INTERNATIONAL JOURNAL OF ACADEMIC RESEARCH IN PROGRESSIVE EDUCATION & DEVELOPMENT

OPEN ACCESS JOURNAL

Vol 15, Issue 4, (2026) E-ISSN: 2226-6348

Self-Reported Generative AI Literacy among Undergraduates in Shandong, China: Dimensional Patterns and Educational Implications

Xinyue Kang

Department of Educational Studies, Faculty of Human Development, Universiti Pendidikan Sultan Idris (UPSI), Perak, Malaysia

Norazlin Mohd Rusdin

Department of Educational Studies, Faculty of Human Development, Universiti Pendidikan Sultan Idris (UPSI) Tanjong Malim, Perak, Malaysia

Corresponding Author Email: norazlin@fpm.upsi.edu.my

DOI Link: http://dx.doi.org/10.6007/IJARPED/v15-i4/29256

Published Online: 09 November 2026

Abstract

University responses to generative artificial intelligence (GenAI) require evidence about the specific capabilities students perceive themselves to possess. This cross-sectional study examined overall, domain, and item-level self-reports among 616 undergraduates recruited through 18 universities in Shandong, China. A contextually adapted 12-item measure assessed awareness, use, evaluation, and ethics. Descriptive statistics and response distributions were interpreted alongside internal-consistency and confirmatory factor analyses. Overall literacy averaged 3.495 on a five-point scale (SD = 0.650); domain means were 3.688 for ethics, 3.477 for awareness, 3.423 for evaluation, and 3.392 for use. Cronbach's alpha was .900 overall and .821–.874 across domains. Positive-direction responses accounted for 61.2%–64.9% on ethics items, 45.9%–48.1% on use items, and 45.5%–50.8% on evaluation items. Midpoint responses represented 37.5% for proficient academic use and 38.8% for evaluating capabilities and limitations. Item distributions thus qualified the positive overall mean by locating specific areas of more reserved self-appraisal. The findings support connecting practical application with evaluation and examining how ethical commitments translate into academic decisions. They provide contextual evidence for planning follow-up assessment, while voluntary recruitment and self-report measurement limit population and proficiency claims.

Keywords: Generative Artificial Intelligence, Ai Literacy, Undergraduate Students, Self-Report, Higher Education

Introduction

Generative artificial intelligence (GenAI) has brought the production of text, explanations, and ideas into everyday discussions of university learning. Following the public release of ChatGPT in November 2022, interest in its educational applications increased, including brainstorming, writing assistance, and feedback (Chan & Hu, 2023). These applications extend students' opportunities to obtain academic support, but they also require decisions about the accuracy,

suitability, and acceptable use of generated material. Consequently, the educational significance of GenAI concerns both what these systems can produce and how students understand and judge their contributions to learning (Miao & Holmes, 2023).

The development of AI literacy research provides a conceptual foundation for examining these demands. Before the recent expansion of conversational GenAI, Long and Magerko (2020) conceptualised AI literacy through competencies enabling critical evaluation, effective interaction, and informed engagement with AI. Ng et al. (2021) subsequently organised the field around knowing and understanding, using and applying, evaluating and creating, and ethical issues. Together, these perspectives position literacy as a combination of knowledge, practical engagement, judgement, and responsibility. They also provide a basis for distinguishing familiarity with an application from the broader capabilities required to use it meaningfully in an educational setting.

GenAI makes this multidimensional perspective particularly relevant because students can obtain a fluent response while still needing to decide whether its content deserves acceptance. In a learning task, operating a chatbot, evaluating its explanation, and deciding whether its use complies with academic expectations involve related but distinguishable judgements. UNESCO's guidance places human capacity and responsible educational integration at the centre of these concerns (Miao & Holmes, 2023), while Chan's (2023) policy framework connects institutional responses with teaching and assessment. These perspectives suggest that universities need evidence about specific literacy domains when planning support, rather than relying solely on students' access to tools or willingness to use them.

Empirical findings reinforce the distinction between positive engagement and comprehensive capability. In a survey of 399 undergraduate and postgraduate students in Hong Kong, Chan and Hu (2023) found generally positive perceptions of GenAI alongside concerns about accuracy, privacy, and ethical issues. Evidence from educational provision likewise indicates that initial orientation may address some needs more readily than others. Sullivan et al. (2024) reported increased confidence and better understanding of university policy after a 90-minute workshop, while identifying a continuing need to develop skills for assessing generated outputs. Read together, these studies support examining what students feel able to do within particular activities; favourable attitudes or increased confidence alone do not establish that every relevant capability is equally developed.

Recent research has extended this discussion towards contextual literacy assessment. O'Dea et al. (2026) examined knowledge, application, evaluation, and ethics among 234 university students in the United Kingdom and Hong Kong. Their study provides relevant evidence beyond general attitudes, although its questionnaire differs from the measure adopted here. Assessment research also makes the distinction between perceived and demonstrated capability increasingly consequential. Jin et al. (2025) developed the Generative AI Literacy Assessment Test and found that its scores predicted performance in GenAI-supported tasks more effectively than the self-reported measures examined. Thus, self-report surveys can document learners' appraisals and inform further assessment, but their results require interpretation within the boundaries of the questions asked.

The present study adopts the awareness–use–evaluation–ethics framework operationalised by Wang et al. (2023) in a 12-item AI literacy scale and contextually adapted here to GenAI in university learning. Awareness concerns recognising AI and its applications; use concerns practical engagement; evaluation concerns judging capabilities and suitability; and ethics concerns responsible conduct and associated risks. This framework offers a coherent organisation for descriptive inquiry without assuming that the four domains develop uniformly. An overall mean summarises the direction of students' appraisals, but it cannot reveal whether a particular statement attracts broad endorsement, substantial midpoint responding, or predominantly negative responses. Examining response composition therefore adds information relevant to selecting follow-up assessment questions.

Against this background, this study examines self-reported GenAI literacy among 616 undergraduates recruited through 18 universities in Shandong Province, China. It supplies contextual evidence from students across four academic years and three broad academic fields, without treating the retained volunteers as representative of the provincial undergraduate population. The analysis connects overall scores, four domain scores, and individual response distributions to identify where aggregate summaries conceal more reserved or differentiated appraisals. Its contribution is an educationally focused interpretation of perceived capabilities, rather than a direct estimate of proficiency or an explanation of causal mechanisms. This is relevant as universities move from introducing GenAI tools towards deciding which forms of application, evaluation, and responsible judgement need closer assessment (Chan, 2023; Sullivan et al., 2024).

The study addresses three research questions:
RQ1: What are the overall descriptive characteristics of self-reported GenAI literacy among the participating undergraduates?
RQ2: What descriptive patterns are evident across awareness, use, evaluation, and ethics?
RQ3: What item-level response patterns characterise students' perceived capabilities within these domains?

Literature Review

Multidimensional Literacy and Measurement Scope

AI literacy frameworks organise capabilities in different ways. Ng et al. (2021) identified understanding, application, evaluation and creation, and ethical considerations, while Ng et al. (2024) subsequently developed a questionnaire addressing affective, behavioural, cognitive, and ethical learning. These approaches share an interest in multidimensional engagement but differ in content and organisation. The choice of framework consequently determines which aspects of literacy a study can meaningfully describe.

Reviews reinforce the need to make this choice explicit. Laupichler et al. (2022) reviewed AI literacy in higher and adult education, while Almatrafi et al. (2024) examined conceptualisations, constructs, implementation, and assessment. Lintner (2024) focused specifically on literacy scales and their measurement evidence. Taken together, these reviews show that literacy research encompasses both conceptual and assessment decisions. An instrument's label alone does not establish that it measures the same construct as another scale. Carolus et al. (2023) illustrate this variation through the Meta AI Literacy Scale, which distinguishes AI literacy from related competencies such as self-efficacy and self-

management. This distinction supports specifying the boundaries of the present measure: its four domains describe perceived awareness, use, evaluation, and ethics, rather than a comprehensive assessment of students' confidence or psychological resources.

The present analysis adopts the awareness–use–evaluation–ethics organisation. Awareness concerns recognising AI applications and understanding their role. Use concerns applying them to learning activities. Evaluation concerns judging applications, capabilities, and outputs. Ethics concerns responsible conduct and attention to associated risks. Adapting these domains to GenAI retains a connection with general AI literacy while directing students' appraisals towards technologies encountered in academic learning. It does not establish exhaustive coverage of prompt design, factual verification, disclosure, or every other specialised GenAI skill.

Student Experiences and Evidence of Capability

Studies of students' experiences supply important context without substituting for competency assessment. Chan and Hu (2023) found that students recognised potential educational benefits while also expressing concerns. Almassaad et al. (2024) examined utilisation and perceptions through technology acceptance and task–technology fit perspectives. These studies address how students experience and appraise the technology. They do not make perceived usefulness equivalent to the capability to evaluate a particular output.

Research on GenAI literacy also examines contextual variation. O'Dea et al. (2026) analysed four dimensions using descriptive statistics and group comparisons. Their ethics questions concerned ethical considerations when using AI-provided data and recommendations. The current scale instead includes reported ethical adherence, privacy attention, and vigilance concerning misuse. These related but different formulations illustrate why shared domain names do not establish interchangeable scores. Contextual studies can complement one another by explaining their item coverage rather than treating mean rankings as universal benchmarks.

Other evidence concerns how literacy relates to surrounding variables and how it can be measured. Bewersdorff et al. (2025) synthesised associations with affective, behavioural, cognitive, and contextual variables. Hornberger et al. (2025a) examined a knowledge-based assessment across Germany, the United Kingdom, and the United States. These contributions address different questions: correlates place literacy within a wider explanatory context, whereas assessment research examines the evidence provided by scores. A descriptive self-report study occupies a narrower role by documenting how particular respondents appraise specified capabilities.

This role becomes more informative when results retain item content. A use score can combine perceived proficiency with an existing assistant and perceived difficulty learning a new tool. An evaluation score can combine judging limitations with comparing options and selecting an application. Domain means capture their shared direction, but item distributions reveal the statements underlying that direction. Neither a high total nor a domain ranking establishes that every student possesses the same combination of capabilities.

Literacy Education and Responsible Engagement

Educational support requires attention to both application and judgement. Pinski and Benlian (2024) reviewed AI literacy for users, including components, learning methods, and effects. Their user-oriented focus is relevant to university students who need to engage competently with systems without necessarily becoming AI developers. Sullivan et al. (2024) provide instructional evidence: a short workshop improved reported confidence and policy understanding, while the authors identified a continuing need to develop output-evaluation skills. That distinction cautions against assuming that all aspects of literacy develop together after initial orientation.

Policy and pedagogical research gives this distinction practical relevance. Chan (2023) connected AI policy with teaching and assessment, while Miao and Holmes (2023) emphasised human capacity in educational integration. Ahmad et al. (2026), reviewing AI-supported personalised research writing, also highlighted pedagogical design and learner engagement. Together, these perspectives support embedding literacy within learning activities that require decisions and justification, rather than limiting provision to demonstrations of tool functions.

For a survey to inform such design, its implications should remain connected to the questions asked. A self-rating of evaluation can motivate a subsequent task requiring students to examine an output. A privacy-risk statement can motivate a scenario requiring a decision about sharing information. The instructional activity then generates evidence beyond the original appraisal. This sequence gives descriptive data a practical role while leaving the effectiveness of any proposed teaching strategy open to evaluation.

Research Gap and the Present Study

Existing research provides multidimensional frameworks, student surveys, and approaches to assessment and instruction. The gap addressed here concerns the educational interpretation of descriptive evidence: an overall mean and domain ordering do not reveal whether a particular capability statement attracts broad endorsement, substantial midpoint responding, or negative appraisal. Without that distinction, universities may infer a uniform training need from an aggregate score or interpret all reserved responses as deficits. Item distributions offer a means of examining this problem within a specified sample and measure.

This study contributes a contextual account of 616 undergraduates across four academic years and three broad fields, recruited through 18 Shandong universities. It connects aggregate summaries with the response composition of specific capability statements and uses these patterns to identify questions for follow-up assessment. This adds interpretive detail to multidimensional self-report research without proposing a new construct or claiming provincial representativeness. The three research questions organise the analysis from overall description to domains and individual items.

Methodology

Design and Participants

A cross-sectional online survey was administered from 14 April to 12 June 2026. This article reports the descriptive literacy component of a broader doctoral survey of digital conditions, technology anxiety, and GenAI literacy. Only demographic characteristics and the 12 literacy

items are analysed here. Eligible participants were full-time undergraduates aged 18 years or above at participating universities in Shandong Province.

Institutional selection used proportionate regional stratification. The frame comprised 72 undergraduate institutions: 24 in Eastern, 36 in Central, and 12 in Western Shandong. Eighteen were randomly selected within strata, with six, nine, and three institutions selected, respectively. Students were then recruited voluntarily through institutional contacts and class-based or online communication channels. Proportional allocation concerned institutions, not individual students according to enrolment. The procedure therefore did not constitute simple random sampling of students.

Instrument and Scoring

The literacy measure was adapted from Wang et al. (2023). It contained three items each for awareness (AWA1–AWA3), use (USE1–USE3), evaluation (EVA1–EVA3), and ethics (ETH1–ETH3). The Chinese wording was retained with contextual adjustments referring to GenAI technologies in university learning, including assistants such as ChatGPT. This was contextual adaptation of an existing measure, not the development of a new instrument or a separate translation and back-translation of the literacy items.

Responses ranged from one (strongly disagree) to five (strongly agree). Negatively worded AWA2, USE2, and ETH2 were reverse-scored once as six minus the original response. Domain scores were means of their three scored items, and the overall score was the mean of all 12. Higher values indicated stronger perceived literacy. Because each domain contained three items, the overall mean also equalled the mean of the four domain scores.

A pilot involving 64 students with similar characteristics preceded the main survey; these respondents were excluded from the main sample. The pilot overall alpha was .895, with domain coefficients of .752–.830. Main-sample reliability and supporting factor-analytic evidence are reported in the findings.

Ethical Considerations

Ethical approval was obtained from the Human Research Ethics Committee of Universiti Pendidikan Sultan Idris, Malaysia, before data collection. Participants were informed of the study's purpose, the voluntary nature of participation, and the confidentiality of their responses. Informed consent was obtained before questionnaire completion. Findings are reported in aggregate form, and the analytical dataset used for this article excluded IP addresses and other administrative metadata.

Data Screening

The broader survey contained 80 substantive items and three demographic questions. Screening used multiple response-quality indicators, consistent with the general concern that careless responding may take different forms (Meade & Craig, 2012). The thresholds below were study-specific rules rather than criteria prescribed by that source. Of 852 submissions, 10 were excluded for IP locations outside Shandong and 148 for identical responses across at least 40 consecutive substantive items. IP location indicated geographical consistency rather than independently verifying enrolment.

A further 55 submissions showed content-based inconsistencies on three or more of the five negatively worded items in the broader questionnaire. These were assessed against positively worded statements addressing closely corresponding content; agreement with a negative statement alone was not classified as inconsistent. Among submissions with one or two such inconsistencies, 23 were excluded because completion took less than 120 seconds or the response pattern contained more than 30 consecutive identical answers. Each excluded submission was counted once.

The final sample contained 616 respondents, representing 72.3% of submissions. This is retention after screening, not a survey response rate, because invitation coverage was not established. All retained literacy items were complete integer responses between one and five; no imputation was required. The present analysis did not apply additional exclusions to the retained sample.

Analysis

Frequencies and percentages described participant characteristics. Means, sample standard deviations, and observed ranges described overall and domain scores. Item analysis reported means, standard deviations, and the complete five-category distributions. For compact presentation, scored responses were also grouped into positive direction (four or five), midpoint (three), and negative direction (one or two). For reverse-scored items, positive direction corresponds to disagreement with the original negative statement. These groups indicate response direction, not proficiency levels.

Cronbach's alpha assessed overall and domain internal consistency. Corrected item–total correlations were examined within each three-item domain and across the full 12-item score, together with alpha if an item was deleted. Supporting measurement evidence came from maximum-likelihood confirmatory factor analysis in AMOS 31. The correlated four-factor model assessed the item–domain structure; the second-order model examined a common factor underlying the four domains. Composite reliability and average variance extracted were interpreted at their respective measurement levels. These analyses used the same main-study sample and do not constitute independent cross-validation.

Descriptive statistics, five-category item frequencies, and five internal-consistency analyses were obtained using IBM SPSS Statistics 30 and checked against independent Python calculations from the retained Excel dataset. Results agreed to reported precision. Sample standard deviations are reported. Combined-category percentages were calculated from summed frequencies before rounding.

Analyses are unweighted sample descriptions. No significance tests of domain differences, demographic comparisons, or causal relationships were performed. Mean rankings are descriptive; the scale midpoint is a response reference rather than a competency threshold. No population confidence intervals or design-adjusted estimates are reported.

Findings

Participant Characteristics

The sample included 315 male students (51.1%) and 301 female students (48.9%). First-year students formed the largest academic-year group (35.6%). Humanities, natural sciences, and engineering accounted for 37.0%, 32.8%, and 30.2%, respectively (Table 1).

Table 1
Participant characteristics

Variable	Category	n	%
Gender	Male	315	51.1
Gender	Female	301	48.9
Academic year	First year	219	35.6
Academic year	Second year	159	25.8
Academic year	Third year	144	23.4
Academic year	Fourth year	94	15.3
Academic field	Humanities	228	37.0
Academic field	Natural sciences	202	32.8
Academic field	Engineering	186	30.2

Note. N = 616. Gender, year, and field each form a separate distribution. Percentages may not sum to 100 because of rounding.

Reliability and Supporting Measurement Evidence

The overall 12-item score yielded Cronbach's alpha of .900; domain coefficients ranged from .821 to .874 (Table 2). Within-domain corrected item–total correlations were .642–.771, compared with .486–.687 for the overall score. Removing any individual item reduced alpha in its domain and in the full score. All 12 items were retained. Each reliability analysis included 616 cases.

Table 2
Reliability and convergent validity

Domain	Items	Alpha	CR	AVE
Awareness	3	.840	.840	.638
Use	3	.837	.838	.636
Evaluation	3	.874	.874	.699
Ethics	3	.821	.821	.605

Note. N = 616. CR = composite reliability; AVE = average variance extracted. Domain CR and AVE concern item-level measurement in the second-order model. Overall 12-item alpha = .900. Higher-order CR = .859 and AVE = .605 concern the four domain loadings.

Supporting CFA indicated adequate fit for the correlated four-factor model, $\chi^2(48) = 126.282$, $p < .001$, $\chi^2/df = 2.631$, CFI = .980, and TLI = .973; standardised item loadings were .693–.860.

The second-order model also fit adequately, $\chi^2(50) = 130.822$, $p < .001$, $\chi^2/df = 2.616$, CFI = .979, TLI = .973, and RMSEA = .051. Its domain loadings were .766, .834, .795, and .710 for awareness, use, evaluation, and ethics, respectively. These results supported the specified scores within this sample. Item-based alpha and higher-order CR and AVE describe different measurement levels; neither establishes the model's unique superiority.

Overall and Domain Scores

For RQ1, overall literacy averaged 3.495 (SD = 0.650), with observed scores from 1.42 to 5.00. Overall skewness was −0.348 and excess kurtosis was −0.031. The mean exceeded the response midpoint, indicating the average direction of self-appraisal rather than an externally established competency level (Table 3).

Table 3
Overall and domain score descriptions

Score	Mean	SD	Minimum	Maximum	Skewness	Kurtosis
Overall	3.495	0.650	1.42	5.00	−0.348	−0.031
Awareness	3.477	0.843	1.00	5.00	−0.436	−0.484
Use	3.392	0.825	1.33	5.00	−0.307	−0.486
Evaluation	3.423	0.835	1.00	5.00	−0.340	0.035
Ethics	3.688	0.782	1.33	5.00	−0.723	0.282

Note. N = 616; all five scores had complete observations. Higher scores indicate stronger perceived literacy on the one-to-five scale. Observed minimum and maximum values are rounded to two decimal places. Kurtosis is excess kurtosis as reported by SPSS. Mean ordering was not tested for statistical significance.

For RQ2, ethics had the highest mean, followed by awareness, evaluation, and use. The ethics–use difference was approximately 0.296 points. Evaluation and use were numerically close, differing by approximately 0.030 points. Domain standard deviations ranged from 0.782 to 0.843. These comparisons describe the sample means rather than statistically established differences.

Item Response Patterns

For RQ3, item means ranged from 3.362 to 3.701. Table 4 presents content summaries, means, standard deviations, positive-direction responses, and midpoint percentages. The three original negative statements were reverse-scored; their summaries describe the scored direction. Appendix A provides the complete five-category frequencies.

Table 4
Item descriptions and positive-direction responses

Item	Content summary	Mean	SD	Positive n	Positive %	Midpoint %
AWA1	Distinguish AI-based from non-AI applications	3.394	0.975	301	48.9	32.3
AWA2	Understand how GenAI can help	3.500	0.968	343	55.7	27.3
AWA3	Recognise AI's role within an application	3.537	0.964	333	54.1	30.5
USE1	Use AI assistants proficiently for academic tasks	3.388	0.951	283	45.9	37.5
USE2	Less difficulty learning a new GenAI tool	3.362	0.934	289	46.9	34.9
USE3	Use GenAI to improve academic efficiency	3.427	0.967	296	48.1	35.2
EVA1	Evaluate AI capabilities and limitations	3.390	0.961	280	45.5	38.8
EVA2	Compare and choose among AI-generated options	3.459	0.904	313	50.8	35.7
EVA3	Select a task-appropriate AI tool	3.419	0.935	301	48.9	35.1
ETH1	Follow ethical principles	3.667	0.862	377	61.2	31.5
ETH2	Attend to personal data and privacy risks	3.695	0.956	400	64.9	22.7
ETH3	Remain alert to risks of GenAI misuse	3.701	0.915	389	63.1	27.1

Note. N = 616. AWA = awareness; USE = use; EVA = evaluation; ETH = ethics. AWA2, USE2, and ETH2 were originally negatively worded and are reported after reverse-scoring. Content summaries are not verbatim administered wording. Positive direction = scored 4 or 5; midpoint = 3. For the three reverse-scored items, positive direction corresponds to disagreement with the original negative statement. These categories describe response direction, not proficiency levels.

Awareness item means ranged from 3.394 to 3.537. Recognising AI's role within an application had the highest mean in this domain, whereas distinguishing AI-based from non-AI applications had the lowest. Positive-direction responses ranged from 48.9% to 55.7%. For the latter item, 199 students selected the midpoint (32.3%) and 116 a negative-direction response (18.8%).

Use item means ranged from 3.362 to 3.427. The reverse-scored new-tool learning item had the lowest mean across all 12 statements. Positive-direction responses ranged from 45.9% to 48.1%. For proficiency in using an assistant for academic tasks, 231 students selected the midpoint (37.5%), 283 a positive-direction response (45.9%), and 102 a negative-direction response (16.6%).

Evaluation item means ranged from 3.390 to 3.459. Comparing and choosing among AI-generated options had the highest mean within the domain. Evaluating capabilities and limitations had the lowest positive-direction proportion of all items (45.5%). Its midpoint group comprised 239 students (38.8%), while 97 selected negative-direction responses (15.7%). Selecting a task-appropriate tool received positive-direction responses from 301 students (48.9%).

Ethics showed a consistent pattern across reported adherence to principles, privacy attention, and vigilance concerning misuse. Means ranged from 3.667 to 3.701 and positive-direction proportions from 61.2% to 64.9%. Vigilance concerning misuse had the highest mean; the reverse-scored privacy item had the highest positive-direction proportion. Each ethics item had a higher sample mean and positive-direction proportion than any item in the other domains. These are descriptive ethical self-appraisals, not observed behaviour.

Discussion

Interpreting Overall and Domain Scores

The main finding is the combination of a positive overall mean with uneven endorsement of particular capability statements. Ethics received the highest domain mean, but practical use and evaluation contained substantial midpoint groups. For the three use items, fewer than half selected a positive-direction response. The aggregate score therefore summarised the direction of students' appraisals while the distributions revealed where that appraisal was more reserved. This distinction is useful for deciding what to assess next.

These results support using multidimensional frameworks as a way to organise inquiry rather than assuming that the domains develop uniformly. The awareness–use–evaluation–ethics scale distinguishes related aspects of engagement (Wang et al., 2023). Here, numerical proximity between use and evaluation suggests treating them as connected instructional questions: students must apply a tool to a purpose and judge whether its contribution is suitable. Their close means do not support a sharply ordered training priority.

The contribution is interpretive rather than a new hierarchy of literacy. A five-point mean can conceal different response compositions, and common response options do not make the capability statements equally demanding. As Lintner's (2024) review of scales makes clear, measurement evidence is central to literacy research. The present findings should consequently guide content-specific follow-up assessment, rather than support comparisons with unrelated surveys using a common numerical benchmark.

Ethical Self-Appraisal and Situated Judgement

Ethical adherence, privacy attention, and vigilance concerning misuse all received comparatively positive self-appraisals. Their consistency across three statements makes the descriptive pattern broader than a single favourable item. However, these statements concern reported responsibility and attention to risk, whereas the use and evaluation items ask about carrying out particular activities. Differences in the demands of the statements are therefore relevant alongside possible differences in perceived capability.

One possible explanation is that recognising a responsibility is easier to endorse than judging one's proficiency in a practical task. Social desirability and familiarity with ethical expectations

could also contribute. The survey did not separately measure these explanations. The appropriate educational interpretation is that ethical commitments provide an entry point for examining decisions in context, rather than that ethics instruction has already succeeded or become unnecessary.

This distinction matters when situating the result within earlier research. O'Dea et al. (2026) identified concerns about students' understanding of AI ethics using questions tied to AI-provided data and recommendations. The present ethics mean cannot resolve those concerns because its item coverage differs. A student may endorse vigilance while still requiring support in deciding whether to share information or use a generated recommendation. The two forms of evidence raise complementary questions about general commitment and situated judgement.

Chan's (2023) policy framework connects responsibility with teaching and assessment. For the present sample, this suggests following broad ethical endorsements with cases requiring students to justify a choice. Such assessment would examine what the survey leaves open: how students apply reported principles when convenience, privacy, attribution, and assignment rules intersect.

Application and Evaluation as Related Assessment Questions

The response composition for proficient academic use and evaluation of limitations is especially informative. Midpoint responses accounted for 37.5% and 38.8%, respectively, compared with negative-direction proportions of 16.6% and 15.7%. Thus, the relatively modest positive endorsement was accompanied mainly by a large middle group, rather than predominantly negative self-appraisal. Treating these respondents as uniformly incapable would obscure the pattern actually observed.

An intermediate rating can reflect partial capability, limited experience, uncertainty about the statement, or variation across tasks. For example, students may feel comfortable generating ideas but less certain about checking a specialised explanation. These are possible interpretations rather than subgroups established by the survey. Follow-up assessment should distinguish them before institutions decide whether students need introductory exposure, guided practice, evaluative instruction, or feedback on self-assessment.

Perception studies document why this distinction is relevant. Chan and Hu (2023) reported perceived benefits alongside concerns; Almassaad et al. (2024) examined utilisation, benefits, and challenges. Engagement and perceived usefulness provide context for literacy education but do not establish competence in assessing a particular output. The present item-level evidence narrows the educational question from general acceptance to the activities for which respondents express more reserved appraisals.

Instructional evidence likewise suggests attending to specific outcomes. Sullivan et al. (2024) found improved confidence and policy understanding after a workshop while identifying further development needs in output evaluation. The present survey does not evaluate that intervention, but its use and evaluation responses support examining whether instruction develops both practical engagement and judgement, rather than assuming that greater confidence represents comprehensive capability.

Contribution to Educational Assessment

Overall, domain, and item-level summaries answer different questions. The overall mean establishes broad self-appraisal; the domains organise areas for curriculum planning; and the distributions locate statements requiring closer inquiry. For evaluating limitations, for example, the sizeable midpoint group gives educators a reason to investigate students' decision processes. The added value of item analysis lies in this connection between a reported pattern and a specific next assessment question.

Knowledge-based instruments provide complementary evidence (Hornberger et al., 2025a, 2025b). More specifically, Jin et al. (2025) developed the Generative AI Literacy Assessment Test (GLAT) and found that its scores predicted performance in GenAI-supported tasks more effectively than the self-reported measures examined. This strengthens the rationale for combining perceived-capability questionnaires with assessments of GenAI knowledge and application. Such measures should be aligned with the intended learning objectives before differences between self-ratings and performance are interpreted as inaccurate self-appraisal.

This approach gives descriptive surveys a bounded educational role. They can help select assessment content and plan support, while subsequent tasks establish what learners can demonstrate. The present analysis supplies the first form of evidence; it does not identify individual learner profiles or evaluate the effectiveness of a teaching strategy.

Educational Implications

Link Practical Application with Evaluation

The use and evaluation results support a short learning sequence that connects application with judgement. Students could complete a permitted course task with a GenAI assistant, retain the relevant interaction, and explain which contribution they accepted, revised, or rejected. A rubric could address task suitability, checking of claims, and justification of the final response. This corresponds to the questionnaire's practical and evaluative content while producing observable evidence beyond self-appraisal.

Ou et al. (2024) provide a relevant pedagogical example through a self-regulated learning-based micro-curriculum for doctoral academic writing. Their analysis connected critical engagement with GenAI to developing prompting strategies, reconsidering ethical issues, and recognising text ownership. Although this advanced writing context differs from the present undergraduate sample, it illustrates how application and critical reflection can be addressed within the same learning sequence.

For students selecting the midpoint, an initial task could determine the required support. Difficulty choosing a tool may call for guided comparison; accepting unsupported claims may call for source-checking practice; competent performance accompanied by a cautious rating may call for feedback on self-assessment. These are possible instructional responses to subsequent evidence, not learner categories inferred from the current data.

Apply Ethical Commitments to Academic Decisions

The relatively positive ethics ratings justify using reported commitments as a starting point for discussion. Cases could require students to decide what personal information to withhold,

whether assistance complies with assignment rules, and how its contribution should be acknowledged. Assessment should consider the reasons given and the context of the decision. This extends the surveyed concerns about ethical adherence, privacy, and misuse into decisions that can be examined.

Such activities accord with the emphasis on human capacity in Miao and Holmes (2023). Expectations should be explicit before students are assessed, with examples of acceptable use and opportunities to discuss ambiguous cases. Strong ethical self-ratings alone provide no reason to remove ethics from instruction; equally, they provide no evidence that the respondents acted irresponsibly.

Use Follow Up Evidence to Differentiate Support

Universities could pilot the learning sequence with self-ratings and a small set of aligned tasks before and after instruction. The self-ratings would show perceived change, while the tasks would show changes in demonstrated performance. Keeping these outcomes distinct would help evaluate whether provision improves confidence, capability, or both.

The emphasis on learning design is consistent with Ahmad et al. (2026), whose review concerns AI-supported personalised research writing. For the current broader undergraduate sample, the practical priority is to test locally whether guided application and judgement address the capabilities identified for follow-up. An evaluated pilot would provide a stronger basis for scaling provision than the domain ordering alone.

Limitations and Future Research

The findings describe retained volunteers rather than all undergraduates in Shandong or China. Although universities were randomly selected within regional strata, student recruitment did not establish proportional representation or a known individual selection probability. Invitation coverage was unavailable. Unweighted descriptions also do not account for potential clustering within universities. These features limit population inference. Screening removed 27.7% of submissions using the full 80-item questionnaire. IP location may not perfectly reflect enrolment, and negatively worded items can produce inconsistent patterns for several reasons. The results are conditional on the applied rules; excluded responses were not reanalysed to estimate sensitivity to alternative thresholds. Reporting the screening process is therefore important to understanding the retained sample.

The adapted scale assesses perceived literacy and does not exhaust GenAI-specific competencies. It does not directly test factual verification, prompt quality, disclosure decisions, or actual privacy behaviour. Main-sample reliability and CFA support the specified scores but do not replace content validation, independent cross-validation, or measurement invariance. Reverse wording appears in three domains but not evaluation; possible wording effects were not isolated.

Domain and item comparisons are descriptive. Common response options do not imply equal item difficulty, and no inferential tests were conducted to establish differences. Midpoint responses were retained without assigning a single explanation. Cross-sectional data also cannot establish development or instructional effects. Experience with GenAI and prior training were not analysed as explanations of these patterns.

Future research should combine aligned self-ratings with practical tasks and investigate the meaning of midpoint responses through brief interviews or task-based reflection. Studies addressing demographic, institutional, or domain comparisons should first examine measurement comparability and use appropriate paired or clustered analyses. Longitudinal and evaluated instructional designs could then examine whether targeted provision changes perceived and demonstrated capability.

Conclusion

Among 616 undergraduates recruited through 18 Shandong universities, self-reported GenAI literacy showed a positive overall mean alongside more differentiated item responses. Ethics received the highest domain mean and comparatively positive responses across all three statements; use and evaluation contained sizeable midpoint groups. The principal contribution is showing how item distributions qualify aggregate summaries and identify specific questions for educational assessment. Universities can use these questions to connect guided application, evaluation of outputs, and situated ethical decisions, while evaluating subsequent performance rather than inferring proficiency from self-report alone.

References

Ahmad, N. H., Hanid, M. F. A., & Zakaria, M. I. (2026). Understanding AI-supported personalised learning in research writing: A systematic review of technologies, pedagogy, and learning outcomes. *International Journal of Academic Research in Progressive Education and Development, 15*(2), 65–96. https://doi.org/10.6007/IJARPED/v15-i2/28013

Almassaad, A., Alajlan, H., & Alebaikan, R. (2024). Student perceptions of generative artificial intelligence: Investigating utilization, benefits, and challenges in higher education. *Systems, 12*(10), Article 385. https://doi.org/10.3390/systems12100385

Almatrafi, O., Johri, A., & Lee, H. (2024). A systematic review of AI literacy conceptualization, constructs, and implementation and assessment efforts (2019–2023). *Computers and Education Open, 6*, Article 100173. https://doi.org/10.1016/j.caeo.2024.100173

Bewersdorff, A., Nerdel, C., & Zhai, X. (2025). How AI literacy correlates with affective, behavioral, cognitive and contextual variables: A systematic review. *Computers and Education: Artificial Intelligence, 9*, Article 100493. https://doi.org/10.1016/j.caeai.2025.100493

Carolus, A., Koch, M. J., Straka, S., Latoschik, M. E., & Wienrich, C. (2023). MAILS – Meta AI literacy scale: Development and testing of an AI literacy questionnaire based on well-founded competency models and psychological change- and meta-competencies. *Computers in Human Behavior: Artificial Humans, 1*(2), Article 100014. https://doi.org/10.1016/j.chbah.2023.100014

Chan, C. K. Y. (2023). A comprehensive AI policy education framework for university teaching and learning. *International Journal of Educational Technology in Higher Education, 20*, Article 38. https://doi.org/10.1186/s41239-023-00408-3

Chan, C. K. Y., & Hu, W. (2023). Students' voices on generative AI: Perceptions, benefits, and challenges in higher education. *International Journal of Educational Technology in Higher Education, 20*, Article 43. https://doi.org/10.1186/s41239-023-00411-8

Hornberger, M., Bewersdorff, A., Schiff, D. S., & Nerdel, C. (2025a). A multinational assessment of AI literacy among university students in Germany, the UK, and the US. *Computers in Human Behavior: Artificial Humans, 4*, Article 100132. https://doi.org/10.1016/j.chbah.2025.100132

Hornberger, M., Bewersdorff, A., Schiff, D. S., & Nerdel, C. (2025b). Development and validation of a short AI literacy test (AILIT-S) for university students. *Computers in Human Behavior: Artificial Humans, 5*, Article 100176. https://doi.org/10.1016/j.chbah.2025.100176

Jin, Y., Martinez-Maldonado, R., Gašević, D., & Yan, L. (2025). GLAT: The generative AI literacy assessment test. *Computers and Education: Artificial Intelligence, 9*, Article 100436. https://doi.org/10.1016/j.caeai.2025.100436

Laupichler, M. C., Aster, A., Schirch, J., & Raupach, T. (2022). Artificial intelligence literacy in higher and adult education: A scoping literature review. *Computers and Education: Artificial Intelligence, 3*, Article 100101. https://doi.org/10.1016/j.caeai.2022.100101

Lintner, T. (2024). A systematic review of AI literacy scales. *npj Science of Learning, 9*, Article 50. https://doi.org/10.1038/s41539-024-00264-4

Long, D., & Magerko, B. (2020). What is AI literacy? Competencies and design considerations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1–16). Association for Computing Machinery. https://doi.org/10.1145/3313831.3376727

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods, 17*(3), 437–455. https://doi.org/10.1037/a0028085

Miao, F., & Holmes, W. (2023). *Guidance for generative AI in education and research*. UNESCO. https://doi.org/10.54675/EWZM9535

Ng, D. T. K., Leung, J. K. L., Chu, S. K. W., & Qiao, M. S. (2021). Conceptualizing AI literacy: An exploratory review. *Computers and Education: Artificial Intelligence, 2*, Article 100041. https://doi.org/10.1016/j.caeai.2021.100041

Ng, D. T. K., Wu, W., Leung, J. K. L., Chiu, T. K. F., & Chu, S. K. W. (2024). Design and validation of the AI literacy questionnaire: The affective, behavioural, cognitive and ethical approach. *British Journal of Educational Technology, 55*(3), 1082–1104. https://doi.org/10.1111/bjet.13411

O'Dea, X., Ng, D. T. K., O'Dea, M., & Shkuratskyy, V. (2026). Factors affecting university students' generative AI literacy: Evidence and evaluation in the UK and Hong Kong contexts. *Policy Futures in Education, 24*(1), 13–34. https://doi.org/10.1177/14782103241287401

Ou, A. W., Khuder, B., Franzetti, S., & Negretti, R. (2024). Conceptualising and cultivating critical GAI literacy in doctoral academic writing. *Journal of Second Language Writing, 66*, Article 101156. https://doi.org/10.1016/j.jslw.2024.101156

Pinski, M., & Benlian, A. (2024). AI literacy for users – A comprehensive review and future research directions of learning methods, components, and effects. *Computers in Human Behavior: Artificial Humans, 2*(1), Article 100062. https://doi.org/10.1016/j.chbah.2024.100062

Sullivan, M., McAuley, M., Degiorgio, D., & McLaughlan, P. (2024). Improving students' generative AI literacy: A single workshop can improve confidence and understanding. *Journal of Applied Learning and Teaching, 7*(2), 88–97. https://doi.org/10.37074/jalt.2024.7.2.7

Wang, B., Rau, P. L. P., & Yuan, T. (2023). Measuring user competence in using artificial intelligence: Validity and reliability of artificial intelligence literacy scale. *Behaviour & Information Technology, 42*(9), 1324–1337. https://doi.org/10.1080/0144929X.2022.2072768