컴퓨터교육학회 논문지 2026년 제29권 제6호
https://doi.org/10.32431/kace.2026.29.6.010

공학계열 대학생의 생성형 AI 활용 과업-전략 프로파일 측정도구 개발 및 타당화

Development and Validation of a Task-Strategy Scale for Profiling Generative AI Use Among Undergraduates at an Engineering-Oriented University

이대영† · 유정††
Daeyeong Lee† · Liu Tang††

요약

본 연구는 공학계열 대학생의 생성형 AI 학습활용을 과업과 전략의 두 축에서 체계적으로 측정할 수 있는 도구를 개발하고 타당화하는 데 목적이 있다. 선행연구 검토를 통해 예비 문항을 구성한 후 전문가 5인의 내용타당도 검토를 거쳐 문항을 정제하였으며, 공학계열 대학생 201명의 응답 자료를 수집하여 문항분석, 탐색적 요인분석, 확인적 요인분석을 수행하였다. 그 결과 4요인 15문항의 측정도구를 확정하였으며, 과업 차원에서는 텍스트 기반 과업 활용과 미디어 및 데이터 과업 활용, 전략 차원에서는 비판적 검증 전략과 프롬프트 및 선택 전략으로 구성되었다. 본 도구는 총점이 아닌 4개 하위요인의 점수 조합을 통해 학습자의 생성형 AI 활용 양상을 다차원적으로 프로파일링할 수 있도록 설계되었으며, 향후 맞춤형 교육지원과 AI 리터러시 교육의 기초 자료로 활용될 수 있을 것으로 기대된다.

주제어 생성형 인공지능, 공학교육, 측정도구 개발, 과업-전략 프로파일

ABSTRACT

This study aimed to develop and validate a scale for systematically assessing undergraduates' use of generative AI for learning along two dimensions: task and strategy, at an engineering-oriented university. Preliminary items were generated through a review of prior research and refined through content validity evaluation by a panel of five experts. Survey data from 201 undergraduates were analyzed using item analysis, exploratory factor analysis (EFA), and confirmatory factor analysis (CFA). The final scale comprises 15 items across four factors, with text-based task use and media and data task use representing the task dimension, and critical verification strategy and prompting and selection strategy representing the strategy dimension. Rather than yielding a single composite score, the scale was designed to generate multidimensional profiles of learners' generative AI use based on combinations of scores across the four subscales. The scale is expected to provide a foundation for tailoring educational support and informing AI literacy education.

Keywords Generative Artificial Intelligence, Engineering Education, Scale Development, Task-Strategy Scale

†정회원 한국공학대학교 전공교육혁신센터 연구교수
††정회원 한국공학대학교 전공교육혁신센터 연구교수(교신저자)

논문투고 2026년 03월 13일
심사완료 2026년 05월 11일
게재확정 2026년 05월 21일
발행일자 2026년 06월 30일

1. Introduction

Generative artificial intelligence (GenAI) has rapidly spread across higher education since 2022. Undergraduate students increasingly use these tools to accomplish a wide range of academic activities, including assignment completion, study support, and idea generation. The rapid diffusion has generated both considerable optimism regarding gains in learning efficiency and accessibility, and growing concern over issues of academic integrity, plagiarism, information reliability, and potential degradation of learning quality. Previous studies on students' GenAI use have shown that students primarily use these tools for brainstorming, text summarization, and information retrieval, while also reporting concerns about ethical issues and the need for clearer institutional guidelines [1-3]. Furthermore, emerging evidence suggests that GenAI use varies significantly depending on the academic stage and task type, indicating that a single measure of overall usage frequency is insufficient to fully explain how university students engage with GenAI [2,3].

Scholarly efforts to assess university students' use of generative AI have continued to grow. However, previous research has been limited by a fragmented structure, proceeding largely along two separate lines of inquiry. One strand has focused on the academic tasks for which students primarily use AI, such as writing, idea generation, and exam preparation [1,2]. The other has examined strategic use, including how learners formulate prompts and evaluate hallucinations in AI-generated outputs, often without considering specific task contexts [4-8]. This fragmented approach makes it difficult to capture the integrated ways in which learners perform academic tasks while simultaneously checking the reliability of AI-generated outputs and regulating their interaction with AI.

This limitation is even more apparent in engineering-oriented learning contexts. In tasks requiring high accuracy, such as code generation and data processing [2,3], verification behaviors for checking logical errors in AI-generated outputs are likely to be closely intertwined with prompt refinement and selection behaviors for deriving results that meet specific task demands. Accordingly, GenAI use needs to be understood not by separating what learners do (task) from how they perform and regulate it (strategy), but by considering how these two aspects operate together in actual learning performance. Nevertheless, no standardized instrument has yet been developed that captures students' task characteristics and strategic use behaviors within a unified multidimensional structure, particularly in engineering-oriented university contexts.

In response to this gap, this study aimed to develop and validate a multidimensional scale that comprehensively measures how undergraduates at an engineering-oriented university use generative AI in the course of disciplinary learning. Specifically, the scale was designed to capture two task-use factors, text-based task use and media and data task use, as well as two strategy factors, prompting and selection strategy and critical verification strategy. By developing a scale based on an integrated four-factor model encompassing both task and strategy, this study sought to move beyond simple frequency-based assessment and to establish a basis for multidimensional profiling of learners' generative AI use.

To achieve this aim, the study addressed the following research questions:

RQ1. What is the factor structure of the scale designed to measure the task–strategy profile of generative AI use among undergraduates at an engineering-oriented university?

RQ2. Does the developed scale demonstrate adequate reliability and validity?

2. Theoretical Background

2.1 Generative AI Use in Engineering Education

Research on generative AI use in higher education initially focused on how learners used generative AI for academic purposes. These studies found that students used generative AI for a range of activities, including assignment completion, information retrieval, and idea generation [1,9]. At the same time, learners were found to recognize both the convenience and usefulness of generative AI while also expressing concerns about output reliability and ethical use. Issues related to verification have also remained prominent in the literature. These include concerns about the accuracy of AI-generated

outputs, the opacity of information sources, and the possibility that learners may accept AI-generated content uncritically [1,4,7]. Taken together, these findings suggest that a more precise understanding of generative AI use in this learning environment requires going beyond frequency-based indicators to capture both task-related use patterns and strategic regulatory behaviors.

Much of the existing research on generative AI use has adopted a technology acceptance perspective, focusing primarily on intention to use and frequency of use [10,11]. However, this perspective is limited in explaining the specific academic tasks for which learners use generative AI and how they regulate their interactions with it during task performance.

The need for more precise measurement becomes particularly evident in engineering education. Compared with students in non-engineering disciplines, engineering undergraduates use generative AI more frequently for producing technical outputs [2-3], especially in tasks that require high accuracy, such as code generation and data analysis. Given the comparatively high proportion of accuracy-critical tasks in engineering curricula, it is necessary to consider not only the diversity of task types but also the importance of verifying AI-generated outputs.

2.2 Task Types in Generative AI Use

A task-type perspective on generative AI use offers a multidimensional way of understanding the academic tasks for which learners use generative AI. It is based on the premise that generative AI use may take different forms depending on the nature of the task and the learning context [2,3]. Prior studies have repeatedly shown that generative AI use varies according to task characteristics. Specifically, Ravšelj et al. [1] and Oh [9] found that university students' AI use tends to concentrate on text-based tasks such as information retrieval, drafting, and summarization. Singer-Freeman et al. [2] also found that generative AI use varies across academic tasks in both frequency and pattern, with relatively frequent use for study preparation tasks and comparatively lower use for writing and data-related tasks. In addition, Stöhr et al. [3] derived usage profiles by measuring the degree of students' reliance on AI across multiple tasks, showing that task type can serve as a key criterion for distinguishing patterns of AI use. Taken together, these findings suggest that task type should be understood not as a simple frequency variable, but as a multidimensional construct.

However, most previous studies, even when providing lists of tasks, have not systematically reflected them in a measurement structure [1-2,9]. In particular, non-text tasks frequently performed by engineering learners, such as data analysis, visual material generation, and structural formatting of documents, have not been sufficiently addressed. Thus, prior task classifications have focused mainly on general text-based writing and have therefore been limited in covering the non-text tasks distinctive to engineering education.

In the context of engineering education, generative AI use can be broadly understood in terms of two task types. Prior research has shown that university students often use generative AI for text-based academic activities, such as information retrieval, drafting, summarization, and learning support [1,2,9], which in this study are conceptualized as text-based task use. At the same time, engineering learners also use generative AI for more structured forms of output production, including code generation, data analysis, visual or video material generation, and document formatting, which are conceptualized here as media and data task use. The latter category is particularly relevant to engineering education and also includes document formatting, as it involves rule-based structural processing rather than simple text generation.

2.3 Advanced Use of Generative AI

The qualitative differences in how learners engage with generative AI cannot be adequately explained by usage volume or task categories alone. Even when performing identical tasks, the quality of AI use and the reliability of its outputs vary considerably depending on how learners construct and refine their prompts, how they select among multiple generated outputs, and how they verify the accuracy and sourcing of AI-generated content [4-8,11-14]. Recent research on AI literacy has increasingly emphasized critical judgment and self-regulatory behaviors as core dimensions of AI use competence, approaching these capabilities from a multidimensional perspective [12-14]. However,

few studies have examined these capacities together with the concrete interaction strategies learners use during actual academic task performance [2,3].

Research on prompting behavior has directly addressed the strategic aspects of interaction with generative AI. Gibreel and Arpaci [6] proposed prompt construction and adjustment ability as a measurable competence dimension, while Lee and Palmer [5] showed that prompt design and revision are not merely input techniques but strategic acts involving goal specification, constraint setting, and the comparison and selection of outputs. These findings suggest that the ability to regulate the generative phase of AI interaction represents a distinct and assessable strategic construct.

Alongside the forms of control during the generative phase, verification and critical evaluation of outputs also need to be treated as a separate strategic dimension. Yan et al. [8] identified information verification behavior as a key outcome variable and demonstrated its importance in the context of generative AI use. Domestic studies have likewise highlighted the risks of uncritical acceptance of AI-generated content and the need for source checking and reliability assessment [4,7]. Overall, previous studies suggest that strategic behavior in generative AI use should not be treated as a single dimension, but should instead be conceptualized in terms of prompting and selection during generation and critical verification after generation.

2.4 Synthesis of Previous Studies

Previous research on generative AI use can be grouped into four main lines of inquiry. First, AI literacy-centered studies have examined AI use competence from cognitive, behavioral, ethical, and evaluative perspectives [12-14]. Second, AI use behavior studies have explained intention to use and frequency of use from a technology acceptance perspective [10,11]. Third, task type-centered studies have analyzed the specific academic tasks for which learners use generative AI [1-3,9]. Fourth, studies on prompting strategies, verification, and evaluative behavior have highlighted the importance of prompt refinement [5,6] and the need for critical evaluation and information verification of AI-generated content [4,7,8]. Table 1 presents a thematic synthesis of these strands of research, their major findings, and their implications for item development and measurement structure design in this study.

Table 1. Thematic Synthesis of Key Studies on GenAI

Study	Research Stream	Key Findings	Implications
Ng et al. (2024)	AI Literacy	Multidimensional AI literacy validated via EFA/CFA.	Link literacy to GenAI tasks and verification.
Im et al. (2024)	AI Literacy	Reliable multi-factor AI literacy scale for Korean college students.	Complement literacy with task–strategy profiling.
Ranieri et al. (2025)	AI Literacy	Critical AI literacy scale developed and refined.	Offers a useful model for item refinement and validation.
Kaya & Adıgüzel (2025)	AI Use Behavior	Acceptance factors predicted intention and varied across groups.	Acceptance alone is insufficient for task–strategy profiling.
Noh et al. (2024)	AI Use Behavior	Usefulness and ease of use predicted continuance intention.	Add task- and strategy-based measures beyond continuance models.
Ravšelj et al. (2025)	Task Type	Use centered on brainstorming, summarizing, and information search.	Distinguish task types and include post-output verification.
Oh, E. (2025)	Task Type	Main uses included assignments, exam study, and idea generation.	Supports task-category measurement and institutional support.
Singer-Freeman et al. (2025)	Task Type	Use varied by academic stage and task type.	Supports task differentiation and stage-sensitive interpretation.
Stöhr et al. (2024)	Task Type	Use and attitudes varied by gender, level, and field.	Reflect contextual differences in profiling and guidance.
Lee, Y. (2023)	Verification Behavior	High use but moderate trust, with misinformation-related concerns.	Supports critical verification and risk-mitigation guidance.
Kim & Kang (2025)	Verification Behavior	Benefits coexisted with ethical, privacy, and misuse concerns.	Positions ethics and verification as key strategy constructs.
Yan et al. (2026)	Verification Behavior	GenAI literacy was linked to information verification behavior.	Supports verification as a distinct outcome and strategy component.
Gibreel & Arpaci (2025)	Prompting	PECS showed stable structure and high reliability.	Supports prompting competence as a measurable construct.
Lee & Palmer (2025)	Prompting	Prompt engineering was identified as a teachable skill.	Supports strategy-focused assessment and highlights skill gaps.

3. Research Method

3.1 Research Procedure

To develop and validate the Task–Strategy Scale for Profiling Generative AI Use, this study proceeded in five phases.

First, generative AI use in academic learning contexts was organized into two broad domains, task-related use and strategy-related use, and an initial item pool was constructed on that basis. Items for the task-related domain were developed with reference to Singer-Freeman et al. [2] and Ravšelj et al. [1], yielding 15 candidate items across four subfactors. Items for the strategy-related domain were developed with reference to Lee and Palmer [5], Gibreel and Arpaci [6], and Yan et al. [8], yielding 10 candidate items across three subfactors. Second, to establish content validity, a panel of five experts in education, communication, and educational measurement reviewed the initial items. Based on the quantitative results and qualitative feedback, abstract terms that were not appropriate for the student level were revised, and the items were reorganized to reflect actual generative AI use contexts in engineering education. Third, the items for the main survey were finalized, and data were collected from 201 undergraduates enrolled at an engineering-oriented university. Fourth, item analysis, exploratory factor analysis (EFA), and confirmatory factor analysis (CFA) were conducted to examine the integrated factor structure. Fifth, reliability and both convergent and discriminant validity were assessed for the finalized factor structure.

3.2 Scale Development

3.2.1 Preliminary Item Generation

Preliminary item generation began with a broad initial pool of candidate items designed to cover the conceptual components of both task-related use and strategy-related use as comprehensively as possible. The initial item pool was then refined through expert review and empirical analysis (see Table 2).

The task-related domain was conceptualized as a categorization of specific academic activities in which learners integrate generative AI into disciplinary learning. The initial pool for this domain consisted of 15 items across four preliminary subfactors: Writing and Editing, Idea Generation, Learning Support, and Technical and Creation.

The strategy-related domain was conceptualized as a pattern of active human–AI collaborative behavior through which learners optimize learning outcomes by refining prompts iteratively and verifying outputs. The initial pool for this domain consisted of 10 items across three preliminary subfactors: Prompt Engineering, Verification Behavior, and Critical Choice.

Table 2. Preliminary Item Pool and Construct Definitions

Construct	Proposed Subfactor	Definition	No.of Items	Item Codes
Task-related Domain	Writing and Editing	AI-supported drafting and refinement of academic writing.	3	TA11 – TA13
	Idea Generation	AI-assisted ideation to expand concepts and reduce cognitive fixation.	2	TA21 – TA22
	Learning Support	AI as personalized scaffolding for understanding complex content.	4	TA31 – TA34
	Technical and Creation	AI-enabled production of non-text artifacts for engineering tasks.	6	TA41 – TA46
Strategy-related Domain	Prompt Engineering	Strategic prompt formulation and iteration to optimize AI outputs.	4	AU11 – AU14
	Verification Behavior	Cross-checking AI outputs with authoritative sources to ensure accuracy.	3	AU21 – AU23
	Critical Choice	Evaluating AI options and selecting the best solution for the task.	3	AU31 – AU33

3.2.2 Content Validity Review

A content validity review was conducted to verify whether the survey items developed by the research team adequately reflected the purpose and target population of the study. The composition of the expert panel is presented in Table 3. The panel was asked to evaluate the content validity of all 25 items, including 15 items from the task-related domain and 10 items from the strategy-related domain.

To support the experts' evaluation, the review materials included both the original English source items and a preliminary Korean translation prepared by the research team, as several of the constructs represented in these items have not yet been clearly established in the Korean context. Experts rated

each item on a five-point Likert scale ranging from 1 (not appropriate at all) to 5 (very appropriate). For any item rated as not appropriate at all, not appropriate, or neutral, experts were asked to provide the reason for the rating and suggest directions for revision or improvement. In addition, they were invited to record any recommendations regarding item addition, integration, separation, deletion, or modification in an open-ended comment section.

Table 3. Characteristics of the Expert Panel (N=5)

Degree	Specialization	AI Research Experience (Years)	Scale/Instrument Development Experience (Years)
PhD	Gifted Education	1	3
PhD	Communication	2	4
PhD	Educational Technology	2	3
PhD	Educational Administration	1	2
PhD	Educational Measurement and Evaluation	3	10

3.2.3 Survey Administration

Data were collected via an online survey administered from February 4 to February 12, 2026, targeting currently enrolled students at T University, a specialized four-year technology institution located in the Seoul metropolitan area. Prior to participation, all respondents were informed of the study's purpose, the anonymous nature of their responses, the scope of personal data collection, and the intended use of the data. A total of 206 responses were received. Of these, five were excluded due to either insincere response patterns or non-consent to personal data collection, yielding a final analytical sample of 201 participants.

While the majority of the sample (86.1%) belonged to traditional engineering colleges, it also included students from the Business School (n = 12) and the Design Engineering School (n = 16). Importantly, these specific schools offer STEM-integrated majors such as IT Management, Data Science Management, and Industrial Design Engineering. Therefore, rather than being strictly limited to traditional engineering disciplines, the sample is more accurately defined as undergraduates sharing a cohesive, engineering-oriented educational context. The demographic characteristics of the study sample are presented in Table 4.

Table 4. Participant Characteristics (N=201)

Characteristic	Category	n (%)
Gender	Male	138 (68.7)
	Female	63 (31.3)
Academic Year	Freshman	21 (10.4)
	Sophomore	76 (37.8)
	Junior	35 (17.4)
	Senior or above	69 (34.3)
College	SW College	45 (22.4)
	IT & Semiconductor Convergence	36 (17.9)
	Smart Mechanical Convergence	46 (22.9)
	Advanced Convergence College	46 (22.9)
	Business School	12 (6.0)
	Design Engineering School	16 (8.0)

3.3 Data Analysis

All analyses were conducted in R (version 4.3.3) using the psych and lavaan packages. The data were analyzed in four steps. First, descriptive statistics (means and standard deviations) were computed for all items, and skewness and kurtosis were examined to assess univariate normality.

Second, exploratory factor analysis (EFA) was conducted to examine the factor structure of the scale. Given the ordinal nature of the five-point Likert-type responses, a polychoric correlation matrix was used. Factors were extracted using Principal Axis Factoring (PAF), and an oblique rotation method (Oblimin) was applied. The factorability of the data was assessed using the Kaiser–Meyer–Olkin (KMO) measure and Bartlett's test of sphericity.

Third, confirmatory factor analysis (CFA) was conducted to evaluate the validity of the four-factor structure derived from the EFA. The weighted least squares mean- and variance-adjusted (WLSMV) estimator was used because it is robust to multivariate non-normality in ordinal data. Model fit was evaluated using the Comparative Fit Index (CFI), Tucker–Lewis Index (TLI), Root Mean Square Error of Approximation (RMSEA), and Standardized Root Mean Square Residual (SRMR).

Fourth, the reliability and validity of the final scale were examined. Internal consistency reliability was assessed using Cronbach's alpha. Convergent validity

was evaluated by calculating Composite Reliability (CR) and Average Variance Extracted (AVE). Discriminant validity was assessed primarily using the Fornell–Larcker criterion [15]. When inter-factor correlations were high or the Fornell–Larcker criterion was borderline, the Heterotrait–Monotrait ratio (HTMT) was additionally calculated. An HTMT value below .85 was used as the criterion for discriminant validity [16].

4. Results

4.1 Content Validity Results

A single round of ratings was completed by five experts for the 25 preliminary items, including 15 items from the task-related domain and 10 items from the strategy-related domain. Content validity was evaluated primarily using the Item-level Content Validity Index (I-CVI) and Aiken's V, summarized at the construct level. The level of expert agreement was additionally examined using the median IQR and the proportion of items with IQR $\leq$ 1 (see Table 5).

For the task-related domain, the mean I-CVI (S-CVI/Ave) was .875, the S-CVI/UA was .562, the mean Aiken's V was .875, and the proportion of items with IQR $\leq$ 1 was .875. For the strategy-related domain, the mean I-CVI (S-CVI/Ave) was .920, the S-CVI/UA was .800, the mean Aiken's V was .870, and the proportion of items with IQR $\leq$ 1 was .800.

The S-CVI/UA reflects the proportion of items on which all experts reached agreement. The relatively low S-CVI/UA of .562 for the task-related domain suggests that some items elicited differing judgments across experts. Accordingly, five items were revised based on expert feedback: TA22, TA32, TA46, AU23, and AU33. These items showed either lower than average Aiken's V values at the construct level or relatively large IQR values. Revisions focused primarily on clarifying terminology, removing redundant expressions, and improving item comprehensibility for respondents.

Overall, both domains demonstrated acceptable content validity. The S-CVI/Ave ranged from .875 to .920, mean Aiken's V ranged from .870 to .875, and the proportion of items with IQR values of 1 or lower ranged from .800 to .875. Accordingly, the preliminary item pool was considered to demonstrate acceptable content validity [17].

Table 5. Summary of Content Validity Indices by Construct

Construct	No. of Items	S-CVI/Ave	S-CVI/UA	Mean Aiken's V	Prop. IQR ≤ 1
Task-related domain	15	.875	.562	.875	.875
Strategy-related domain	10	.920	.800	.870	.800

4.2 Descriptive Statistics and Item Normality

Descriptive statistics and normality checks were conducted for the 25 preliminary items retained after the content validity review (see Table 6). Item means ranged from 2.60 to 4.46, and standard deviations ranged from 0.57 to 1.53. To assess item-level normality, skewness and kurtosis were examined. The absolute values of skewness ranged from 0.25 to 1.58, and the absolute values of kurtosis ranged from 0.04 to 2.77. Based on the commonly used criteria of absolute skewness below 3.0 and absolute kurtosis below 10.0, no extreme non-normality was observed at the item level.

Corrected item-total correlations (ITC) were then computed within each initial theoretical subfactor. ITC values ranged from .241 to .751. Following the criterion suggested by Nunnally and Bernstein [18], according to which an ITC of .30 or higher indicates an acceptable item, AU12 ("I can write effective prompts to get desired results.") was removed because it did not meet this threshold.

Table 6. Descriptive Statistics and Item Analysis for the Initial Item Pool

Item	M	SD	Skew	Kurt	ITC
TA11	4.11	0.754	−0.827	1.392	.304
TA12	4.28	0.734	−0.958	1.018	.400
TA13	3.47	1.273	−0.552	−0.787	.354
TA21	3.97	1.040	−1.060	0.678	.478
TA22	3.86	1.192	−0.932	−0.043	.478
TA31	4.46	0.691	−1.173	1.106	.451
TA32	4.27	0.849	−1.576	2.332	.460
TA33	4.09	0.978	−1.229	1.323	.379
TA34	4.00	1.090	−1.180	0.870	.552
TA41	3.79	1.168	−0.899	−0.313	.526
TA42	3.27	1.399	−0.378	−1.180	.459
TA43	2.60	1.531	0.250	−1.535	.546
TA44	3.58	1.213	−0.657	−0.438	.574
TA45	3.19	1.357	−0.352	−1.096	.626
TA46	3.55	1.204	−0.769	−0.191	.535

Item	M	SD	Skew	Kurt	ITC
AU11	4.18	0.649	−0.643	1.236	.300
AU12	4.46	0.566	−0.606	0.368	.241
AU13	4.40	0.642	−1.065	2.766	.385
AU14	4.41	0.594	−0.434	−0.671	.443
AU21	4.01	1.022	−0.946	0.167	.652
AU22	4.05	0.873	−1.042	1.378	.751
AU23	4.18	0.904	−1.100	0.943	.665
AU31	4.22	0.710	−0.860	1.139	.398
AU32	4.04	0.845	−1.071	1.725	.428
AU33	4.02	0.755	−0.808	0.944	.462

Note. M = mean; SD = standard deviation; Skew = skewness; Kurt = kurtosis; ITC = corrected item–total correlation. Item AU12 (bold) was removed due to low ITC (< .30).

4.3 Exploratory Factor Analysis (EFA) Results

After AU12 was removed based on the ITC criterion in Section 4.2, the remaining 24 items were examined for factorability. The KMO measure of sampling adequacy was .800, and Bartlett's test of sphericity was significant, (276) = 1709.122, p < .001, supporting the suitability of the data for factor analysis.

EFA was conducted on the 24 candidate items. The number of factors to retain was determined on the basis of theoretical interpretability, the scree plot, parallel analysis, and eigenvalues. Items were removed iteratively based primarily on low primary loadings (< .40) and salient cross-loadings (loading differences < .20), in conjunction with theoretical interpretability. Communalities were used as a secondary diagnostic indicator to flag items for further review rather than as a strict deletion rule. Accordingly, a small number of items with lower communalities were retained to preserve content coverage and maintain an identifiable factor with an adequate number of indicators. Using these criteria, nine items were removed, resulting in a final four-factor solution with 15 items (see Table 7).

Table 7. Deleted Items and Reasons for Deletion

Item	Item statement	Reasons for Deletion
TA12	I use AI to organize my writing.	Content redundancy within the factor
TA21	I use AI to generate ideas (e.g., brainstorming).	Salient cross–loading
TA22	I use AI to create plans.	Salient cross–loading
TA31	I use AI to explain concepts.	Redundant content / Cross–loading
TA32	I use AI for tutoring.	Redundant content / Cross–loading
TA41	I use AI to generate code.	Low primary loading
TA46	I use AI to conduct a literature review.	Low primary loading
AU13	I can iterate prompts to improve output.	Redundant content and/ or cross–loading
AU33	I can evaluate the capabilities and limitations of a generative AI application after using it for a while.	Redundant content and/ or cross–loading

Note. AU12 was removed at the item analysis stage due to low ITC (< .30).

The EFA yielded a four-factor solution with eigenvalues of 1.211, 2.717, 1.140, and 1.845 (see Table 8). The four factors accounted for 8.1%, 18.1%, 7.6%, and 12.3% of the total variance, respectively, with a cumulative explained variance of 46.1%. Item communalities ranged from .219 to .729, and all retained items had factor loadings of .40 or higher. Based on the content of the retained items and their loading patterns, the four factors were labeled as follows: Text-based Task Use (3 items), Media and Data Task Use (5 items), Prompting and Selection Strategy (4 items), and Critical Verification Strategy (3 items).

Table 8. EFA Pattern Matrix, Communalities, and Variance Explained for Retained Items

Item	Factor 1 (TU)	Factor 2 (MU)	Factor 3 (PS)	Factor 4 (CS)	h^2
TA11	**.665**	−.085	.064	−.007	.451
TA33	**.560**	.040	−.009	.072	.315
TA34	**.586**	.171	.001	.020	.388
TA13	.177	**.601**	.008	.010	.391
TA42	−.043	**.721**	.006	.093	.526
TA43	−.098	**.855**	−.084	.017	.687
TA44	.196	**.694**	−.010	−.034	.519
TA45	.043	**.678**	−.004	.008	.480
AU21	.022	−.010	.068	**.735**	.557
AU22	−.065	.003	.030	**.854**	.729
AU23	−.007	−.022	.012	**.724**	.525
AU11	.064	−.032	**.407**	.007	.269
AU14	.001	−.022	**.466**	.026	.309
AU31	−.009	.008	**.547**	−.010	.360
AU32	.003	−.013	**.401**	.018	.219
Eigen value	1.211	2.717	1.140	1.845	Cumulative % =46.1
% of Variance	8.1	18.1	7.6	12.3	

Note. PAF = principal axis factoring. TU = text–based task use; MU = media and data task use; PS = prompting and selection strategy; CS = critical verification strategy. Pattern coefficients are reported from an oblimin–rotated PAF solution based on polychoric correlations. Primary factor loadings are shown in bold. h^2= communality

4.4 Confirmatory Factor Analysis (CFA) Results

CFA was conducted to evaluate the validity of the four-factor measurement model derived from the EFA. Because the scale used an ordinal five-point Likert-type responses format, the WLSMV estimator was applied. Before finalizing the model, item pairs with theoretically justifiable residual covariance were identified on the basis of item content and wording. Residual covariances were allowed for two theoretically justified item pairs (TA42–TA43 and AU22–AU23).

Model fit was evaluated using X^2, CFI, TLI, RMSEA, and SRMR. The final model yielded the following fit indices: $X^2(82) = 158.782$, $p < .001$, CFI = .923, TLI = .901, RMSEA = .068, and SRMR = .065. Both CFI and TLI exceeded .90, and both RMSEA and SRMR were below .08, indicating that the four-factor model showed acceptable fit to the data.

Standardized factor loadings were also examined to determine whether each item loaded significantly on its intended factor. All items had standardized loadings of .40 or higher, and all loadings were statistically significant ($p < .001$). Detailed CFA results are presented in Table 9.

Table 9. Standardized Factor Loadings, Residual Variances, and R^2 from the CFA

Item	Item	Standardized Loading (λ)	Residual Variance	R^2
Text-based Task Use	TA11	.562	.684	.316
	TA33	.588	.654	.346
	TA34	.726	.472	.528
Media and Data Task Use	TA13	.695	.517	.483
	TA42	.653	.574	.426
	TA43	.715	.489	.511
	TA44	.797	.366	.634
	TA45	.740	.453	.547
Prompting and Selection Strategy	AU11	.549	.698	.302
	AU14	.464	.785	.215
	AU31	.544	.705	.295
	AU32	.612	.626	.374
Critical Verification Strategy	AU21	.885	.217	.783
	AU22	.730	.467	.533
	AU23	.625	.609	.391

Note. All standardized factor loadings were significant at $p < .001$. R^2 represents item reliability (squared multiple correlation). Residual variance was calculated as $1 - X^2$.

4.5 Reliability and Validity Results

To evaluate the reliability and validity of the measurement model, Cronbach's alpha (α), Composite Reliability (CR), Average Variance Extracted (AVE), and inter-factor correlations were computed. Discriminant validity was examined using the Fornell–Larcker criterion by comparing the square root of each factor's AVE (diagonal elements) with the inter-factor correlations (off-diagonal elements) (see Table 10).

In terms of reliability, media and data task use showed good internal consistency (α= .853), and critical verification strategy also demonstrated acceptable reliability (α= .828). In contrast, text-based task use (α= .654) and prompting and selection strategy (α= .623) showed relatively lower values. These lower values indicate that the two subscales should be interpreted with caution; their substantive and theoretical implications are examined in the Discussion.

With respect to convergent validity, media and data task use (AVE = .520) and critical verification strategy (AVE = .569) met the conventional .50 criterion. In contrast, text-based task use (α= .654, AVE = .397) and prompting and selection strategy (α= .623, AVE = .297) showed comparatively lower internal consistency and convergent validity indices, and these results should be interpreted with caution (see Table 10). However, all four factors showed CR values above .60, ranging from .626 to .844. Following the criterion suggested by Fornell and Larcker [15], convergent validity was therefore considered acceptable at the overall construct level.

Inter-factor correlations ranged from .294 to .582. Within the task domain, media and data task use and text-based task use showed the highest correlation (r = .582), which is theoretically interpretable given that both involve academic task performance using generative AI. Within the strategy domain, critical verification strategy and prompting and selection strategy were also moderately correlated (r = .560), suggesting that prompt refinement and strategic selection may co-occur with verification behavior in actual use.

Regarding discriminant validity, the square root of each factor's AVE exceeded or approximated the corresponding inter-factor correlations for most factor pairs, indicating general support for discriminant validity. One borderline case was observed: the correlation between critical verification strategy and prompting and selection strategy (r = .560) slightly exceeded the square root of the AVE for prompting and selection strategy (.545). To

examine this case further, the Heterotrait–Monotrait ratio (HTMT) was additionally calculated. Using the criterion of HTMT ≤ .85 [16], the HTMT value for this factor pair was .535, supporting adequate discriminant validity between critical verification strategy and prompting and selection strategy.

Table 10. Reliability, Convergent Validity, and Discriminant Validity of the Four Factors

Factor	Text-based Task Use	Media and Data Task Use	Critical Verification Strategy	Prompting and Selection Strategy
TU	**.630**			
MU	.582	**.721**		
CS	.294	.360	**.754**	
PS	.436	.422	.560	**.545**
α	0.654	0.853	0.828	0.623
CR	0.661	0.844	0.795	0.626
AVE	0.397	0.520	0.569	0.297

Note. Bold diagonal values represent the square root of the average variance extracted (AVE). Off-diagonal values represent latent factor correlations estimated in the CFA. All correlations were significant at $p < .01$. α= Cronbach's alpha; CR = composite reliability; AVE = average variance extracted. TU = text-based task use; MU = media and data task use; CS = critical verification strategy; PS = prompting and selection strategy.

4.6 Final Scale

Following the expert content validity review and the EFA and CFA of the main survey data, the final version of the Task-Strategy Scale for Profiling Generative AI Use was established as a 15-item scale with a four-factor structure (see Table 11). The final factor structure consists of two task-related factors, Media and Data task use (5 items) and Text-based task use (3 items), and two strategy-related factors, Prompting and selection strategy (4 items) and Critical verification strategy (3 items).

The overall internal consistency of the 15-item scale was acceptable (Cronbach's α = .836). However, the final scale was designed to be interpreted not through a single total score, but through combinations of scores across the four subscales. In this respect, the scale supports multidimensional profiling of learners' generative AI use.

5. Discussion and Conclusion

Table 11. Final Structure of the Task – Strategy Scale for Profiling Generative AI Use

Domain	Sub Factor	Item	Item statement
Task Profile	Text-based Task Use	TA11	I use AI to create drafts.
		TA33	I use AI to summarize information related to the learning content.
		TA34	I use AI to check my understanding of the learning content.
	Media and Data Task Use	TA42	I use AI to create images or visual materials.
		TA43	I use AI to create or edit video materials.
		TA44	I use AI to analyze various forms of data (numerical, textual, or visual) and derive key insights.
		TA45	I use AI to conduct statistical analysis.
		TA13	I use AI to format text or documents (e.g., adjusting font size, margins, and citation styles).
Strategy Profile	Prompting and Selection Strategy	AU11	I specify the purpose or format of the task to the AI to minimize unnecessary follow-up questions.
		AU14	I refine and elaborate my prompts based on the responses provided by generative AI.
		AU31	I can select the most appropriate solution for problem-solving from the various alternatives provided by generative AI.
		AU32	I can select and utilize the most appropriate generative AI tool tailored to the specific characteristics of the task.
	Critical Verification Strategy	AU21	I verify the sources and accuracy of information generated by GenAI.
		AU22	I cross-verify AI-generated content using external sources or reliable evidence.
		AU23	Rather than blindly accepting generative AI responses, I critically check for errors and evaluate their reliability.

The present study aimed to develop and validate a scale for profiling undergraduates' use of generative AI in learning along two axes, task (what learners do) and strategy (how they perform and regulate), at an engineering-oriented university. An initial item pool was developed on the basis of prior research and refined through expert content validity review, with attention to both the clarity of item wording and the comprehensiveness of construct coverage. Main survey data (n = 201) were then analyzed through item analysis, exploratory factor analysis (EFA), confirmatory factor analysis (CFA), and tests of reliability and validity. As a result, the final version of the scale was established as a 15-item instrument with a four-factor structure: media and data task use (5 items), text-based task use (3 items), critical verification strategy (3 items), and prompting and selection strategy (4 items). The implications of these findings are as follows.

At the institutional level, students' generative

AI use needs to be assessed beyond single indicators such as use versus non-use or overall frequency of use. Instead, it should be examined separately in terms of task dimensions, such as text-based learning and writing versus media and data production, and strategy dimensions, such as verification and critical evaluation versus prompt optimization and strategic selection. This is important because generative AI use is not only a matter of efficiency, but is also closely related to learners' metacognitive and regulatory processes.

In practice, this task–strategy distinction has implications for curriculum design and learning support. Learners who score high on media and data task use but low on critical verification strategy may benefit from interventions that preserve the strengths of output-oriented performance while strengthening verification routines, such as source checking, cross-verification, and evidence-based judgment. Conversely, learners who score high on text-based task use but low on prompting and selection strategy may benefit from targeted support in strategic AI interaction, including question design, iterative prompt revision, and principled tool selection.

More broadly, if the four subscale scores are used as the basis for latent profile analysis or cluster analysis rather than being combined into a single total score, it becomes possible to identify distinct learner profiles. For instance, if a profile group in a programming practicum shows high media and data task use but low critical verification strategy, instructors could assign a mandatory code review task requiring students to analyze vulnerabilities in AI-generated code, annotate problematic sections, and submit their evaluations. In this way, instructors may be able to develop teaching and learning guidelines, as well as AI ethics guidance, that are more responsive to the characteristics of different learner groups.

As an evaluation tool, the scale may also be used to assess the effectiveness of generative AI literacy education programs. By administering the scale before and after instruction, researchers and educators can examine how learners' task-related use and strategic use change over time. Such evidence may also provide a basis for curriculum revision and program improvement.

From a profiling perspective, the scale was designed not to rank learners on the basis of a single total score, but to describe their generative AI use through combinations of scores across the four subscales. Future studies may therefore use the four-factor score patterns to derive empirically grounded usage profiles and examine how their distribution and associated educational needs vary by academic major, year of study, and level of achievement. This would further extend the educational usefulness of the scale.

Furthermore, the relatively lower internal consistency observed for text-based task use and prompting and selection strategy is worth noting. This pattern may reflect the functional heterogeneity of student behaviors rather than a simple measurement flaw. Text-based task use includes activities such as drafting, summarization, and comprehension checking, which serve different cognitive purposes, from generating content to monitoring understanding. Similarly, prompting and selection strategy includes both the refinement of language prompts and the selection of appropriate AI tools. These behaviors are related, but they may not always occur together in the same way across students or tasks.

This heterogeneity is especially plausible in engineering-oriented environments, which require students to navigate diverse academic demands from code generation to structured writing. In such contexts, students' generative AI use tends to be task-adaptive rather than a uniform trait; a learner might meticulously refine prompts for coding but rely mostly on tool selection for a literature review. Because these sub-behaviors do not always co-occur uniformly, these two subscales are more appropriately understood as broad behavioral domains. This interpretation aligns with the scale's design as a multidimensional profiling tool, though future research could enhance precision by further delineating these subcomponents.

Several limitations should be acknowledged. First, the data were drawn from a single institution, which limits the generalizability of the findings. Although the sample was drawn predominantly from engineering-related colleges, it also included students from the Business School and the Design Engineering School; accordingly, the findings are most directly applicable to undergraduates at engineering-oriented universities rather than

to engineering majors in the strict disciplinary sense. Future research should revalidate the scale with students from a broader range of institutions and academic fields. Second, EFA and CFA were conducted sequentially on the same sample (n = 201), without cross-validation via sample splitting or an independent replication sample. Future studies should examine the stability of the four-factor structure through cross-validation procedures, such as split-sample validation or replication with an independent sample. Third, the reliability and convergent validity of text-based task use (α = .654, AVE = .397) and prompting and selection strategy (α = .623, AVE = .297) were comparatively low. As discussed above, this likely reflects the fact that the two factors capture relatively broad and task-adaptive behavioral domains; future studies should nevertheless seek to improve measurement precision by expanding the item pool or by differentiating subfactors more clearly. Fourth, the present scale relied on self-reported measures of students' generative AI use. Self-reports capture perceived rather than directly observed behavior, and are therefore subject to recall error and social desirability bias. In the specific context of generative AI use, this concern is particularly salient because behaviors framed as academically desirable, such as critical verification and transparent disclosure, may be over-reported, whereas behaviors such as overreliance on AI-generated outputs may be under-reported. As a result, self-report data may not fully correspond to students' actual use of generative AI. Future research should therefore complement self-report measures with behavioral evidence such as system log data, prompt histories, trace data from learning platforms, and direct observation of task performance. Such behavioral data could, for example, directly corroborate the prompting and selection strategy factor through prompt-history analysis, or validate the critical verification strategy factor through interaction logs, thereby strengthening methodological triangulation and the external validity of the scale.

Despite these limitations, this study contributes to the literature by developing a scale that systematically assesses undergraduates' generative AI use for learning at an engineering-oriented university from an integrated and multidimensional perspective.

참고문헌

[1] Ravšelj, D., Keržič, D., Tomaževič, N., Umek, L., Brezovar, N., et al. (2025). Higher education students' perceptions of ChatGPT: A global study of early reactions. *PLOS ONE, 20*(2), e0315011. https://doi.org/10.1371/journal.pone.0315011

[2] Singer-Freeman, K., Verbeke, K., & Barre, B. (2025). Generative AI usage among university students depends on academic level and task. *Higher Learning Research Communications, 15*(2), 1–24. https://doi.org/10.18870/hlrc.v15i2.1616

[3] Stöhr, C., Ou, A., & Malmström, H. (2024). Perceptions and usage of AI chatbots among students in higher education across genders, academic levels and fields of study. *Computers and Education: Artificial Intelligence, 7*, 100259. https://doi.org/10.1016/j.caeai.2024.100259

[4] Lee, Y. (2023). Analysis of college students' perception about generative artificial intelligence ChatGPT. *The Treatise on The Plastic Media, 26*(4), 46–55. https://doi.org/10.35280/KOTPM.2023.26.4.6

[5] Lee, D., & Palmer, E. (2025). Prompt engineering in higher education: A systematic review to help inform curricula. *International Journal of Educational Technology in Higher Education, 22*(1), 7. https://doi.org/10.1186/s41239-025-00503-7

[6] Gibreel, O., & Arpaci, İ. (2025). *Development and validation of the prompt engineering competence scale (PECS)*. Information Development. Advance online publication. https://doi.org/10.1177/02666669251336455

[7] Kim, C., & Kang, K. (2025). Analysis of college students' recognition and ethical awareness of ChatGPT. *Journal of The Korea Society of Computer and Information, 30*(6), 155–163. https://doi.org/10.9708/jksci.2025.30.06.155

[8] Yan, W., Liu, Y., Mamaeva, V., Dong, F., Tao, G., Li, R., & Yang, H. (2026). Generative AI literacy: Scale development and its influence on privacy protection behaviors and information verification behaviors. *Telecommunications Policy, 50*(2), 103117. https://doi.org/10.1016/j.telpol.2025.103117

[9] Oh, E. (2025). College students' AI utilization and perception. *Journal of the Korean Society for Library and Information Science, 59*(1), 671–692. https://doi.org/10.4275/KSLIS.2025.59.1.671

[10] Kaya, M., & Adıgüzel, T. (2025). Exploring the acceptance of ChatGPT in higher education: A comprehensive quantitative study of university students and faculty. *Frontiers in Education, 10*, 1652292. https://doi.org/10.3389/feduc.2025.1652292

[11] Noh, H., Kim, H., & Ku, Y. (2024). Intention to continue using ChatGPT as a learning tool for college students: Based on the technology acceptance model. *The Journal of the Convergence on Culture Technology, 10*(3), 933–942. https://doi.org/10.17703/JCCT.2024.10.3.933

[12] Ng, D., Wu, W., Leung, J., Chiu, T., & Chu, S. (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
[13] Im, C., Cho, Y., & Jeong, D. (2024). A study on developing of AI literacy scale for college students. *Journal of Information Technology and Architecture, 21*(3), 285–296. https://doi.org/10.22865/jita.2024.21.3.285
[14] Ranieri, M., Biagini, G., & Cuomo, S. (2025). AI literacy in higher education: A systematic approach to questionnaire development and validation. *International Journal of Digital Literacy and Digital Competence, 16*(1), Article 388469. https://doi.org/10.4018/IJDLDC.388469
[15] Fornell, C., & Larcker, D. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research, 18*(1), 39–50. https://doi.org/10.2307/3151312
[16] Henseler, J., Ringle, C., & Sarstedt, M. (2015). A new criterion for assessing discriminant validity in variance-based structural equation modeling. *Journal of the Academy of Marketing Science, 43*(1), 115–135. https://doi.org/10.1007/s11747-014-0403-8
[17] Polit, D., & Beck, C. (2006). The content validity index: Are you sure you know what's being reported? Critique and recommendations. *Research in Nursing and Health, 29*(5), 489–497. https://doi.org/10.1002/nur.20147
[18] Nunnally, J., & Bernstein, I. (1994). *Psychometric theory* (3rd ed.). McGraw-Hill.

이대영

· 2018년 건국대학교 유아교육과(교육학사)
· 2021년 건국대학교 교육공학전공(교육학석사)
· 2024년 건국대학교 교육공학전공(교육학박사)
· 2024년~현재 한국공학대학교 전공교육혁신센터 연구교수

관심분야 : AI, 빅데이터분석, 에듀테크, 핵심역량
2vs0@tukorea.ac.kr

유 정

· 2017년 중국 베이징외국어대학교 컴퓨터학과(공학사)
· 2019년 성균관대학교 교육학전공(교육학석사)
· 2024년 성균관대학교 고등교육전공(교육학박사)
· 2025년~현재 한국공학대학교 전공교육혁신센터 연구교수

관심분야 : AIED, 이러닝, 고등교육정책, 고등교육 국제화, 비교교육
liuting0113@tukorea.ac.kr