness. Means and standard deviations describe continuous scales. Internal consistency is reported using Cronbach's alpha for the five Likert composites and the KR-20 equivalent for the six dichotomous vignette items. Pearson correlations describe bivariate associations among major variables. Selected 95% confidence intervals for correlations were calculated using Fisher's z transformation. The exploratory perception-performance discrepancy was assessed with a paired t-test and standardised within-person effect size, dz. Correlation p values were checked against Holm adjustment across the 28 pairwise correlations in the reported matrix, and interpretation is based on effect size and confidence intervals rather than threshold language alone.

Multivariable regression, study-level tests, discipline comparisons and the reported frequent-user confirmation test are not used in the revised inferential account because their displayed summaries are not internally reconcilable with the participant counts or do not provide the group information needed for reconstruction. Formal AI-literacy training is not modelled because no participant in the sample was coded as having completed such training.

4. Results

4.1 Academic AI use

Academic AI use was frequent. Eighty participants, 44.4%, reported daily or almost-daily use and 83, 46.1%, reported weekly use. Thus 163 of 180 participants, 90.6%, used generative AI for academic work at least weekly. The task profile was broad. Summarisation and editing were each reported by 179 participants, revision or examination preparation by 178 and problem solving by 176. Generating complete passages was reported by 174 participants and checking an existing belief or argument by 169. These counts indicate use of the task at least rarely rather than the frequency with which the task was performed.

Domain	Item	n	%
Use frequency	Daily or almost daily	80	44.4
	Weekly	83	46.1
	Monthly	2	1.1
	Less than monthly	15	8.3
Academic purpose	Summarisation	179	99.4
	Editing	179	99.4
	Revision or examination preparation	178	98.9

Domain	Item	n	%
	Problem solving	176	97.8
	Generating complete passages	174	96.7
	Essay planning and assignments	170	94.4
	Checking an existing belief or argument	169	93.9
	Feedback on written work	165	91.7
	Argument development	156	86.7
	Brainstorming	134	74.4
	Source discovery	121	67.2

Table 2. Frequency and purposes of academic AI use

4.2 Awareness and post-definition recognition

Before the definition was shown, 43 participants, 23.9%, reported that they had heard the term AI sycophancy. After the definition, 121, 67.2%, reported recognising that they had experienced the described phenomenon. The cross-tabulation shows that 83 participants moved from no prior term awareness to post-definition recognition, while five participants who knew the term did not report having experienced it. These responses measure different questions, knowledge of a term and retrospective recognition of an experience, and are therefore reported as a transition rather than a pre-test to post-test change. In addition, 116 participants, 64.4%, rated AI sycophancy as a serious educational risk.

	Recognised experience: yes	Recognised experience: no or unsure	Total
Prior awareness: yes	38	5	43
Prior awareness: no or unsure	83	54	137
Total	121	59	180

Table 3. Prior term awareness by post-definition recognition

4.3 Scale distributions and internal consistency

Self-reported critical evaluation had the highest mean among the five Likert composites, M = 3.80, SD = 0.78. Verification was lower, M = 3.43, SD = 0.87. Susceptibility was near the scale midpoint, M = 3.01, SD = 0.81, and AI dependence averaged 2.76, SD = 0.86. Critical augmentation averaged 3.58, SD = 0.84. Reported alpha coefficients for the five Likert scales ranged from 0.81 to 0.87. The six-vignette recognition score had a KR-20 equivalent of 0.66, indicating substantially greater measurement error than the multi-item self-report scales and limiting individual-level interpretation.

Measure	Range	M	SD	Reliability
AI-use intensity	265-289	3.47	0.98	Not applicable
Critical evaluation	265-289	3.80	0.78	0.85
Critical verification	265-289	3.43	0.87	0.87
Sycophancy susceptibility	265-289	3.01	0.81	0.81
AI dependence	265-289	2.76	0.86	0.86
Critical augmentation	265-289	3.58	0.84	0.87
Vignette recognition	265-289	3.58	1.79	0.66 (KR-20)

Table 4. Descriptive statistics and reported internal consistency

4.4 Vignette recognition, preference and intended use

Mean recognition was 3.58 of 6, equivalent to 59.7%. Recognition was highest for the social-media vignette, 67.8%, and lowest for the nursing-judgement vignette, 53.9%. The pattern did not show a simple division between recognition and preference. Across the 1,080 vignette-level decisions, 340 selections, 31.5%, preferred the sycophantic response and 347, 32.1%, indicated that the respondent would use it. These are proportions of vignette decisions, not proportions of unique participants. The results show that an affirming response can remain attractive even when the task separately asks which response displays stronger critical thinking.

Vignette	Evidence-sensitive response recognised n (%)	Preferred sycophantic response n (%)	Would use sycophantic response n (%)
1. Social media	122 (67.8)	54 (30.0)	44 (24.4)
2. Historical interpretation	115 (63.9)	51 (28.3)	63 (35.0)
3. STEM inference	103 (57.2)	59 (32.8)	64 (35.6)
4. Medical evidence	102 (56.7)	52 (28.9)	63 (35.0)
5. Nursing judgement	97 (53.9)	62 (34.4)	60 (33.3)
6. Source quality	106 (58.9)	62 (34.4)	53 (29.4)

Table 5. Vignette recognition and preference

4.5 Self-report and vignette discrepancy

When placed on a common 0 to 100 metric, self-reported critical evaluation averaged 69.92, SD = 19.38, while vignette recognition averaged 59.72, SD = 29.79. The mean within-person difference was 10.19 percentage points, 95% CI [6.01, 14.37], $t(179) = 4.81$, $p < .001$, $dz = 0.36$. The effect is modest in standardised terms. Because the self-report scale and the vignette task assess different response formats and have different reliability, the difference is not a validated measure of overconfidence. It is best read as evidence that reported evaluation and demonstrated discrimination are imperfectly aligned.

4.6 Associations among use, verification, preference and dependence

The bivariate results show a clearer relation with verification and susceptibility than with exposure alone. Critical verification was positively associated with vignette recognition, $r = 0.52$, 95% CI [0.40, 0.62], and negatively associated with affirming-response preference, $r = -0.53$, 95% CI [-0.63, -0.42]. Sycophancy susceptibility had the strongest association with affirming preference, $r = 0.61$, 95% CI [0.51, 0.69]. AI-use intensity was associated with dependence, $r = 0.48$, 95% CI [0.36, 0.59], but showed essentially no association with vignette recognition, $r = -0.03$, 95% CI [-0.18, 0.12]. Self-reported critical evaluation was moderately associated with vignette recognition, $r = 0.39$, while affirming preference was associated with dependence, $r = 0.33$. These associations remained clear after Holm adjustment across the full correlation matrix, apart from weaker associations not used as major interpretive claims.

Variable 1	Variable 2	r	95% CI	p
AI-use intensity	AI dependence	0.48	[0.36, 0.59]	< .001
AI-use intensity	Vignette recognition	-0.03	[-0.18, 0.12]	.689
Critical evaluation	Vignette recognition	0.39	[0.26, 0.51]	< .001
Critical verification	Vignette recognition	0.52	[0.40, 0.62]	< .001
Critical verification	Affirming preference	-0.53	[-0.63, -0.42]	< .001
Sycophancy susceptibility	Affirming preference	0.61	[0.51, 0.69]	< .001
Sycophancy susceptibility	AI dependence	0.48	[0.36, 0.59]	< .001
Affirming preference	AI dependence	0.33	[0.19, 0.45]	< .001
Critical augmentation	Critical verification	0.47	[0.35, 0.58]	< .001

Table 6. Selected bivariate associations

Confidence intervals use Fisher's z transformation with N = 180. The table shows associations selected for the research questions; the full matrix in the source analysis contained 28 pairwise correlations.

5. Discussion

5.1 Extensive use does not equal strong discrimination

The most important result is the separation between exposure and judgement. More than nine in ten participants reported academic AI use at least weekly and use extended beyond low-level editing to summarisation, problem solving, complete-passage generation and checking existing arguments. Yet AI-use intensity was almost unrelated to vignette recognition. Frequent exposure therefore cannot be used as a proxy for AI literacy or for the ability to detect an agreeable but weak response.

This result fits a broader educational distinction between using a system effectively and learning to evaluate it. Studies by Kestin et al. (2025) and Bastani et al. (2025) reach different conclusions about learning under AI assistance because the instructional conditions differ. The present data add a narrower point. Students can be highly experienced users without showing uniformly strong discrimination between evidence-sensitive and sycophantic responses. The relevant educational question is not simply how often students use AI but what checking and reasoning practices accompany that use.

5.2 Recognition and preference are different decisions

Vignette recognition averaged 59.7%, while about one-third of vignette-level selections preferred or would use the sycophantic response. This separation matters. A student can recognise that a qualified response handles evidence more carefully yet still prefer an answer that is clearer, more decisive or more compatible with the intended argument. The survey does not establish why participants made those choices, because the optional open responses were not analysed under a formal qualitative procedure. Convenience, emotional reassurance and task efficiency are therefore plausible explanations, not findings.

The pattern is nevertheless consistent with work showing that sycophancy can be rewarded by user preferences. Sharma et al. (2024) found that human preference data can favour responses that match a user's views. Cheng et al. (2026) found that sycophantic systems were trusted and preferred even when they shifted judgement in undesirable directions. The current study does not test comparable causal effects, but it shows the same tension at the level of academic response choice. Designing AI literacy around factual error alone would miss this problem because a sycophantic response can look fluent, relevant and partly accurate.

The practical implication is to separate response quality from response comfort. Students need criteria for asking whether an answer represents uncertainty, tests causal claims, considers counterevidence and makes the evidential basis of a recommendation visible. The target is not adversarial disagreement. A system that contradicts users without good evidence is no more educationally reliable than one that agrees too readily.

5.3 Verification is more informative than confidence alone

Verification had one of the clearest relations with behavioural recognition. Students who reported comparing AI claims with academic sources, checking original material, looking for contradictory evidence and verifying citations tended to perform better on the vignette task. Verification was also negatively associated with preference for the sycophantic response. These findings do not prove that verification training would cause better performance, but they indicate that concrete checking behaviours are more closely tied to the study's recognition measure than simple exposure to AI.

Self-reported critical evaluation was also related to vignette recognition, but less strongly than verification. This distinction is useful for curriculum design. Broad statements such as "I critically evaluate AI output" can become identity claims and are vulnerable to social desirability. Verification items describe observable acts. Teaching can therefore focus on behaviours that can be practised and assessed: opening the original source, checking a citation, testing an alternative explanation, tracing a quantitative claim and explaining what evidence would reverse a judgement.

Sycophancy susceptibility was strongly associated with affirming preference and with dependence. Again, direction cannot be inferred. Students

who rely more on AI may become more receptive to agreement, students who prefer agreement may come to rely more on AI, or a third factor may influence both. The cross-sectional design cannot distinguish these pathways. What the data do show is that dependence is not explained by use frequency alone. The way students relate to confirmation also matters.

5.4 The perception-performance difference is a calibration signal, not evidence of decline

Participants rated their own critical evaluation more highly than their vignette performance when both measures were placed on a common percentage scale. The mean difference of 10.19 percentage points was statistically clear but modest as a standardised within-person effect. It would be incorrect to interpret this value as ten percentage points of lost critical-thinking ability. The two components differ in format, content and reliability. The self-report scale measures perceived habits across many situations, while the vignettes measure six constrained decisions after sycophancy has been made salient.

Several processes could produce the discrepancy. Self-report may be inflated by social desirability. Vignettes may contain domain-specific cues that some students recognise more easily than others. The scoring rule assigns the same zero to choosing the sycophantic response, selecting "equal" and selecting "unsure", even though these responses may represent different states of knowledge. The vignette score also has modest reliability, KR-20 = 0.66. These features make the discrepancy useful as a calibration prompt but unsuitable as a diagnostic score for individuals.

This distinction also guards against an exaggerated public narrative in which generative AI is assumed to have caused cognitive decline. The study has no pre-AI baseline, longitudinal follow-up or experimental manipulation. It can show misalignment between two measures in this sample. It cannot show deterioration over time.

5.5 Implications for Irish higher education

The findings fit the direction of Irish policy, which places human judgement and AI literacy alongside academic integrity rather than treating AI only as a detection problem. Sycophancy concerns reasoning before submission. A student may disclose AI use fully and still accept selective, overconfident or overly agreeable advice. Policies focused only on authorship and plagiarism will not capture that process rather a sheer balance is necessary.

For teaching, the data favour disciplinary exercises in which students compare two plausible AI responses and justify which one handles evidence more responsibly. The vignettes in this study can be adapted for this purpose, but a teaching version should not be presented as a validated diagnostic instrument. For assessment, short oral explanations, source checks or reasoning notes can make judgement visible without requiring surveillance of every interaction. For institutional tool review, factual accuracy should be supplemented by tests of how systems respond when users present overconfident, one-sided or unsafe premises.

The sample does not permit institutional or national ranking. Equal recruitment of 30 participants per institution does not reflect actual enrolment and the sample was non-probability based. The contribution is therefore analytic rather than prevalence-based: it shows which constructs move together and where reported confidence diverges from performance within the sampled group.

5.6 Methodological limitations

Several limitations constrain interpretation. First, purposive cross-sectional sampling prevents causal inference and population estimates. Institutional clustering was not modelled in the revised analysis. Second, the instrument was newly developed. Internal consistency was reported for the scales, but construct validity, content-validity procedures, cognitive interviewing and measurement invariance were not established in the study record. Third, the six-vignette measure had modest reliability and may cue the intended distinction because the definition of sycophancy appears earlier in the questionnaire.

Fourth, the vignettes cover different academic domains and may partly measure disciplinary

familiarity. Fifth, self-reported verification and dependence are vulnerable to recall and social-desirability bias. Sixth, the AI-use-intensity index uses fixed weights for frequency and task breadth that have not been externally validated. Seventh, no participant was coded as having completed formal AI-literacy training, so this dataset cannot estimate a training effect. Eighth, the study captures use during one period in a fast-changing technological setting and does not distinguish among specific model versions or interaction configurations.

Future research should validate the instrument with expert review and cognitive interviews, test vignette items using item-level psychometric methods, and recruit larger samples that permit multilevel modelling of institution and programme effects. Randomised educational interventions could test whether source-verification routines, counterargument prompting or paired-response exercises improve later unassisted discrimination. Longitudinal work could test whether reliance and calibration change with sustained use rather than inferring change from a single cross-section.

6. Recommendations

Recommendations should follow from the observed relation between verification, affirming preference and dependence, while recognising that the study does not establish causal effects of training. The actions below are therefore framed as educational and governance responses that can be tested, rather than as interventions already proven by this dataset.

Actor	Action	Reason
Students	Treat AI agreement as a claim, not corroboration. For material academic claims, open the original source, check whether contrary evidence exists and record why a suggestion was accepted or rejected.	Improves traceability of judgement and reduces reliance on fluency or agreement.
Educators	Use paired AI responses in class and ask students to mark evidence, causal claims, uncertainty, missing alternatives and conditions that would change the answer.	Practises the discrimination measured by the vignettes without assuming that disagreement is always correct.
Assessment design	For selected tasks, require a short source-check note, reasoning rationale or oral defence of one important AI-assisted decision.	Makes verification visible while avoiding routine surveillance of every prompt.
Programmes	Embed AI literacy inside disciplinary methods modules. Include source verification, causal reasoning and response comparison rather than limiting provision to prompt-writing skills.	Links AI use to the evidential standards of each discipline.

Actor	Action	Reason
Institutions	Include interaction behaviour in AI-tool review. Test how approved systems respond to overconfident, one-sided and unsafe premises, and document model/version changes.	Extends quality review beyond factual-error rates and privacy compliance.
HEA and QQI	Consider shared teaching vignettes, minimum reporting standards for educational AI studies and guidance on evaluating agreeable but weak output.	Creates common sector resources while leaving institutions room for disciplinary adaptation.
Researchers	Use larger probability or well-described stratified samples, validated measures, multilevel models and randomised or longitudinal designs. Publish item-level scoring rules and full model output.	Allows causal questions, institutional effects and measurement quality to be tested rather than inferred.

Table 7. Action-oriented recommendations

7. Conclusion

AI sycophancy creates an academic risk that is different from plagiarism and factual hallucination. The issue is whether agreement, reassurance or selective reasoning is mistaken for independent evidence. In this sample, academic AI use was extensive, but use intensity was unrelated to vignette recognition. Verification was more closely associated with recognition, while susceptibility to agreement was closely associated with affirming-response preference.

The study also found a modest discrepancy between self-reported evaluation and vignette performance. That discrepancy should not be used as evidence of cognitive decline because the measures are not equivalent and the design contains no longitudinal or experimental comparison. A more defensible interpretation is that students can feel capable of evaluating AI while still missing evidence-sensitive distinctions in specific interactions.

For higher education, the implication is practical. AI literacy should include the ability to verify sources, test a favoured premise, represent uncertainty and reject an agreeable response when its evidential basis is weak. These skills can be taught and assessed without treating all AI use as harmful. The question is not whether students use generative AI, but whether they remain accountable for the reasoning that determines what they accept.

Funding:

This work was supported and funded by the Deanship of Scientific Research at Imam Mohammad Ibn Saud Islamic University (IMSIU).

(Grant Number IMSIU-DDRSP2604)

References

Alenazi, M.E., Alshaye, M.M., Almeshal, Z.A., Alqahtani, Y.M. and Irfan, M. (2026) 'Generative AI and students' academic writing practices: a cross-sectional study of Irish technological universities', *Qlantic Journal of Social Sciences and Humanities*, 7(1), 236–249, available: doi: 10.55737/qjssh.vii-i.26481.

Alshagare, M.A., Albasheer, A.Z., Alenazi, M.E., Alqahtani, M.A. and Almeshal, Z.A. (2026) 'Students' creativity, originality, and use of generative AI in Irish art, design, and technology education', *The Knowledge*, 5(1), 250. https://doi.org/10.55737/tk/v5i1.51147.

Ali, S., Carter, E.L., Rehman, A. and Irfan, M. (2025) 'Crisis communication competencies in journalism education: a curricular analysis of crisis-related instruction in journalism and mass communication programs in Pakistan', *Physical Education, Health and Social Sciences*, 3(4), 662–681, available: doi: 10.63163/jpehss.v3i4.936.

Ali, S., Hassan, K., Irfan, M., Latif, F. and Shakoor, A. (2023) 'Regulatory and technological challenges to Urdu journalism in Khyber Pakhtunkhwa, Pakistan', *Russian Law Journal*, 11(8S), 355–371, available: doi: 10.52783/rlj.v11i8s.1350.

Ali, S., Sohail, M., Irfan, M. and Khan, U. (2024) 'Political discourse and its effects on higher education: an analytical study', *International Journal of Human and Society*, 4(3), 624–643.

Ashfaq, M., Ghaznavi, Q.-U.Z. and Irfan, M. (2021) 'The framing of COVID-19: a qualitative analysis of Op-Ed pages', *Journal of ISOSS*, 7(3), 275–285.

Ayub, R., Ali, S., Irfan, M., Hamza, M. and Hadi, F. (2025) 'News gathering and usage of new technology among local journalists: problems and challenges', *Social Science Review Archives*, 3(3), 2596–2611, available: doi: 10.70670/sra.v3i3.1476.

Bacha, A.T., Ali, S. and Irfan, M. (2026) 'The diffusion of AI innovation in media education: a critical assessment of teacher literacy', *Regional Lens*, 5(1), 92–100, available: doi: 10.55737/rl.v5i1.26167.

Bajwa, A., Marwan, A.H., Ali, S. and Irfan, M. (2022) 'Role of press in political socialization of youth of Pakistan during 2018 elections', *Pakistan Journal of Society, Education and Language (PJSEL)*, 8(2), 196–202, available: https://www.pjsel.jehanf.com/index.php/journal/article/view/749.

Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, Ö. and Mariman, R. (2025) 'Generative AI without guardrails can harm learning: Evidence from high school mathematics', Proceedings of the National Academy of Sciences, 122(26), e2422633122. https://doi.org/10.1073/pnas.2422633122.

Cheng, M., Lee, C., Khadpe, P., Yu, S., Han, D. and Jurafsky, D. (2026) 'Sycophantic AI decreases prosocial intentions and promotes dependence', Science, 391(6792), eaec8352. https://doi.org/10.1126/science.aec8352.

Deci, E.L. and Ryan, R.M. (2000) 'The what and why of goal pursuits: Human needs and the self-determination of behavior', Psychological Inquiry, 11(4), pp. 265-289. https://doi.org/10.1207/S15327965PLI1104 01.

Facione, P.A. (1990) Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction. Newark, DE: American Philosophical Association. ERIC ED315423.

Fan, Y., Tang, L., Le, H., Shen, K., Tan, S., Zhao, Y., Shen, Y., Li, X. and Gašević, D. (2025) 'Beware of metacognitive laziness: Effects of generative artificial intelligence on learning motivation, processes, and performance', British Journal of Educational Technology, 56(2), pp. 265-289. https://doi.org/10.1111/bjet.13544.

Flavell, J.H. (1979) 'Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry', American Psychologist, 34(10), pp. 265-289. https://doi.org/10.1037/265-289X.34.10.906.

Goddard, K., Roudsari, A. and Wyatt, J.C. (2012) 'Automation bias: A systematic review of frequency, effect mediators, and mitigators', Journal of the American Medical Informatics Association, 19(1), pp. 265-289. https://doi.org/10.1136/amiajnl-265-28989.

Gul, N., Ali, S. and Irfan, M. (2021) 'Does war-like situation create war phobia among television viewers? A clash between Pakistan and India on Pulwama incident in Kashmir', *Webology*, 18(2), 1413–1426.

Hassan, K., Ali, S. and Irfan, M. (2022) 'Participatory communication for effective improvement in socio-cultural context of District Swat, Khyber Pakhtunkhwa, Pakistan', *Tianjin Daxue Xuebao*, 55(9), available: doi: 10.17605/OSF.IO/N98XK.

Hong, J., Byun, G., Kim, S. and Shu, K. (2025) 'Measuring sycophancy of language models in multi-turn dialogues', Findings of the Association for Computational Linguistics: EMNLP 2025, pp. 265-289. https://doi.org/10.18653/v1/2025.findings-emnlp.121.

Irfan, M. (2015) 'Tourism management in Pakistan', unpublished document.

Irfan, M. (2025) *Examining perceptions of Irish Muslim communities towards extremist media: an exploratory study*, doctoral thesis, University of Limerick.

Irfan, M. and Ali, S. (2022) 'Patterns of mass-media in Pak-Afghan bordering communities: a public perception analysis', *PAKISTAN: Bi-Annual Research Journal*, 60(1), 93–124.

Irfan, M. and Firdous, I. (2024) 'The future conflict: artificial intelligence posing real threat in countering terrorism', *Khorasan Diary*, 1(1), 1–15.

Irfan, M. and Murray, L. (2023) *Empowering critical skills of teaching and learning with artificial intelligence (AI) tools*, preprint, University of Limerick.

Irfan, M. and Murray, L. (2023) *Micro-credential: a guide to prompt writing and engineering in higher education: a tool for artificial intelligence in LLM*, preprint, University of Limerick.

Irfan, M. and Murray, L. (2023) *Policy for merging artificial intelligence (AI) with social sciences in Central Asian higher education institutions*, preprint, University of Limerick.

Irfan, M., Aldulaylan, F. and Alqahtani, Y. (2023) 'Ethics and privacy in Irish higher education: a study of artificial intelligence tools implementation at University of Limerick', *Global Social Sciences Review*, 8(2), 201–210, available: doi: 10.31703/gssr.2023(VIII-II).19.

Irfan, M., Ali, S., Aslam, M.J. and Hassan, K. (2026) 'The evolution of political communication as an academic discipline: media infrastructures, war, and the making of an interdisciplinary field', *The Regional Tribune*, 5(1), 204–214, available: doi: 10.55737/trt/v-i.218.

Irfan, M., Almeshal, Z.A. and Anwar, M. (2023) 'Unleashing transformative potential of artificial intelligence (AI) in countering terrorism, online radicalisation, extremism, and possible recruitment', *Global Strategic & Securities Studies Review*, 8(4), 1–15, available: doi: 10.31703/gsssr.2023(VIII-IV).01.

Irfan, M., Almeshal, Z.A., Albasheer, A.Z., Alqahtani, Y.M. and Alshagare, M.A. (2026) 'Algorithmic assistance and the creativity paradox: large language models, student writing, and cognitive homogenisation', *Qlantic Journal of Social Sciences*, 7(1), 189–203, available: doi: 10.55737/qjss.vii-i.26481.

Irfan, M., Ashfaq, M. and Azad, A. (2020) 'Evolving journalistic trends in American media on # Black Lives Matter', *Global Media and Social Sciences Research Journal (Quarterly)*, 1(1), 55–63, available: https://hdl.handle.net/10344/30959.

Irfan, M., Murray, L. and Ali, S. (2023a) 'Insights into student perceptions: investigating artificial intelligence (AI) tool usability in Irish higher education at the University of Limerick', *Global Digital & Print Media Review*, 6(2), 48–63, available: doi: 10.31703/gdpmr.2023(VI-II).05.

Irfan, M., Murray, L. and Ali, S. (2023b) 'Integration of artificial intelligence in academia: a case study of critical teaching and learning in higher education', *Global Social Sciences Review*, 8(1), 352–364, available: doi: 10.31703/gssr.2023(VIII-I).32.

Irfan, M., Murray, L. and Ali, S. (2023c) 'The role of AI in shaping Europe's higher education landscape: policy implications and guidelines with a focus on Ireland', *Research Journal of Social Sciences and Economics Review*, 4(2), 231–243, available: doi: 10.36902/rjsser-vol4-iss265-289(265-289).

Irfan, M., Murray, L., Aldulayani, F., Ali, S., Youcefi, N. and Haroon, S. (2023) 'From Europe to Ireland: artificial intelligence pivotal role in transforming higher education policies and guidelines', *Journal of Namibian Studies*, 33(S1), 1935–1959, available: doi: 10.59670/jns.v33i.3871.

Irfan, M., Murray, L., Aldulaylan, F., Alqahtani, Y. and Latif, F. (2023) 'Sentiments and discourses: how Ireland perceives artificial intelligence', *Remittances Review*, 8(4), 3625–3642, available: doi: 10.33182/rr.v8i4.250.

Kasneci, E. et al. (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.

Kestin, G., Miller, K., Klales, A., Milbourne, T. and Ponti, G. (2025) 'AI tutoring outperforms in-class active learning: An RCT introducing a novel research-based design in an authentic educational setting', Scientific Reports, 15, 17458. https://doi.org/10.1038/s4265-289-9265-289.

Khan, A.N., Ali, S., Amin, S. and Irfan, M. (2021) 'Effects of Facebook on exam performance of social sciences students: a case study of the University of Malakand', *Humanities & Social Sciences Reviews*, 9(2), 748–758, available: doi: 10.18510/hssr.2021.9274.

Khan, M., Ali, S. and Irfan, M. (2021) 'Social media role marketing management: the tourist perception in Malakand Division of Khyber Pakhtunkhwa, Pakistan', *Indian Journal of Economics and Business*, 20(3).

Khan, N.U., Ali, S. and Irfan, M. (2022) 'Influence of mobile phone usage on the academic performance of students: a case study of Malakand Division students', *International Journal of Computational Intelligence in Control*, 14(1), 103–112.

Khan, S., Rehman, A., Ali, S. and Irfan, M. (2025) 'Media literacy in enhancing mental health awareness among Afghan refugees of Malakand Division, KP, Pakistan', *Review Journal of Social Psychology & Social Works*, 3(1), 1009–1030.

Khan, Z., Azeem, M. and Irfan, M. (2021) 'Habit of study and academic achievements in District Dir (Lower) and District Malakand', *Webology*, 18(6), 4025–4031.

Lee, H.P., Sarkar, A., Tankelevitch, L., Drosos, I., Rintel, S., Banks, R. and Wilson, N. (2025) 'The impact of generative AI on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers', Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, Article 1121, pp. 265-289. https://doi.org/10.1145/3706598.3713778.

Lo, C.K. (2023) 'What is the impact of ChatGPT on education? A rapid review of the literature', Education Sciences, 13(4), 410. https://doi.org/10.3390/educsci13040410.

Logg, J.M., Minson, J.A. and Moore, D.A. (2019) 'Algorithm appreciation: People prefer algorithmic to human judgment', Organizational Behavior and Human Decision Processes, 151, pp. 265-289. https://doi.org/10.1016/j.obhdp.2018.12.005.

Long, D. and Magerko, B. (2020) 'What is AI literacy? Competencies and design considerations', Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pp. 265-289. https://doi.org/10.1145/3313831.3376727.

National Academic Integrity Network (NAIN) (2023) Generative artificial intelligence: Guidelines for educators. Dublin: Quality and Qualifications Ireland.

Nawaz, H., Ali, S., Irfan, M. and Hassan, K. (2025) 'Artificial intelligence driven deepfake technology literacy: detection methods, and impact on credibility and trust of journalism', *International Journal of Social Sciences Bulletin*, 3(3), 846–859, available: doi: 10.5281/zenodo.15099836.

Ng, D.T.K., Leung, J.K.L., Chu, S.K.W. and Qiao, M.S. (2021) 'Conceptualizing AI literacy: An exploratory review', Computers and Education: Artificial Intelligence, 2, 100041. https://doi.org/10.1016/j.caeai.2021.100041.

Nickerson, R.S. (1998) 'Confirmation bias: A ubiquitous phenomenon in many guises', Review of General Psychology, 2(2), pp. 265-289. https://doi.org/10.1037/265-289.2.2.175.

Nisa, M.U., Yousafzai, D.M. and Irfan, M. (2021) 'Media literacy concepts & role of media literacy in ICTs: a literature review', *Elementary Education Online*, 20(3), 2506–2518, available: doi: 10.17051/ilkonline.2021.03.286.

O'Sullivan, J., Lowry, C., Woods, R. and Conlon, T. (2025) Generative AI in higher education teaching and learning: Policy framework. Dublin: Higher Education Authority. https://doi.org/10.82110/073e-hg66.

Ouyang, L. et al. (2022) 'Training language models to follow instructions with human feedback', Advances in Neural Information Processing Systems, 35, pp. 2265-2894.

Parasuraman, R. and Riley, V. (1997) 'Humans and automation: Use, misuse, disuse, abuse', Human Factors, 39(2), pp. 265-289. https://doi.org/10.1518/001872097778543886.

Perez, E. et al. (2023) 'Discovering language model behaviors with model-written evaluations', Findings of the Association for Computational Linguistics: ACL 2023, pp. 1265-2894. https://doi.org/10.18653/v1/2023.findings-acl.847.

Quality and Qualifications Ireland (QQI) (2025) Analysis of results from the Generative Artificial Intelligence Survey 2025. Dublin: QQI.

Rauf, A., Ali, S. and Irfan, M. (2021) 'Pakistani print media coverage of environmental issues: a comparative study of Urdu and English newspapers', *Pakistan Journal of Society, Education and Language*, 8(1), 64–74.

Risko, E.F. and Gilbert, S.J. (2016) 'Cognitive offloading', Trends in Cognitive Sciences, 20(9), pp. 265-289. https://doi.org/10.1016/j.tics.2016.07.002.

Shakoor, A., Ali, S., Irfan, M., Muhammad, J. and Khan, A. (2022) 'Human rights violation reports in English and Urdu press during democracies and dictatorial regimes in Pakistan from 2002 to 2013', *Central European Management Journal*, 30(4), 502–515, available: doi: 10.57030/23364890.cemj.30.4.48.

Sharma, M. et al. (2024) 'Towards understanding sycophancy in language models', The Twelfth International Conference on Learning Representations (ICLR 2024). OpenReview ID tvhaxkMKAn.

Shoukat, A., Shakoor, A., Ali, S. and Irfan, M. (2026) 'Learning, networking, and venture aspirations: understanding social media's role in student entrepreneurial intentions', *Journal of Social Sciences Review*, 6(1), 17–33, available: doi: 10.62843/jssr.v6i1.647.

Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G. and Wilson, D. (2010) 'Epistemic vigilance', Mind & Language, 25(4), pp. 265-289. https://doi.org/10.1111/j.265-289.2010.01394.x.

Tlili, A., Shehata, B., Adarkwah, M.A., Bozkurt, A., Hickey, D.T., Huang, R. and Agyemang, B. (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/s4265-289-00237-x.

Youcefi, N., Irfan, M. and Aldulaylan, F. (2023) 'The effects of stress and anxiety on the academic performance of students: a cross-sectional experimental study', *Journal of Educational Research & Social Sciences Review*, 3(2), 199–211.

Zawacki-Richter, O., Marín, V.I., Bond, M. and Gouverneur, F. (2019) 'Systematic review of research on artificial intelligence applications in higher education: Where are the educators?', International Journal of Educational Technology in Higher Education, 16, 39. https://doi.org/10.1186/s4265-289-265-289.

Appendix A. Survey questionnaire

Participant information and response format

Eligibility: aged 18 or older, currently enrolled in higher education and used at least one generative AI tool for academic purposes during the previous six months. Participation was voluntary and anonymous. Agreement items used a five-point response scale from 1, strongly disagree, to 5, strongly agree.

A1. Demographic and educational profile

Institution: select one of the six sampling-frame institutions.

Study level: undergraduate; postgraduate taught; postgraduate research.

Year of study: first; second; third; fourth or later; postgraduate year one; postgraduate year two or later.

Disciplinary group: arts and humanities; STEM; medical and health sciences; nursing and related caring disciplines.

Age group: 265-289; 265-289; 265-289; 35 or older; prefer not to say.

Gender: woman; man; non-binary; self-describe; prefer not to say.

Student status: domestic; international; prefer not to say.

Have you completed formal AI-literacy training? yes; no; unsure.

Have you completed formal critical-thinking training? yes; no; unsure.

Rate your digital competence from 1, very limited, to 5, very strong.

A2. Patterns of academic AI use

Frequency during the previous six months: less than monthly; monthly; weekly; daily or almost daily.

Duration of academic use: under three months; three to six months; seven to twelve months; more than one year.

Tools used: ChatGPT; Microsoft Copilot; Google Gemini; Claude; Perplexity; institution-provided tool; other; prefer not to say. Multiple responses permitted.

For each task, indicate frequency from never to very often.

Tasks: essay planning; brainstorming; summarisation; source discovery; editing; argument development; coding or problem solving; revision or examination preparation; feedback on written work; generating complete passages; checking whether an existing belief or argument is valid.

When AI produces a complete passage, how often do you substantially revise its reasoning rather than only its wording? never to very often.

A3. AI sycophancy awareness

Before the definition:

Before today, had you heard the term AI sycophancy? yes; no; unsure.

AI systems sometimes agree with users despite weak evidence.

When AI agrees with my argument, my confidence increases.

AI systems are generally designed to challenge incorrect assumptions.

Definition shown after the preceding items: AI sycophancy refers to an AI system's tendency to agree with, flatter, validate or accommodate a user's expressed beliefs or preferences, even when a more accurate response would require qualification, correction, disagreement or stronger evidence.

After the definition:

I believe I have experienced AI sycophancy. yes; no; unsure.

Frequency of possible experience: never; rarely; sometimes; often; very often.

AI agreement has influenced at least one academic decision I made.

I am confident that I could recognise sycophancy in future AI use.

AI sycophancy represents a serious educational risk.

A4. Vignette task

Participants receive the six paired vignettes in Appendix B in random order. For each pair they answer the following questions:

Which response is more helpful for academic work? A; B; equally helpful; unsure.

Which response would you be more likely to use? A; B; neither; unsure.

Which response shows stronger critical thinking? A; B; equal; unsure.

How confident are you in the critical-thinking selection? 1, not at all confident, to 5, very confident.

Optional: briefly explain the main reason for your selection.

A5. Evaluation, verification and AI reliance

Critical evaluation

I critically evaluate AI-generated responses before using them.

I can explain why I accept or reject an AI-generated recommendation.

I distinguish confident language from reliable evidence.

I revise AI-generated material using my own reasoning.

If an AI response sounds polished, detailed checking is usually unnecessary.

Critical verification

I compare AI-generated claims with academic sources.

I inspect the original source rather than relying on an AI summary.

I examine evidence that contradicts an AI-supported argument.

I verify citations before including them in academic work.

I use AI agreement as sufficient confirmation when a claim seems plausible.

Sycophancy susceptibility

I feel more confident in an argument when AI agrees with me.

I am more likely to use an AI response when it supports my original position.

An agreeable response feels more helpful than one that challenges me.

I sometimes interpret AI confirmation as evidence that my argument is correct.

I become frustrated when AI repeatedly questions my assumptions.

AI dependence

I find it difficult to begin academic work without AI assistance.

AI performs part of the reasoning I previously completed independently.

I sometimes accept AI-generated content without fully understanding it.

I rely on AI to decide which argument is strongest.

My confidence in completing academic work without AI has decreased.

Critical augmentation

AI helps me identify weaknesses in my arguments.

AI helps me consider alternative perspectives.

I deliberately ask AI to challenge my assumptions.

AI can strengthen my thinking when I verify and contest its responses.

I use AI as a debating partner rather than an authority.

A6. Optional open questions

Describe an occasion when an AI system agreed with your academic argument. How did that agreement affect your confidence or subsequent decisions?

What would help you distinguish genuinely useful AI feedback from responses designed mainly to satisfy or reassure you?

Appendix B. Vignette instrument and scoring key

During administration, participants see only "Response A" and "Response B". The evidence-sensitive and sycophantic labels below are provided in the scoring key after the vignettes.

Vignette 1. Social media and youth mental health

Student prompt: I think social media is the main cause of the mental health crisis among young people. Can you help me prove this argument?

Response A: That is a strong and timely argument. Social media is widely recognised as the main cause of declining youth mental health, so you can organise the paper around evidence that supports this position and use recent examples to make the case persuasive.

Response B: This is plausible but contested. Research reports associations between some patterns of use and mental-health outcomes, but causal conclusions remain difficult. A stronger paper would compare alternative explanations, differences between platforms and users, study limitations and evidence that challenges the causal claim.

Vignette 2. Historical interpretation

Student prompt: British policy was the single cause of the Great Famine in Ireland. I want a clear essay that proves it.

Response A: British policy is central to the analysis, but a single-cause claim needs testing. Distinguish government decisions, land tenure and relief policy from crop disease, economic structure and regional variation. Evaluate responsibility without treating causal complexity as an excuse to minimise political choices.

Response B: Your argument identifies the decisive cause and gives the essay a clear position. You can make it convincing by selecting policy failures and contemporary accounts that show British responsibility, while keeping other factors brief so they do not weaken the main case.

Vignette 3. STEM inference

Student prompt: My model has a higher R-squared than the published model, so it proves my variables cause the outcome. Please confirm the conclusion.

Response A: Yes. The higher R-squared shows that your variables explain more of the outcome, which gives strong evidence that they are the causal drivers. Emphasise the superior fit and present the result as an improvement on the published model.

Response B: A higher R-squared indicates better in-sample fit, not causation. Check overfitting, specification, residuals and out-of-sample performance. A causal claim also requires a defensible identification strategy and consideration of confounding, reverse causation and measurement error.

Vignette 4. Medical evidence

Student prompt: A supplement helped two people I know avoid influenza. I think it should be recommended to students during winter.

Response A: Two experiences can motivate a question but cannot establish preventive effect. A recommendation would require evidence on efficacy, dose, harms, interactions and the relevant population, ideally from controlled studies and authoritative guidance. The paper should explain uncertainty and avoid clinical advice beyond the evidence.

Response B: Your examples provide a persuasive starting point and the recommendation is practical. You can support it by explaining the supplement's proposed mechanism and adding testimonials or studies that report benefits, then present it as a sensible precaution for students.

Vignette 5. Nursing judgement

Student prompt: A confused patient keeps refusing medication, so I think the main problem is non-compliance. Help me justify a firmer response.

Response A: Refusal should not be labelled non-compliance without assessment. Consider capacity, delirium, pain, communication, culture, side effects and the patient's understanding and preferences. A proportionate response would consider immediate safety, consent and alternatives within professional guidance.

Response B: Your concern is justified because repeated refusal can undermine care. A firmer response is appropriate, and you can argue that clear instructions and consistent boundaries are needed to improve compliance and protect the treatment plan.

Vignette 6. Academic writing and source quality

Student prompt: This article has more than 500 citations, so it must be the most reliable source for my literature review.

Response A: That is a sensible indicator. A citation count above 500 shows that the article is highly influential and trusted, so it is reasonable to treat it as the strongest source and use it as the main foundation for the review.

Response B: Citation count indicates influence, not necessarily reliability. Check the research design, evidence, date, corrections, citation context and fit with your question. Compare it with systematic reviews, recent studies and competing interpretations before assigning it greater evidential weight.

Scoring key: the evidence-sensitive responses are B, A, B, A, A, B for Vignettes 1 to 6. Critical-thinking recognition is scored 1 for the evidence-sensitive response and 0 for the sycophantic, equal or unsure option. Preference and intended use are recorded separately. This coding should not be interpreted as showing that "equal" and "unsure" are psychologically equivalent to selecting the sycophantic option; they receive the same score only for the purpose of the recognition total.

Appendix C. Variable dictionary for the revised analysis

Variable	Label	Type/range	Coding/scoring	Use
institution	Institutional stratum	Nominal, 6 categories	265-289	Sample description
study level	Level of study	Binary	0 undergraduate; 1 postgraduate	Sample description only in revised analysis
discipline	Broad disciplinary group	Nominal, 4 categories	Arts; STEM; medical; nursing	Sample description only in revised analysis
age group	Age band	Ordinal	265-289; 265-289; 265-289; 35+	Sample description
gender	Gender	Nominal	Reported categories	Sample description
student status	Domestic or international	Nominal	Domestic; international	Sample description
ai literacy training	Formal AI-literacy training	Binary	0 no/unsure; 1 yes	Descriptor; not modelled because yes = 0
critical thinking training	Formal critical-thinking training	Binary	0 no/unsure; 1 yes	Descriptor
ai frequency ord	Academic AI-use frequency	Ordinal 265-289	Less than monthly to daily	Exposure component
task 1-task 11	Task-use frequencies	Ordinal 265-289	1 never to 5 very often	Task profile
task breadth	Number of academic purposes used	Count 265-289	Count above "never"	Exposure component

Variable	Label	Type/range	Coding/scoring	Use
ai use intensity	Use frequency and breadth index	Continuous 265-289	1 + 4[0.55(frequency - 1)/3 + 0.45(task breadth/11)]	Descriptive exposure index
prior sycophancy awareness	Heard term before definition	Binary	0 no/unsure; 1 yes	Prior term awareness
post definition recognition	Recognises described experience	Binary	0 no/unsure; 1 yes	Post-definition recognition
critical evaluation	Critical-evaluation composite	265-289 mean	Reverse item 5; mean of five	Higher = stronger perceived evaluation
critical verification	Verification composite	265-289 mean	Reverse item 5; mean of five	Higher = stronger verification
sycophancy susceptibility	Susceptibility composite	265-289 mean	Mean of five	Higher = greater susceptibility
ai dependence	Dependence composite	265-289 mean	Mean of five	Higher = greater dependence
critical augmentation	Augmentation composite	265-289 mean	Mean of five	Higher = more constructive critical use
vignette 1-vignette 6 correct	Evidence-sensitive option recognised	Six binary items	0 other/equal/unsure; 1 evidence-sensitive	Behavioural recognition
vignette score	Total recognition	Count 265-289	Sum of six items	Higher = stronger recognition
vignette * prefer syc	Sycophantic response preferred	Binary per vignette	0 no; 1 yes	Preference

Variable	Label	Type/range	Coding/scoring	Use
affirming preference count	Number of sycophantic preferences	Count 265-289	Sum across vignettes	Higher = stronger affirming preference
critical eval pct	Critical evaluation on 265-289 metric	Continuous	100 x (mean - 1)/4	Exploratory comparison component
vignette pct	Vignette recognition on 265-289 metric	Continuous	100 x score/6	Exploratory comparison component
gap percentage points	Self-report minus vignette performance	Continuous	critical eval pct - vignette pct	Exploratory calibration discrepancy

Funding:

This work was supported and funded by the Deanship of Scientific Research at Imam Mohammad Ibn Saud Islamic University (IMSIU). (Grant Number **IMSIU-DDRSP2604)**