

Profiles of generative AI chatbot use for schoolwork and achievement among Indonesian 15-year-olds: A latent class and multilevel analysis of PISA 2025 data

Umar Yeni Suyanto

umaryenisuyanto@uny.ac.id

Universitas Negeri Yogyakarta

Nur Rahmi Akbarini

Universitas Sebelas Maret

Febrika Yogie Hermanto

Universitas Negeri Surabaya

Indra Febrianto

Universitas Negeri Yogyakarta

Dilla Octavianingrum

Institut Seni Indonesia Yogyakarta

Research Article

Keywords: Generative artificial intelligence, Chatbots, Latent class analysis, PISA 2025, Secondary education, Indonesia, Multilevel models

Posted Date: October 20th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11408706/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Background

Secondary students now use generative artificial intelligence chatbots for schoolwork, yet evidence on how this use relates to learning comes mainly from short experiments with university students. Large-scale assessment data from middle-income countries have not been used to examine it.

Methods

We analyzed data from the 2025 cycle of the Programme for International Student Assessment for 14,168 Indonesian 15-year-olds in 419 schools, representing about 4.0 million students. Weighted latent class analysis of four purpose-specific items identified profiles of chatbot use for schoolwork. We related the profiles to science and reading achievement in two-level models that used sampling weights, replicate weights, and ten plausible values per domain, controlled for student and school background, and tested whether metacognitive strategy use moderates the associations. We examined school practice in evaluating generated content as an exploratory question.

Results

Four profiles emerged: rare users (28.3%), selective learning-support users (35.0%), uniform monthly users (7.2%), and intensive users (29.5%). The profiles differed in how often students used chatbots rather than in the purposes they pursued, so the hypothesized offloading profile did not emerge. Rare users were more often enrolled in lower secondary and private schools, had lower socioeconomic status, and reported less frequent metacognitive strategy use. After background was controlled, rare users scored about 7 points (0.1 standard deviation) lower than selective users in science and reading, whereas intensive users did not differ significantly from selective users. Metacognitive strategy use did not moderate these associations, and an exploratory interaction between rare use and school practice did not hold when school composition was controlled.

Conclusions

Intensive chatbot use was not associated with lower achievement among Indonesian 15-year-olds. The small disadvantage of rare users may reflect selection rather than an effect of use, but it suggests that schools need to look beyond frequent users when they consider how to guide chatbot use. The cross-sectional design and self-reported use limit causal interpretation.

1. Introduction

Since the public release of ChatGPT in November 2022, educators have weighed the promise of large language models for personalized support against their risks to academic integrity and independent thinking (Kasneci et al., 2023; Yan et al., 2024). Meanwhile, students have already made these tools part of their schoolwork. In Indonesia, 53% of 15-year-olds reported in the Programme for International Student Assessment (PISA) 2025 that they used artificial intelligence (AI) chatbots at least once a week to help them learn, and between 30% and 39% used them weekly to research a new topic, summarize an assigned text, or draft a writing assignment (OECD, 2026a). For teachers, the practical question now concerns how students use these tools and with what consequences for learning.

Experimental evidence on that question points in two directions. A systematic review of 69 experimental studies, 62 of which were meta-analyzed, found that ChatGPT interventions improved academic performance, affective-motivational states, and higher-order thinking propensities while reducing mental effort (Deng et al., 2025). Other studies show that the same tools can raise immediate performance while weakening learning. In a field experiment with nearly 1,000 high school students in Turkey, unrestricted access to GPT-4 raised grades on practice problems, yet students performed 17% worse than controls once access was removed; a tutor version with learning safeguards largely mitigated this loss (Bastani et al., 2025). University students who revised essays with ChatGPT obtained higher essay scores but retained no more knowledge than peers who received other forms of support, a pattern the authors called metacognitive laziness (Fan et al., 2025).

Two theoretical lenses help explain this divergence. From the perspective of self-regulated learning, technologies can support students' planning, monitoring, and evaluation of their learning or take over these functions, which may hinder the development of metacognitive skills (Molenaar, 2022; Zimmerman, 2002). Research on cognitive offloading adds that external tools reduce the cognitive demands of a task, often improving performance while diminishing memory for the offloaded content (Grinschgl et al., 2021; Risko & Gilbert, 2016). A chatbot that summarizes a reading or drafts an essay may bypass the processing through which students build understanding. The consequences of chatbot use should therefore depend on the purposes students pursue, on their capacity to regulate their learning, and on whether teachers prepare them to evaluate AI output critically, a core component of AI literacy (Casal-Otero et al., 2023; Ng et al., 2021).

Three gaps limit what educators can draw from current research. First, most studies of students' generative AI use involve university students, small samples, or short interventions (Chan & Hu, 2023; Fan et al., 2025; Stojanov et al., 2024; Strzelecki, 2024). Of the 69 experiments reviewed by Deng et al. (2025), 58 were conducted at the university level, and primary and secondary education accounted for only about 10% of highly cited studies of AI in education (Brezovec et al., 2026). Surveys of secondary students describe how often and why they use chatbots (Annuš, 2025; Callado et al., 2025; Weinhandl et al., 2026) but have not linked this use to measured achievement, and representative analyses from middle-income countries have not been available. Second, studies using PISA data have treated information and communication technology (ICT) mainly in terms of access and frequency of use, and they have often found null or negative associations between ICT use and achievement (Hu et al., 2018; Tan et al., 2026), whereas students' interest and autonomy in ICT related positively to achievement (S. C. Li & Petersen, 2022). Petko et al. (2017) found that the perceived quality of technology use matters more than its quantity, yet purpose-based profiles of AI use have so far been examined only among university students (Stojanov et al., 2024) and in a convenience sample of Swedish secondary students (Hugerth & Hugerth, 2026). Third, a systematic review warns that over-reliance on AI dialogue systems can affect students' cognitive abilities (Zhai et al., 2024), but large-scale studies have not tested whether students' self-regulation or classroom practice in judging AI-generated content changes that relationship.

This study addresses these gaps with data from PISA 2025, which asked 15-year-olds how often they use AI chatbots for specific schoolwork purposes and how often they assessed the quality of AI-generated information in their lessons (OECD, 2026b). We analyze responses from 14,168 students in 419 Indonesian schools, a sample that represents about 4.0 million 15-year-olds. Indonesia is an informative case. Its students' mean scores of 389 in science and 365 in reading lie more than 90 points below the averages of the Organisation for Economic Co-operation and Development (OECD) of 482 and 461 (OECD, 2026b), and its schools must respond to widespread chatbot use with little local evidence to guide them. The study asks four research questions:

RQ1: What profiles of AI chatbot use for schoolwork emerge among Indonesian 15-year-olds, and which student and school characteristics predict membership in each profile?

RQ2: How are these profiles associated with students' science and reading achievement after accounting for student and school background?

RQ3: Does students' self-regulated learning moderate these associations?

RQ4 (exploratory): Is the frequency with which students in a school practice evaluating AI-generated content related to differences in achievement between these profiles?

The study contributes nationally representative evidence that links secondary students' reported chatbot use to measured achievement in a middle-income country, a design-based latent class analysis of whether students differ in the purposes or mainly the frequency of their use, a worked example of design-based latent class and multilevel analysis with replicate weights and plausible values that other secondary analysts of PISA can reuse, a test of whether self-regulated learning moderates these associations, and an exploratory analysis of whether school practice in evaluating AI-generated content is related to them.

2. Literature review

2.1. Generative AI and the regulation of learning

Self-regulated learners set goals, choose strategies, monitor their progress, and adjust their approach when it fails. Zimmerman (2002) organized these activities into three cyclical phases: forethought, performance, and self-reflection. Most other models describe a similar cycle of preparatory, performance, and appraisal phases (Panadero, 2017). The construct matters for achievement. A meta-analysis of studies with school-age students found positive associations between self-regulated learning and academic achievement (Dent & Koenka, 2016), and in online learning, metacognition, time management, and effort regulation related to achievement more consistently than rehearsal or elaboration (Broadbent & Poon, 2015).

Generative AI changes who performs these regulatory functions. Molenaar (2022) described levels of automation that divide control over learning among teachers, learners, and AI. The model distinguishes scaffolds that support learners' own regulation from technologies that regulate learning for them, and Molenaar warned that such offloading may harm the development of metacognitive skills. Järvelä et al. (2023) argued that AI should strengthen learners' own regulation rather than take it over. Recent experiments suggest that AI assistance can take over regulation. In a trial with 1,625 university students, Darvishi et al. (2024) found that students tended to rely on rather than learn from AI assistance in peer feedback tasks, and adding self-monitoring checklists did not improve performance. In essay writing, ChatGPT support altered students' self-regulatory processes without improving knowledge retention (Fan et al., 2025). These studies suggest that students' capacity for self-regulation may shape whether chatbot use supports or replaces their learning. Classroom trials with secondary students add nuance. Chatbot prompts designed to support self-regulated learning gave German students in Grades 7 to 9 no advantage over standard ChatGPT in strategy use, effort, or knowledge, and some students mainly sought correct answers (Fütterer et al., 2026). In Grades 9 and 10, self-reported self-regulation skills did not moderate the effects of brief chatbot guidance (Klar, 2025). Whether self-regulated learning shapes the consequences of everyday chatbot use for schoolwork remains untested.

2.2. Cognitive offloading and the purposes of chatbot use

Cognitive offloading refers to the use of physical action, including the use of external tools, to reduce the cognitive demands of a task (Risko & Gilbert, 2016). Offloading has costs. When people expected to retain access to information online, they remembered where to find it better than the information itself (Sparrow et al., 2011), and offloading in laboratory tasks improved performance while diminishing memory for the offloaded content (Grinschgl et al., 2021). Generative AI extends offloading from storing information to producing it. In a survey of 666 participants, more frequent AI tool use was associated with weaker critical thinking, and this association was mediated by cognitive offloading (Gerlich, 2025). A systematic review of 14 studies concluded that over-reliance on AI dialogue systems affects students' cognitive abilities because users come to favor fast solutions over slower, more effortful ones (Zhai et al., 2024). In an

experiment with 91 university students, those who used ChatGPT rather than a search engine to gather information reported less mental effort and produced lower-quality reasoning (Stadler et al., 2024).

Not all chatbot use involves the same degree of offloading. Students report using chatbots for writing and brainstorming, for research and analysis, and for personalized learning support (Chan & Hu, 2023). Secondary students draw a similar line, judging chatbot use for reference, brainstorming, or clarification acceptable but copying answers as cheating (Chan, 2026). These purposes differ in how much of the cognitive work they transfer to the tool. Asking a chatbot to explain a concept or suggest starting points for a new topic can leave the student responsible for selecting and integrating information. Asking it to summarize an assigned reading or draft a written assignment transfers the processes through which students construct understanding. The direction of influence may also run the other way. Students under greater academic workload and time pressure use ChatGPT more, and that use is associated with procrastination, memory loss, and lower academic performance (Abbas et al., 2024).

2.3. AI literacy and classroom practice

AI literacy encompasses knowing and understanding AI, using and applying it, evaluating and creating with it, and attending to its ethical implications (Ng et al., 2021). The capacity to evaluate AI output is particularly relevant to chatbots, which can produce fluent but inaccurate content that users may over-trust (Park & Nan, 2026). Pritchard (2026) argues that accepting such output without checking it is not a route to knowledge, yet most Portuguese secondary students surveyed before any classroom instruction on ChatGPT judged its outputs accurate (Callado et al., 2025). A systematic review of AI literacy in K-12 education identified a growing number of classroom initiatives but found that few assessed whether students actually understood AI concepts after instruction (Casal-Otero et al., 2023). Zhong and Liu (2025) developed a validated framework and scale for the AI literacy of secondary students. Classroom practice is also where teachers have direct influence. Bastani et al. (2025) showed that a chatbot designed with pedagogical safeguards largely prevented the learning losses that unrestricted access produced. Training students to assess the quality of AI-generated information may serve as a comparable safeguard that teachers can provide without redesigning the tool, but whether it is related to the consequences of chatbot use has not been examined with large-scale data.

2.4. Technology use and achievement in large-scale assessments

Experimental syntheses report positive effects of AI chatbots on learning, mostly from university samples. A meta-analysis of 24 randomized studies found larger effects in higher education than in schools and in shorter interventions, partly reflecting novelty effects (Wu & Yu, 2024). Studies of everyday technology use in representative samples tell a more cautious story. Using PISA 2015 data from 305,414 students in 44 countries, Hu et al. (2018) found that students' academic use of ICT was negatively associated with performance. S. C. Li and Petersen (2022) showed that autonomy and interest in ICT related positively to achievement, whereas the provision of technology in schools did not. Petko et al. (2017) found that students' positive perceptions of educational technology, rather than the frequency of its use, were associated with higher PISA 2012 scores in most participating countries.

Person-centered approaches have begun to describe how students combine different forms of technology use. Applying multilevel latent profile analysis to PISA 2022 data from 302,592 students, Tan et al. (2026) found that students with low ICT use had the highest academic achievement and those with high use had the lowest. For generative AI specifically, Stojanov et al. (2024) identified five profiles of university students' reliance on ChatGPT, ranging from versatile low reliers (38.2%) to assignment delegators (23.1%). Among university language learners, profiles of self-efficacy in AI-assisted writing differed markedly in self-regulated strategy use (Liu & Zhang, 2025). Hugerth and Hugerth (2026) applied latent class analysis to survey data from 1,266 Swedish secondary students and distinguished profiles of AI use oriented toward seeking knowledge from those oriented toward efficiency. To our knowledge, no study has applied this approach to nationally representative data or linked AI use profiles to measured achievement.

2.5. Hypotheses

Drawing on these frameworks, we formulated three hypotheses, which were not preregistered.

- H1: At least three profiles of AI chatbot use will emerge: infrequent users, exploratory users who use chatbots mainly for preliminary research and learning support, and offloading users who use chatbots mainly to summarize readings and draft written work. - H2: Compared with the exploratory profile, the offloading profile will show lower science and reading achievement after student and school background are controlled. - H3: The negative association between the offloading profile and achievement will be weaker among students with stronger self-regulated learning. Research on practice in evaluating AI-generated content is too sparse to support a directional hypothesis. We therefore examined, as an exploratory question (RQ4), whether the frequency with which students in a school practice evaluating AI-generated content is related to differences in achievement between the profiles.

3. Methods

3.1. Data and sample

We used the Indonesian data from PISA 2025 (OECD, 2026b). PISA assesses students aged between 15 years 3 months and 16 years 2 months who are enrolled in grade 7 or above. Its two-stage stratified design first samples schools with probability proportional to size and then samples students within each school. The Indonesian sample comprised 14,168 students in 419 schools and represents about 4.0 million 15-year-olds. Weighted estimates show that 52.2% of students were female, 44.9% attended lower secondary schools, 34.4% attended general upper secondary schools, and 20.7% attended vocational upper secondary schools; 38.8% attended private schools. Indonesia did not administer the optional ICT familiarity questionnaire or the teacher questionnaire, so all measures came from the student questionnaire, the school questionnaire, and the cognitive assessment. The study used de-identified public-use files and required no additional ethical approval or consent to participate.

3.2. Measures

AI chatbot use. Students reported how often they used "Artificial Intelligence chatbots (e.g., ChatGPT) for your school work" for four purposes: to summarize a text they had to read, to conduct preliminary research on a new topic, to draft texts for writing assignments, and to help them learn (items ST438Q01–Q04). Responses ranged from 1 (*never or almost never*) to 5 (*every day or almost every day*). We recoded each item into three categories: rarely (never, or once or twice a year), monthly (once or twice a month), and weekly or more often. This coding distinguishes occasional from habitual use; the original five-category and a binary coding served as robustness checks. PISA also provides a scaled index of these items (AIUSESCH), which we used in a robustness check. Item nonresponse ranged from 3.4% to 4.8%. The 447 students (3.2%) who answered none of the four items were excluded from the latent class analysis, which left an analytic sample of 13,721.

Achievement. Science and reading achievement were each represented by 10 plausible values (PV1SCIE–PV10SCIE and PV1READ–PV10READ). The Indonesian weighted means were 388.7 (standard deviation [SD] = 67.1) for science and 365.2 (SD = 69.4) for reading.

Self-regulated learning. We used two weighted likelihood estimate (WLE) indices from the student questionnaire. The index labeled GOALSET in the public-use file is built from four items on metacognitive strategy use (ST432): planning one's approach to studying, summarizing key points without notes, asking oneself questions to check understanding, and changing one's approach when it is not working. The index labeled SELFREG is built from four items on help seeking (ST433), such as asking teachers or classmates for help. In the Indonesian data, the two indices correlated .99 and .97 with the means of their four items ($\alpha = .80$ for both item sets), so we refer to them as metacognitive strategy use and help seeking. Because metacognitive strategy use maps onto the planning, monitoring, and adjustment processes in Zimmerman's (2002) model, we designated it as the primary moderator for H3, with help seeking as a secondary

moderator. Both indices were standardized with student weights. They describe general study strategies, not regulation during AI-assisted tasks.

Practice in evaluating AI content. Students reported how often during the current school year they had used digital resources in their lessons to "assess the quality of information generated by Artificial Intelligence chatbots" (ST436Q16), on a 4-point scale from *never or almost never* to *very often*. We aggregated responses to the school level as the weighted mean of students' reports, which represents the typical frequency of this practice in each school (school practice), and centered the school means on the grand mean. The means rested on a median of 30 respondents per school, although 28 of 413 schools had fewer than five; the intraclass correlations were $ICC(1) = .12$ and $ICC(2) = .79$ (Lüdtke et al., 2008). It reflects students' reports of classroom exposure, not observed teaching.

Covariates. Student-level covariates were gender, grade relative to the modal grade, study program (lower secondary, general upper secondary, or vocational upper secondary), the PISA index of economic, social, and cultural status (ESCS), and daily hours spent using digital devices at school for learning and for leisure (ST326Q01 and ST326Q04, coded at category midpoints). School-level covariates were the school mean of ESCS, school type (public or private), and location (rural area or small town versus town or city of 15,000 or more inhabitants). Curiosity (CURIO) and growth mindset (ST420Q01) entered only the robustness checks.

3.3. Analytic strategy

Sampling weights and plausible values. All estimates used the final student weight. Except in the multilevel models, standard errors came from the 80 balanced repeated replication weights with Fay's adjustment ($k = 0.5$). Analyses involving achievement were run for each plausible value and combined with Rubin's rules (Jerrim et al., 2017; Rutkowski et al., 2010). The procedure reproduced the estimates the OECD published for Indonesia (OECD, 2026a; Appendix A.1).

Latent class analysis (RQ1). We used latent class analysis because H1 concerns differences in the purposes of use and not only in frequency; mean frequency served as a comparison. We estimated latent class models with one to six classes, using the four AI use items as categorical indicators, by weighted pseudo-maximum likelihood; incomplete response patterns were retained under full information maximum likelihood (Appendix A.1). Following Nylund et al. (2007), Nylund-Gibson and Choi (2018), and Weller et al. (2020), we selected the number of classes by considering the Bayesian information criterion (BIC), the sample-size-adjusted BIC, relative entropy, the bootstrap likelihood ratio test (BLRT) with 100 bootstrap samples, the size of the smallest class (at least 5%), the average posterior probability of each class, bivariate residuals between pairs of items, and the interpretability of the profiles.

To identify which students belonged to each profile, we used the bias-adjusted three-step approach with maximum likelihood correction (Bakk et al., 2013; Vermunt, 2010), which is equivalent to the R3STEP procedure in Mplus (Asparouhov & Muthén, 2014). The predictors were gender, study program, school type, ESCS, metacognitive strategy use, help seeking, curiosity, and daily hours of digital device use for learning and for leisure. Mean achievement in each profile was estimated with the weights proposed by Bolck, Croon, and Hagenaars (BCH; Bakk et al., 2013; Bolck et al., 2004), combined across plausible values and replicate weights. We implemented the estimation in Python and cross-checked the selected solution against the StepMix package (Morin et al., 2025), which produced identical log-likelihood values and class proportions.

Multilevel models (RQ2–RQ4). We fitted two-level random-intercept models with students nested in schools. The school weight entered at level 2 and the conditional student weight, scaled to the cluster size, at level 1 (Mang et al., 2021; Rabe-Hesketh & Skrondal, 2006). We estimated five models for each outcome. M0 was an empty model used to compute the intraclass correlation. M1 added dummy variables for the AI use profiles, with the selective learning-support profile as the reference category. We chose this profile because it is closest to the exploratory use described in H1. Because the hypothesized offloading profile did not emerge (see Results), we tested H2 and H3 through the closest available contrast,

that of the intensive profile, whose members most often used chatbots to summarize readings and draft texts, with the selective profile, a choice we made after the profile solution was known. The contrast of rare with selective users is an additional, unhypothesized comparison. M2 added the student and school covariates. M3 added interactions between the profiles and metacognitive strategy use (RQ3). M4 added cross-level interactions between the profiles and school practice in evaluating AI content, which we treated as an exploratory analysis (RQ4). Metacognitive strategy use was centered on the school mean, and the school mean was added at level 2, so that the M3 interactions reflect within-school differences (Enders & Tofighi, 2007). Main analyses assigned students to their most likely profile; Appendix A.2 describes checks that account for classification error. We tested each set of profile terms jointly with the D1 statistic for multiply imputed data (K. H. Li et al., 1991) and adjusted the profile-interaction tests with the Benjamini–Hochberg procedure (Benjamini & Hochberg, 1995). All models used the same analytic sample of 12,695 students in 413 schools with complete data on the covariates and moderators. We estimated the models by weighted pseudo-maximum likelihood with the WeMix package (Version 4.0.4; Bailey et al., 2024) in R (Version 4.3.3; R Core Team, 2024), which provides robust standard errors that account for the clustering of students within schools. For significant interactions, we computed simple slopes and profile differences at one standard deviation below and above the mean of the moderator, with standard errors derived from the covariance matrix of the fixed effects.

Missing data. Item nonresponse on student covariates ranged from 2.5% (ESCS) to 5.9% (metacognitive strategy use) and was handled by listwise deletion, which left 12,652 students for the three-step model and 12,695 students in 413 schools for the multilevel models. Nonresponse on the practice item was higher (22.1% of students, weighted), rose across the items of its question block, and was more common among lower-achieving students (reading: 328.0 vs. 375.8). We therefore computed the school means from the available responses and added the school proportion of nonresponse, centered on the grand mean, as a covariate in M4. Multiple imputation with school indicators served as a check (Grund et al., 2018; Appendix A.2).

Robustness checks. Appendix A.2 describes the checks on the profile differences, labeled (a) to (j), and on the practice interaction, labeled (h1) to (h7); mathematics served as an additional outcome (Tables A2 and A3). The checks include controls for school composition (Bell et al., 2019) and a random slope for rare users (Heisig & Schaeffer, 2019).

3.4. Use of generative AI

The authors used AI tools to assist with this study. ChatGPT (OpenAI) was used to assist with language editing and clarity. The authors designed the study, made the key analytic decisions, reviewed all code, outputs, and text, verified the procedures by reproducing estimates published by the OECD, and take full responsibility for the content.

4. Results

Six in ten students (60.0%) used AI chatbots for schoolwork at least weekly for at least one of the four purposes, whereas 7.5% never used them for any of the four. Help with learning was the most common weekly purpose (53.5%), followed by summarizing texts (38.8%), preliminary research (31.7%), and drafting texts (30.3%). The four items were moderately to strongly correlated (weighted Spearman $\rho = .49$ to $.63$). Table 1 reports the fit of models with one to six classes. The BIC and aBIC decreased with each additional class, but the decrease flattened after four classes. The BLRT, with 100 bootstrap samples, reached its smallest attainable value ($p = .010$) at every step and did not discriminate between models. The five- and six-class solutions had lower entropy and lower minimum average posterior probabilities, and their additional classes split existing profiles. We therefore retained the four-class solution, whose best log-likelihood was replicated in all 10 of the best starts (entropy = $.788$; average posterior probabilities = $.83$ to $.92$). Its bivariate residuals ranged from 1.4 to 7.6, with three of six item pairs above 3.84 (Table A4), so some association between items remained; the five-class model reduced these residuals but split an existing profile.

Table 1
Fit statistics for latent class models with one to six classes ($n = 13721$)

Classes	LL	Parameters	BIC	aBIC	Δ BIC	Entropy	Smallest class (%)	Min. AvePP	BLRT p
1	-55954.6	8	111985.4	111960.0	—	—	100.0	—	—
2	-47307.7	17	94777.4	94723.4	-17208.0	.824	45.2	.93	.010
3	-44208.9	26	88665.4	88582.8	-6112.0	.790	28.8	.87	.010
4	-43611.2	35	87555.9	87444.7	-1109.5	.788	7.2	.83	.010
5	-43394.1	44	87207.3	87067.5	-348.6	.753	7.0	.77	.010
6	-43289.9	53	87084.6	86916.2	-122.7	.701	5.0	.65	.010
Note. LL = log-likelihood; BIC = Bayesian information criterion; aBIC = sample-size-adjusted BIC; ΔBIC = change from the model with one fewer class; AvePP = average posterior probability; BLRT = bootstrap likelihood ratio test of K versus K - 1 classes (100 bootstrap samples; .010 is the smallest attainable p value). The selected solution is shown in bold.									

Figure 1 displays the four profiles. Rare users (28.3%) almost never used chatbots for preliminary research, summarizing, or drafting (96% to 98% in the rarely category) and seldom used them to help them learn (84% rarely). Selective learning-support users (35.0%) used chatbots mainly to help them learn (62% weekly or more often) and less often to summarize texts (30%), conduct preliminary research (19%), or draft texts (18%). Uniform monthly users (7.2%) reported monthly use for all four purposes (87% to 95%), and intensive users (29.5%) reported weekly or more frequent use for all four purposes (82% to 99%).

The profiles therefore differed mainly in how often students used chatbots rather than in what they used them for. Help with learning was the most frequent purpose in every profile, and no profile combined frequent summarizing and drafting with infrequent use for learning. H1 was thus partly supported: more than three profiles emerged, but none matched the hypothesized offloading pattern. This frequency-based structure held with the original five categories, with a binary coding, and after the 32.7% of students (weighted) who gave identical answers to all four items were excluded; identical answers were most common among uniform monthly users (72%). Because no offloading profile emerged, we tested H2 and H3 with the contrast between intensive and selective users; the contrast between rare and selective users was not anticipated.

Table 2 shows who belonged to each profile. Relative to lower secondary students, students in general and vocational upper secondary programs had higher odds of being selective or intensive rather than rare users (odds ratios [ORs] = 1.59 to 2.28), whereas students in private schools had lower odds (ORs = 0.61 and 0.50). Higher ESCS and more frequent metacognitive strategy use predicted membership in all three profiles other than rare users (ORs = 1.21 to 1.33 per SD), but help seeking did not. Female students were more likely than male students to be selective users (OR = 1.40), and curiosity predicted membership in the intensive and uniform monthly profiles. Rare users also had less exposure to practice in evaluating AI content: 52.1% had practiced it at least sometimes, compared with 78.3% to 84.4% of students in the other profiles. Before background characteristics were taken into account, rare users had the lowest achievement. They scored 23.7 and 25.6 points lower in science than selective and intensive users, respectively, and 26.8 and 28.4 points lower in reading (all $ps < .001$), or 0.35 to 0.41 Indonesian SD.

Table 2
Predictors of profile membership and mean achievement by AI use profile

Profile size, % (SE)	Rare users	Selective learning-support users	Uniform monthly users	Intensive users
	28.3 (1.2)	35.0 (1.1)	7.2 (0.8)	29.5 (1.1)
Odds ratio [95% CI] vs. rare users				
Female	ref.	1.40** [1.08, 1.80]	0.73 [0.46, 1.16]	1.10 [0.92, 1.30]
General upper secondary (vs. lower secondary)	ref.	1.59** [1.17, 2.17]	0.97 [0.67, 1.39]	1.81*** [1.38, 2.38]
Vocational upper secondary (vs. lower secondary)	ref.	1.70** [1.23, 2.36]	0.97 [0.59, 1.59]	2.28*** [1.70, 3.05]
Private school	ref.	0.61*** [0.48, 0.78]	0.67 [0.41, 1.10]	0.50*** [0.40, 0.63]
ESCS (per SD)	ref.	1.21** [1.06, 1.39]	1.28* [1.04, 1.57]	1.27*** [1.15, 1.41]
Metacognitive strategy use (per SD)	ref.	1.21** [1.08, 1.35]	1.22* [1.05, 1.43]	1.33*** [1.20, 1.47]
Help seeking (per SD)	ref.	1.04 [0.92, 1.18]	0.98 [0.85, 1.13]	1.05 [0.96, 1.15]
Curiosity (per SD)	ref.	1.11 [0.98, 1.25]	1.30* [1.03, 1.65]	1.35*** [1.24, 1.48]
Daily hours of digital learning at school	ref.	1.01 [0.97, 1.04]	1.06 [0.99, 1.13]	1.08*** [1.04, 1.12]
Daily hours of digital leisure at school	ref.	1.12** [1.05, 1.20]	1.06 [0.95, 1.17]	1.11*** [1.05, 1.18]
Practiced evaluating AI content at least sometimes, % (SE)	52.1 (2.1)	78.3 (1.6)	82.5 (5.0)	84.4 (1.3)
Mean achievement, BCH-weighted (SE)				
Science	372.4 (3.5)	396.0 (4.5)	396.7 (12.7)	398.0 (2.8)
Reading	347.1 (3.7)	373.9 (4.7)	371.8 (13.7)	375.6 (3.0)
<i>Note.</i> Odds ratios come from the bias-adjusted three-step multinomial regression with maximum likelihood correction (Vermunt, 2010), $n = 12652$; confidence intervals are based on balanced repeated replication. Profile sizes ($n = 13721$) are model-estimated class proportions. The share of students who practiced evaluating AI content (respondents to the item only) and mean achievement are BCH-weighted estimates (Bakk et al., 2013; Bolck et al., 2004) that account for classification error; mean achievement combines 10 plausible values. *$p < .05$. **$p < .01$. ***$p < .001$.				

Achievement varied considerably between schools: 37.3% of the variance in science and 40.8% in reading lay between schools. In the first multilevel model (M1), rare users scored 6.6 points lower than selective users in science and 6.9 points lower in reading (Table 3). These gaps were much smaller than in single-level regressions without covariates (21.0 and 23.3 points), because rare users were concentrated in lower-achieving schools. The gap remained after the student and school covariates were added in M2 (science: -7.3 points, $p = .008$; reading: -7.1 points, $p = .009$; about 0.1 Indonesian SD), and the three profile contrasts were jointly significant (science: $p = .019$; reading: $p = .025$; Table A3). Daily hours of digital device use at school were negatively associated with both outcomes (Table A1), and the covariates explained about half of the between-school variance.

Table 3
Two-level models of science and reading achievement

Predictor	M1	M2	M3	M4
Science				
Intercept	374.76*** (4.69)	467.00*** (13.66)	470.49*** (14.27)	462.97*** (9.25)
Rare users	-6.61* (2.77)	-7.33** (2.73)	-6.95* (2.73)	-10.82*** (2.75)
Uniform monthly users	3.56 (4.43)	3.78 (4.32)	3.87 (4.31)	2.45 (4.34)
Intensive users	-2.52 (2.66)	-1.30 (2.61)	-1.70 (2.60)	-2.07 (2.51)
Metacognitive strategy use (within school)			2.35 (1.88)	
Metacognitive strategy use (school mean)			-14.50 (10.15)	
Rare × metacognitive strategy use			-0.49 (2.60)	
Monthly × metacognitive strategy use			-3.83 (4.71)	
Intensive × metacognitive strategy use			0.06 (2.77)	
School practice in evaluating AI content				28.75*** (8.12)
School proportion of nonresponse on practice item				-132.69*** (11.57)
Rare × school practice				-23.65** (8.21)
Monthly × school practice				-16.00 (13.45)
Intensive × school practice				-7.92 (7.06)
Student and school covariates	No	Yes	Yes	Yes
School-level variance (τ_{00})	1730.8	892.0	877.5	396.6
Student-level variance (σ^2)	2937.7	2843.8	2837.6	2824.3
Variance explained, school level (%)	1.4	49.2	50.0	77.4
Variance explained, student level (%)	0.3	3.5	3.7	4.1
Reading				
Intercept	349.29*** (4.91)	436.99*** (14.06)	440.81*** (14.76)	433.34*** (9.28)
Rare users	-6.90* (2.82)	-7.11** (2.69)	-6.57* (2.71)	-10.35*** (2.73)
Uniform monthly users	1.88 (4.27)	3.66 (4.05)	3.62 (4.04)	2.69 (4.01)
<i>Note.</i> $n = 12695$ students in 413 schools. Unstandardized coefficients with standard errors in parentheses, combined across 10 plausible values. The reference profile is selective learning-support users. Covariate estimates are in Table A1. School practice is the school mean of a 1–4 scale ($SD = 0.366$) and nonresponse a proportion, both centered on the grand mean. *$p < .05$. **$p < .01$. ***$p < .001$.				

Predictor	M1	M2	M3	M4
Science				
Intensive users	-2.33 (2.74)	-0.90 (2.61)	-1.15 (2.58)	-1.50 (2.53)
Metacognitive strategy use (within school)			2.48 (1.80)	
Metacognitive strategy use (school mean)			-15.54 (11.06)	
Rare × metacognitive strategy use			0.24 (2.42)	
Monthly × metacognitive strategy use			-2.92 (4.17)	
Intensive × metacognitive strategy use			-0.86 (2.56)	
School practice in evaluating AI content				28.80*** (8.37)
School proportion of nonresponse on practice item				-144.52*** (12.63)
Rare × school practice				-21.82** (8.05)
Monthly × school practice				-9.99 (13.25)
Intensive × school practice				-7.05 (6.91)
Student and school covariates	No	Yes	Yes	Yes
School-level variance (τ_{00})	2042.3	1045.1	1026.3	455.8
Student-level variance (σ^2)	2992.6	2799.7	2793.5	2781.8
Variance explained, school level (%)	1.3	49.5	50.4	78.0
Variance explained, student level (%)	0.3	6.7	6.9	7.3
<i>Note.</i> $n = 12695$ students in 413 schools. Unstandardized coefficients with standard errors in parentheses, combined across 10 plausible values. The reference profile is selective learning-support users. Covariate estimates are in Table A1. School practice is the school mean of a 1–4 scale ($SD = 0.366$) and nonresponse a proportion, both centered on the grand mean. * $p < .05$. ** $p < .01$. *** $p < .001$.				

Neither intensive users ($ps > .61$) nor uniform monthly users ($ps > .36$) differed from selective users, so H2, tested through the intensive–selective contrast, was not supported. Neither the mean frequency of use nor AIUSESCH was associated with achievement ($ps > .13$). The rare-user gap held in most robustness checks (Table A2). It was unchanged when the school shares of each profile were added (-7.3 and -7.1), and it was 5.5 to 7.0 points with curiosity and growth mindset added, with pseudo-class draws, with the three-class solution, and without the digital-hours covariates. Regressions with school fixed effects, which compare students within the same school, gave larger gaps (-11.4 and -11.8 points; 11.4 to 15.0 points across specifications, particularly with BCH weights), as did single-level regressions (14.4 to 17.9 points). One check weakened the result: when the 31.6% of students in the multilevel sample who gave identical answers to all four items were excluded, the gap was smaller and not statistically significant (science: -5.8 , $p = .058$; reading: -5.0 , $p = .124$). Among rare users, identical answers meant never or only once or twice a year for every purpose, so this check removed many of the students the profile was meant to capture.

Metacognitive strategy use did not moderate the associations between profiles and achievement. The interactions in M3 were not significant, individually (all $ps > .41$) or jointly ($p = .82$ in science and $.88$ in reading), and metacognitive strategy use was not associated with achievement within schools ($ps > .17$); H3 was therefore not supported. In the secondary moderator analysis, help seeking was positively associated with achievement among selective users, and this association was weaker among rare users (science: $b = -7.62$; reading: $b = -7.87$; $ps < .01$; Benjamini–Hochberg adjusted $ps = .03$). This result was not anticipated and is exploratory.

The exploratory analysis of school practice in evaluating AI content (M4) pointed to a more complex pattern. School practice was positively associated with the achievement of selective users ($b = 28.75$ in science and 28.80 in reading per scale point, $ps < .001$), or 10.5 points per SD of the school mean. The interaction between the rare profile and school practice was negative (science: $b = -23.65$; reading: $b = -21.82$; $ps < .01$), whereas the interactions for intensive and uniform monthly users were not significant; jointly, the three interaction terms were significant in science ($p = .030$) but not in reading ($p = .073$). Simple slopes (Fig. 2) were positive for selective users (10.5 points per SD) and intensive users (7.6 and 8.0) but not for rare users (1.9 and 2.6 , $ps > .39$), so the rare-selective gap was about 2 points (not significant) one SD below the mean of practice and 18 to 19 points one SD above it. The difference between intensive and selective users did not vary with practice.

The rare $\times$ practice interaction changed little without the nonresponse covariate, with a peer mean that excluded each student's own report, with a random slope for rare users, in schools with at least five respondents, and with multiply imputed responses ($b = -20.6$ to -25.8 ; Table A3). It was, however, sensitive to school composition. When the school share of rare users and its interaction with the rare profile were added, the rare $\times$ practice interaction fell to -12.4 in science and -10.2 in reading and was no longer significant ($ps > .20$), while the rare-user gap was smaller in schools with a larger share of rare users ($b = 54.3$ and 56.2 , $ps < .03$). After Benjamini–Hochberg adjustment across the 24 profile-interaction coefficients, the rare $\times$ practice terms remained significant (adjusted $ps \leq .04$), but they did not survive Holm adjustment. We therefore treat this interaction as exploratory. Mathematics showed the same pattern for the rare-user gap and the interaction (Table A3).

5. Discussion

Students differed mainly in how often they used chatbots and much less in the purposes they pursued. Students who rarely used chatbots scored slightly lower than selective users, whereas intensive users did not, and neither metacognitive strategy use nor school practice in evaluating AI-generated content showed a robust moderating role.

The hypothesized split between exploratory and offloading use did not appear, so H1 was only partly supported. Studies of university students (Stojanov et al., 2024) and of Swedish secondary students (Hugerth & Hugerth, 2026) found profiles that differed in purpose as well as in level of reliance. Many 15-year-olds who adopt chatbots may apply them to several kinds of tasks, but the measure may also have limited what the analysis could detect, because four frequency items on a common scale favor solutions based on overall frequency, and one in three students gave identical answers. The ordering persisted without these students, but the small uniform monthly profile may largely reflect response style. In this population, how much students use chatbots therefore differentiates them more than what they use chatbots for.

Contrary to H2, intensive users did not score significantly lower than selective users, and the mean frequency of use was not significantly associated with achievement. Equivalence tests with bounds of ± 0.1 SD indicated a negligible intensive-selective difference in the main models (90% CIs: -5.6 to 3.0 points in science and -5.2 to 3.4 in reading) but not after curiosity and growth mindset were controlled, so a small disadvantage cannot be ruled out. This result departs from earlier PISA studies of ICT use (Hu et al., 2018; Tan et al., 2026) and appears to conflict with experiments in which unrestricted chatbot access impaired later unaided performance (Bastani et al., 2025) or did not improve knowledge retention (Fan et al., 2025). The designs differ, however. The experiments examined learning after brief or course-length

periods of chatbot access, whereas PISA linked students' reported everyday use to a test they completed on their own. Cross-sectional data cannot show whether intensive use changes learning.

The gap for rare users is the clearest association in the data, and it fits the null result for mean frequency if the relevant contrast lies near zero use: among students who used chatbots at all, more frequent use was not associated with achievement. Part of the gap lay between schools, because the unadjusted gap of about 21 points in science fell to 6.6 points in the multilevel model, but a gap also remained within schools, where regressions with school fixed effects gave 11 to 15 points (Table A2). Rare use was more common in lower secondary and private schools, among less advantaged students, and among students with less frequent metacognitive strategy use and less exposure to practice in evaluating AI content. Research on the digital divide offers one interpretation: once basic access spreads, inequalities shift to material access, digital skills, and the kinds of use that yield benefits, and these inequalities follow socioeconomic lines (Scheerder et al., 2017; van Deursen & van Dijk, 2019). The data are consistent with this account but cannot demonstrate it, because students were not asked about access to chatbots or reasons for not using them. Curiosity and growth mindset explained little of the gap (from about 7 to about 6 points).

The null moderation by metacognitive strategy use (H3) may reflect the measure as much as the construct. Metacognitive strategy use predicted profile membership, but the differences between profiles were similar for students with more and less frequent strategy use. The PISA index captures general strategy use, whereas experiments examined regulation scaffolds or processes during AI-assisted tasks (Darvishi et al., 2024; Fan et al., 2025). The unanticipated weaker association of help seeking with achievement among rare users did not survive the most conservative correction for multiple testing and needs replication. Testing the role that Molenaar (2022) and Järvelä et al. (2023) assign to self-regulation will require measures specific to AI use, such as how often students check, question, or revise chatbot output, or the quality of their chatbot interactions (Fütterer et al., 2026).

The exploratory analysis of school practice in evaluating AI content should be read with similar caution. Practice was associated with higher achievement among selective and intensive users but not among rare users, which made the rare-user gap larger where practice was more frequent. This pattern weakened substantially when school composition was controlled. Schools with more practice had fewer rare users ($r = -.66$ at the school level), and the rare-user gap was larger where rare users were few, so the interaction may reflect who the rare users are in such schools rather than how practice works. If the pattern replicates, evaluating AI output, a component of AI literacy (Ng et al., 2021), may help students mainly when they apply it to their own chatbot use.

5.1. Implications for practice and policy

The results give no clear indication that intensive chatbot use is linked to lower achievement among Indonesian 15-year-olds, but they also do not show that more use raises achievement: the gap concerned students who rarely used chatbots and may reflect selection rather than an effect of use. Three implications follow. First, because rare users were concentrated in lower secondary, private, and less advantaged settings, schools could examine whether these students lack access to chatbots or guidance in using them; the data cannot show whether rare use reflects constraint or choice; such monitoring could be part of national smart education policies, which are expanding worldwide (Yang et al., 2024). Second, instruction in evaluating AI output, for example tasks in which students compare a chatbot's answer with an assigned source and correct its errors, should reach students who seldom use chatbots and not only frequent users. Third, guided designs such as tutor-like chatbots (Bastani et al., 2025) and teacher-structured chatbot use in lower secondary science lessons (Le et al., 2026) offer models for such instruction, and experiments could test whether they benefit rare users.

5.2. Limitations and future research

First, the cross-sectional design does not permit causal inference, and the associations may reflect selection or reverse causation. Second, chatbot use was measured with four self-reported frequency items, and response styles may have shaped the profiles; the rare-user gap was not statistically significant when students with identical answers were excluded. Third, the main models assigned students to their most likely profile; BCH-weighted within-school regressions gave larger gaps and pseudo-class draws, which are biased toward zero (Bray et al., 2015), slightly smaller ones, so the multilevel estimates are likely conservative. Excluding the 447 students who answered none of the AI use items, who scored 35 to 40 points lower, may also understate the gap. Fourth, the moderators were measured with general indices and a single item, and the practice item was not missing completely at random; its school mean relies on student reports rather than observed teaching. Fifth, Indonesia did not administer the ICT familiarity and teacher questionnaires, so information on access, teacher guidance, school AI policies, and the quality of chatbot use was unavailable. Finally, the results come from one middle-income country and may not transfer to other systems. Longitudinal and experimental studies with secondary students, ideally with logs of chatbot interactions, could establish how use and school practice shape learning.

6. Conclusions

In nationally representative data from Indonesia, 15-year-olds who rarely used AI chatbots for schoolwork had slightly lower science and reading achievement than those who used them selectively (about 0.1 *SD* after accounting for student and school background), whereas intensive users did not differ from selective users. Profiles of use differed mainly in frequency, and neither metacognitive strategy use nor, in exploratory analyses, school practice in evaluating AI content moderated these associations robustly. These findings provide no evidence of a penalty for intensive use in this population, although a small one cannot be ruled out. They identify students who rarely use chatbots, and the question of whether guidance in evaluating AI output can reach them, as priorities for research and practice.

Abbreviations

AI

artificial intelligence

aBIC

sample-size-adjusted Bayesian information criterion

AvePP

average posterior probability

BCH

Bolck, Croon, and Hagnaars

BIC

Bayesian information criterion

BLRT

bootstrap likelihood ratio test

ESCS

index of economic, social, and cultural status

ICC

intraclass correlation coefficient

ICT

information and communication technology

LL

log-likelihood

OECD

Organisation for Economic Co-operation and Development

OR

odds ratio

PISA

Programme for International Student Assessment

RQ

research question

SD

standard deviation

WLE

weighted likelihood estimate.

Declarations

Availability of data and materials

The datasets analyzed during the current study are available from the Organisation for Economic Co-operation and Development (<https://www.oecd.org/en/data/datasets/pisa-2025-database.html>). The analysis data, outputs, and code that support the findings of this study are available in the Open Science Framework repository (https://osf.io/cu7p5/overview?view_only=42dcec64ed3c4f11ac9b7ecfae6b3a1a; anonymized view-only link for peer review).

Competing interests

The authors declare that they have no competing interests.

Funding

This research received no specific funding.

Authors' contributions

The authors' contributions are provided on the title page to preserve anonymity during double-anonymous review.

Acknowledgements

Acknowledgements are provided on the title page to preserve anonymity during double-anonymous review.

References

1. Abbas, M., Jam, F. A., & Khan, T. I. (2024). Is it harmful or helpful? Examining the causes and consequences of generative AI usage among university students. *International Journal of Educational Technology in Higher Education*, 21(1). Article 10. <https://doi.org/10.1186/s41239-024-00444-7>
2. Annuš, N. (2025). Investigation of generative AI adoption in IT-focused vocational secondary school programming education. *Education Sciences*, 15(9). <https://doi.org/10.3390/educsci15091152>. Article 1152.
3. Asparouhov, T., & Muthén, B. (2014). Auxiliary variables in mixture modeling: Three-step approaches using Mplus. *Structural Equation Modeling: A Multidisciplinary Journal*, 21(3), 329–341. <https://doi.org/10.1080/10705511.2014.915181>

4. Bailey, P., Webb, B., Kelley, C., Nguyen, T., & Huo, H. (2024). WeMix: Weighted mixed-effects models using multilevel pseudo maximum likelihood estimation (Version 4.0.4) [Computer software]. <https://CRAN.R-project.org/package=WeMix>
5. Bakk, Z., Tekle, F. B., & Vermunt, J. K. (2013). Estimating the association between latent class membership and external variables using bias-adjusted three-step approaches. *Sociological Methodology*, 43(1), 272–311. <https://doi.org/10.1177/0081175012470644>
6. Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakçı, Ö., & Mariman, R. (2025). Generative AI without guardrails can harm learning: Evidence from high school mathematics. *Proceedings of the National Academy of Sciences*, 122(26), Article e2422633122. <https://doi.org/10.1073/pnas.2422633122>
7. Bell, A., Fairbrother, M., & Jones, K. (2019). Fixed and random effects models: Making an informed choice. *Quality & Quantity*, 53(2), 1051–1074. <https://doi.org/10.1007/s11135-018-0802-x>
8. Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1), 289–300. <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>
9. Bolck, A., Croon, M., & Hagenars, J. (2004). Estimating latent structure models with categorical variables: One-step versus three-step estimators. *Political Analysis*, 12(1), 3–27. <https://doi.org/10.1093/pan/mp001>
10. Bray, B. C., Lanza, S. T., & Tan, X. (2015). Eliminating bias in classify-analyze approaches for latent class analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, 22(1), 1–11. <https://doi.org/10.1080/10705511.2014.935265>
11. Brezovec, E., Zelić, M., & Zagode, A. M. (2026). Stabilizing truth in educational sciences: A systematic review of generative AI in education. *Kybernetes*, 55(13), 1–17. <https://doi.org/10.1108/K-09-2025-2339>
12. Broadbent, J., & Poon, W. L. (2015). Self-regulated learning strategies & academic achievement in online higher education learning environments: A systematic review. *The Internet and Higher Education*, 27, 1–13. <https://doi.org/10.1016/j.iheduc.2015.04.007>
13. Callado, A., Saúde, I., & Araújo, J. L. (2025). Portuguese secondary school students' perceptions regarding the use of ChatGPT. *Contemporary Educational Technology*, 17(4). <https://doi.org/10.30935/cedtech/17516>. Article ep610.
14. Casal-Otero, L., Catala, A., Fernández-Morante, C., Taboada, M., Cebreiro, B., & Barro, S. (2023). AI literacy in K-12: A systematic literature review. *International Journal of STEM Education*, 10(1). <https://doi.org/10.1186/s40594-023-00418-7>. Article 29.
15. Chan, C. K. Y. (2026). Cheating or not cheating? Rethinking AI-giarism and academic integrity through secondary students' ethical reasoning. *Computers & Education*. <https://doi.org/10.1016/j.compedu.2026.105698>. 253, Article 105698.
16. 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(1). <https://doi.org/10.1186/s41239-023-00411-8>. Article 43.
17. Darvishi, A., Khosravi, H., Sadiq, S., Gašević, D., & Siemens, G. (2024). Impact of AI assistance on student agency. *Computers & Education*. <https://doi.org/10.1016/j.compedu.2023.104967>, 210, Article 104967.
18. Deng, R., Jiang, M., Yu, X., Lu, Y., & Liu, S. (2025). Does ChatGPT enhance student learning? A systematic review and meta-analysis of experimental studies. *Computers & Education*, 227, 105224. <https://doi.org/10.1016/j.compedu.2024.105224>
19. Dent, A. L., & Koenka, A. C. (2016). The relation between self-regulated learning and academic achievement across childhood and adolescence: A meta-analysis. *Educational Psychology Review*, 28(3), 425–474. <https://doi.org/10.1007/s10648-015-9320-8>

20. Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods*, 12(2), 121–138. <https://doi.org/10.1037/1082-989X.12.2.121>
21. Fan, Y., Tang, L., Le, H., Shen, K., Tan, S., Zhao, Y., Shen, Y., Li, X., & Gašević, D. (2025). Beware of metacognitive laziness: Effects of generative artificial intelligence on learning motivation, processes, and performance. *British Journal of Educational Technology*, 56(2), 489–530. <https://doi.org/10.1111/bjet.13544>
22. Fütterer, T., Bardach, L., Kuhn, J., Keller, S. D., & Gerjets, P. (2026). Enhancing school students' self-regulated learning through generative AI support: A randomized controlled trial. *Educational Psychology Review*, 38(1). <https://doi.org/10.1007/s10648-026-10133-8>. Article 42.
23. Gerlich, M. (2025). AI tools in society: Impacts on cognitive offloading and the future of critical thinking. *Societies*, 15(1). <https://doi.org/10.3390/soc15010006>. Article 6.
24. Grinschgl, S., Papenmeier, F., & Meyerhoff, H. S. (2021). Consequences of cognitive offloading: Boosting performance but diminishing memory. *Quarterly Journal of Experimental Psychology*, 74(9), 1477–1496. <https://doi.org/10.1177/17470218211008060>
25. Grund, S., Lüdtke, O., & Robitzsch, A. (2018). Multiple imputation of missing data for multilevel models: Simulations and recommendations. *Organizational Research Methods*, 21(1), 111–149. <https://doi.org/10.1177/1094428117703686>
26. Heisig, J. P., & Schaeffer, M. (2019). Why you should always include a random slope for the lower-level variable involved in a cross-level interaction. *European Sociological Review*, 35(2), 258–279. <https://doi.org/10.1093/esr/jcy053>
27. Hu, X., Gong, Y., Lai, C., & Leung, F. K. S. (2018). The relationship between ICT and student literacy in mathematics, reading, and science across 44 countries: A multilevel analysis. *Computers & Education*, 125, 1–13. <https://doi.org/10.1016/j.compedu.2018.05.021>
28. Hugerth, M. W., & Hugerth, L. W. (2026). Seeking knowledge or efficiency: Profiling students' AI-use through survey-based latent class analysis. *Computers and Education: Artificial Intelligence*. <https://doi.org/10.1016/j.caeai.2025.100531>, 10, Article 100531.
29. Järvelä, S., Nguyen, A., & Hadwin, A. (2023). Human and artificial intelligence collaboration for socially shared regulation in learning. *British Journal of Educational Technology*, 54(5), 1057–1076. <https://doi.org/10.1111/bjet.13325>
30. Jerrim, J., Lopez-Agudo, L. A., Marcenaro-Gutierrez, O. D., & Shure, N. (2017). What happens when econometrics and psychometrics collide? An example using the PISA data. *Economics of Education Review*, 61, 51–58. <https://doi.org/10.1016/j.econedurev.2017.09.007>
31. Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103. <https://doi.org/10.1016/j.lindif.2023.102274>. Article 102274.
32. Klar, M. (2025). Using ChatGPT is easy, using it effectively is tough? A mixed methods study on K-12 students' perceptions, interaction patterns, and support for learning with generative AI chatbots. *Smart Learning Environments*, 12(1). <https://doi.org/10.1186/s40561-025-00385-2>. Article 32.
33. Le, N. C., Nguyen, H. T., Nguyen, Q. T. T., Nguyen, L. P. H., Tran, T. V., Tran, T. A., Pham, Q. V., & Nguyen, G. T. (2026). ChatGPT as cognitive scaffolding in 5E-based science instruction: Effects on middle school students' problem-solving competence. *EURASIA Journal of Mathematics Science and Technology Education*, 22(8). <https://doi.org/10.29333/ejmste/19030>. Article em2886.
34. Li, K. H., Raghunathan, T. E., & Rubin, D. B. (1991). Large-sample significance levels from multiply imputed data using moment-based statistics and an F reference distribution. *Journal of the American Statistical Association*, 86(416),

1065–1073. <https://doi.org/10.1080/01621459.1991.10475152>

35. Li, S. C., & Petersen, K. B. (2022). Does ICT matter? Unfolding the complex multilevel structural relationship between technology use and academic achievements in PISA 2015. *Educational Technology & Society*, 25(4), 43–55. [https://doi.org/10.30191/ETS.202210_25\(4\).0004](https://doi.org/10.30191/ETS.202210_25(4).0004)
36. Liu, M., & Zhang, L. J. (2025). Examining language learners' GenAI-assisted writing self-efficacy profiles and the relationship with their writing self-regulated learning strategies. *System*, 134., Article 103826. <https://doi.org/10.1016/j.system.2025.103826>
37. Lüdtke, O., Marsh, H. W., Robitzsch, A., Trautwein, U., Asparouhov, T., & Muthén, B. (2008). The multilevel latent covariate model: A new, more reliable approach to group-level effects in contextual studies. *Psychological Methods*, 13(3), 203–229. <https://doi.org/10.1037/a0012869>
38. Mang, J., Küchenhoff, H., Meinck, S., & Prenzel, M. (2021). Sampling weights in multilevel modelling: An investigation using PISA sampling structures. *Large-scale Assessments in Education*, 9(1). <https://doi.org/10.1186/s40536-021-00099-0>. Article 6.
39. Molenaar, I. (2022). Towards hybrid human-AI learning technologies. *European Journal of Education*, 57(4), 632–645. <https://doi.org/10.1111/ejed.12527>
40. Morin, S., Legault, R., Laliberté, F., Bakk, Z., Giguère, C. É., de la Sablonnière, R., & Lacourse, É. (2025). StepMix: A Python package for pseudo-likelihood estimation of generalized mixture models with external variables. *Journal of Statistical Software*, 113(8), 1–39. <https://doi.org/10.18637/jss.v113.i08>
41. 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*. <https://doi.org/10.1016/j.caeai.2021.100041>, 2, Article 100041.
42. Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling: A Multidisciplinary Journal*, 14(4), 535–569. <https://doi.org/10.1080/10705510701575396>
43. Nylund-Gibson, K., & Choi, A. Y. (2018). Ten frequently asked questions about latent class analysis. *Translational Issues in Psychological Science*, 4(4), 440–461. <https://doi.org/10.1037/tps0000176>
44. OECD (2026a). PISA 2025 results (Volume I): Indonesia [Country note]. OECD Publishing. https://www.oecd.org/en/publications/pisa-2025-results-volume-i-country-notes_2d4ff9ea-en/indonesia_9c880f8a-en.html
45. OECD (2026b). PISA 2025 results (Volume I): Future-ready students. OECD Publishing. <https://doi.org/10.1787/73451bc5-en>
46. Panadero, E. (2017). A review of self-regulated learning: Six models and four directions for research. *Frontiers in Psychology*, 8., Article 422. <https://doi.org/10.3389/fpsyg.2017.00422>
47. Park, S., & Nan, X. (2026). Generative AI and misinformation: A scoping review of the role of generative AI in the generation, detection, mitigation, and impact of misinformation. *AI & Society*, 41(2), 1501–1515. <https://doi.org/10.1007/s00146-025-02620-3>
48. Petko, D., Cantieni, A., & Prasse, D. (2017). Perceived quality of educational technology matters: A secondary analysis of students' ICT use, ICT-related attitudes, and PISA 2012 test scores. *Journal of Educational Computing Research*, 54(8), 1070–1091. <https://doi.org/10.1177/0735633116649373>
49. Pritchard, D. (2026). Generative AI is untrustworthy. *Synthese*, 208(2). <https://doi.org/10.1007/s11229-026-05755-y>. Article 97.
50. R Core Team (2024). R: A language and environment for statistical computing (Version 4.3.3) [Computer software]. R Foundation for Statistical Computing. <https://www.R-project.org/>

51. Rabe-Hesketh, S., & Skrondal, A. (2006). Multilevel modelling of complex survey data. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 169(4), 805–827. <https://doi.org/10.1111/j.1467-985X.2006.00426.x>
52. Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. *Trends in Cognitive Sciences*, 20(9), 676–688. <https://doi.org/10.1016/j.tics.2016.07.002>
53. Rutkowski, L., Gonzalez, E., Joncas, M., & von Davier, M. (2010). International large-scale assessment data: Issues in secondary analysis and reporting. *Educational Researcher*, 39(2), 142–151. <https://doi.org/10.3102/0013189X10363170>
54. Scheerder, A., van Deursen, A., & van Dijk, J. (2017). Determinants of Internet skills, uses and outcomes. A systematic review of the second- and third-level digital divide. *Telematics and Informatics*, 34(8), 1607–1624. <https://doi.org/10.1016/j.tele.2017.07.007>
55. Sparrow, B., Liu, J., & Wegner, D. M. (2011). Google effects on memory: Cognitive consequences of having information at our fingertips. *Science*, 333(6043), 776–778. <https://doi.org/10.1126/science.1207745>
56. Stadler, M., Bannert, M., & Sailer, M. (2024). Cognitive ease at a cost: LLMs reduce mental effort but compromise depth in student scientific inquiry. *Computers in Human Behavior*, 160., Article 108386. <https://doi.org/10.1016/j.chb.2024.108386>
57. Stojanov, A., Liu, Q., & Koh, J. H. L. (2024). University students' self-reported reliance on ChatGPT for learning: A latent profile analysis. *Computers and Education: Artificial Intelligence*, 6., Article 100243. <https://doi.org/10.1016/j.caeai.2024.100243>
58. Strzelecki, A. (2024). To use or not to use ChatGPT in higher education? A study of students' acceptance and use of technology. *Interactive Learning Environments*, 32(9), 5142–5155. <https://doi.org/10.1080/10494820.2023.2209881>
59. Tan, Y., Mendoza, N. B., & Yeung, S. S. S. (2026). Profiling ICT users and social-emotional and academic outcomes: A multilevel latent profile analysis. *European Journal of Psychology of Education*, 41(2). <https://doi.org/10.1007/s10212-026-01098-x>. Article 48.
60. van Deursen, A. J. A. M., & van Dijk, J. A. G. M. (2019). The first-level digital divide shifts from inequalities in physical access to inequalities in material access. *New Media & Society*, 21(2), 354–375. <https://doi.org/10.1177/1461444818797082>
61. Vermunt, J. K. (2010). Latent class modeling with covariates: Two improved three-step approaches. *Political Analysis*, 18(4), 450–469. <https://doi.org/10.1093/pan/mpq025>
62. Weinhandl, R., Andic, B., Wijaya, T. T., Bleckenwegner, V., Riegler, V., Baldinger, S., Mayrhofer, J., & Helm, C. (2026). Beyond calculators: Decoding students' usage behaviour regarding AI chatbots for mathematics using UTAUT. *Journal of Open Innovation: Technology Market and Complexity*, 12(2). <https://doi.org/10.1016/j.joitmc.2026.100773>. Article 100773.
63. Weller, B. E., Bowen, N. K., & Faubert, S. J. (2020). Latent class analysis: A guide to best practice. *Journal of Black Psychology*, 46(4), 287–311. <https://doi.org/10.1177/0095798420930932>
64. Wu, R., & Yu, Z. (2024). Do AI chatbots improve students learning outcomes? Evidence from a meta-analysis. *British Journal of Educational Technology*, 55(1), 10–33. <https://doi.org/10.1111/bjet.13334>
65. Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., Li, X., Jin, Y., & Gašević, D. (2024). Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology*, 55(1), 90–112. <https://doi.org/10.1111/bjet.13370>
66. Yang, J., Sun, Y., Lin, R., & Zhu, H. (2024). Strategic framework and global trends of national smart education policies. *Humanities and Social Sciences Communications*, 11(1). <https://doi.org/10.1057/s41599-024-03668-0>
67. Zhai, C., Wibowo, S., & Li, L. D. (2024). The effects of over-reliance on AI dialogue systems on students' cognitive abilities: A systematic review. *Smart Learning Environments*, 11(1). <https://doi.org/10.1186/s40561-024-00316-7>. Article 28.

68. Zhong, B., & Liu, X. (2025). Evaluating AI literacy of secondary students: Framework and scale development. *Computers & Education*. <https://doi.org/10.1016/j.compedu.2024.105230>. 227, Article 105230.

69. Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. *Theory Into Practice*, 41(2), 64–70. https://doi.org/10.1207/s15430421tip4102_2

Figures

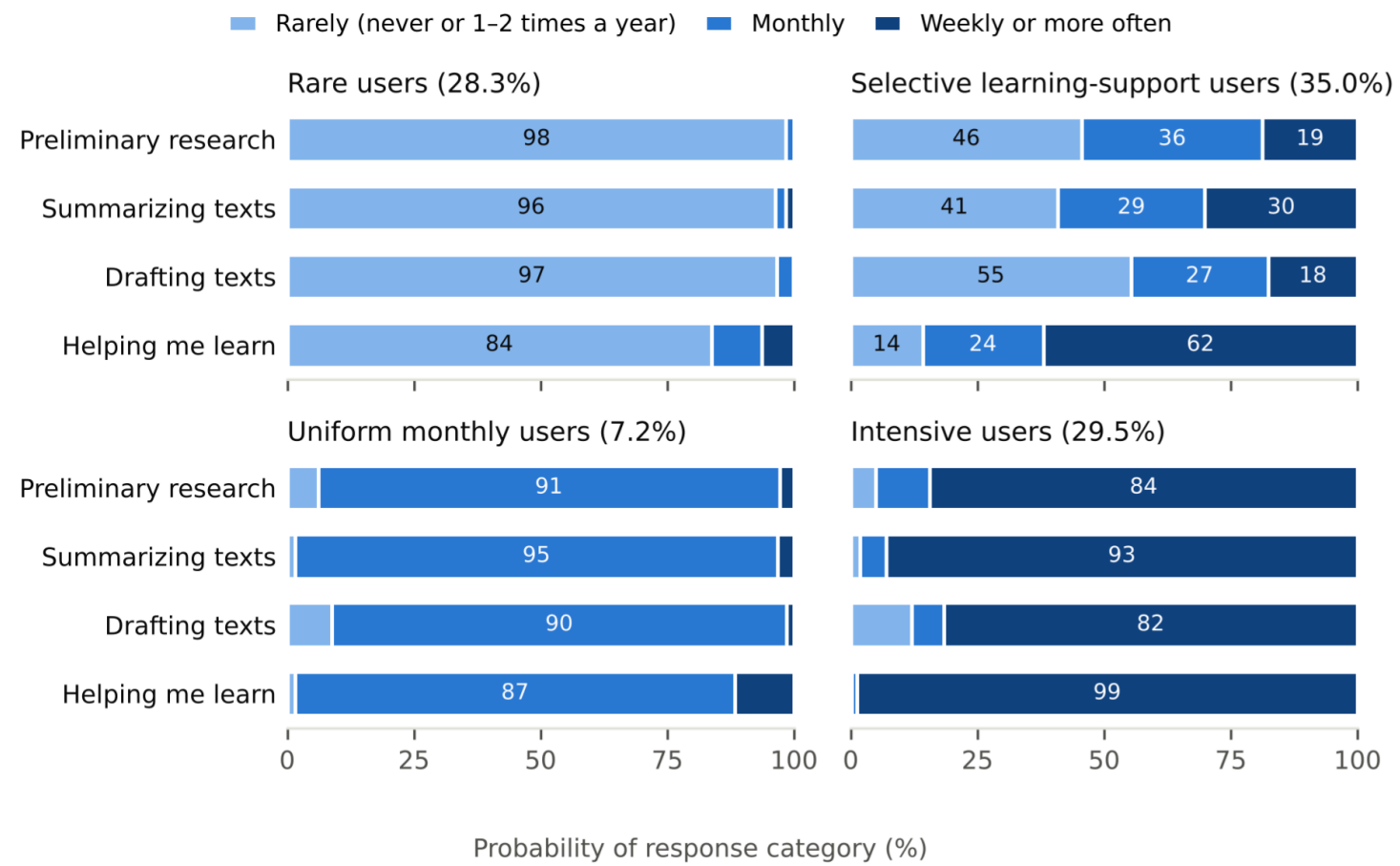


Figure 1

Item response probabilities of the four AI chatbot use profiles. Percentages in parentheses are profile sizes ($n = 13721$); segments below 12% are not labeled.

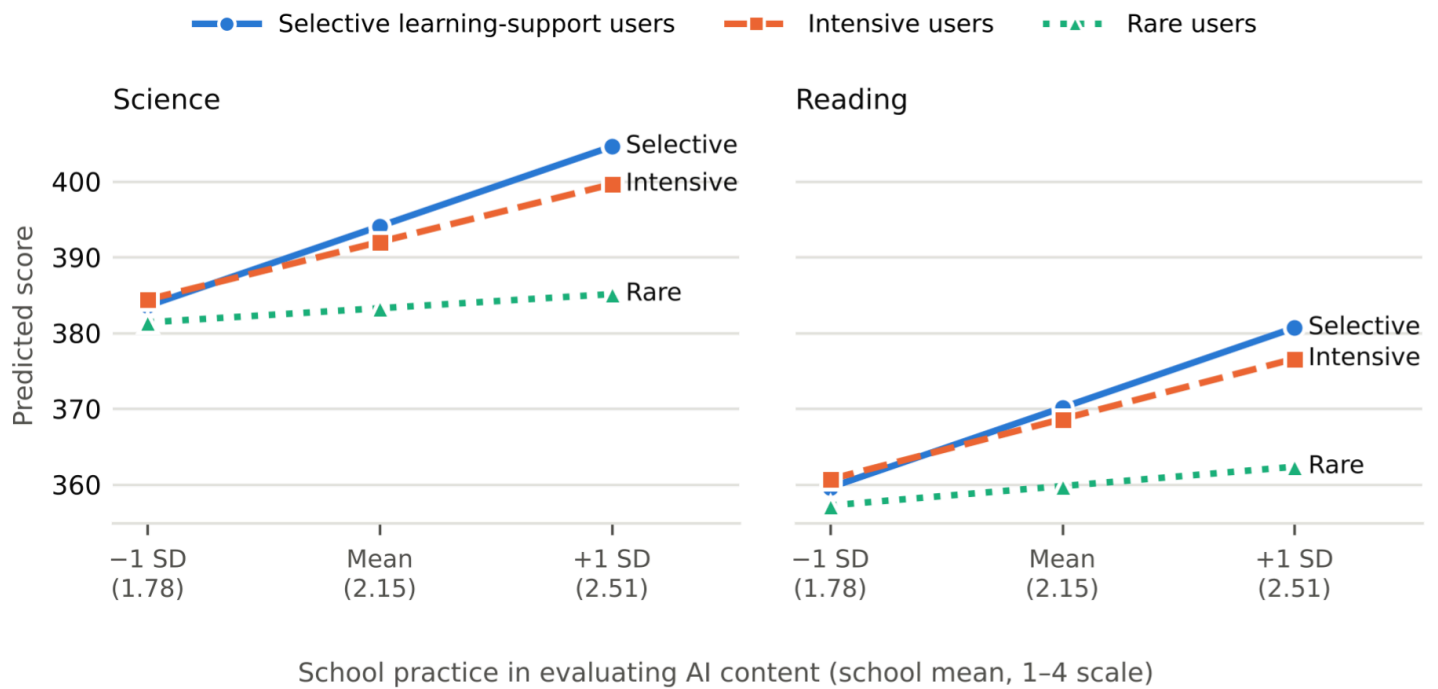


Figure 2

Predicted achievement by AI use profile and school practice in evaluating AI content. Exploratory model M4; predictions hold the covariates at their weighted means; uniform monthly users are omitted for clarity.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [LSAEAdditionalfile1.docx](#)
- [Appendix.docx](#)