Print ISSN: 3104-932X
Online ISSN: 3104-9338

Constructing a Generative AI Framework for Personalised Instructional Resource Development: A Quasi-Experimental Evaluation in a Micro-Lecture Design and Production Course

Guihua Gao*, Hanyan Cai, Shiru Zhu

Yiyang Normal College, Yiyang 413000, Hunan, China

Received: October 8, 2026

Revised:October 17, 2026

Accepted: October 21, 2026

Published online: October 22, 2026

To appear in: *International Journal of Advanced AI Applications,* Vol. 2, No. 11(November 2026)

Corresponding Author: Guihua Gao (2260540996@qq.com)

Abstract. Generative artificial intelligence is promoted as a way of tailoring instruction to learners, yet most reported applications stay within language and general-education settings, and few measure outcomes in courses whose value lies in design and production skill. This paper describes a four-module generative AI framework for resource development that links learner and content modelling, generation, assessment and community iteration to a course-grounded knowledge base, with the teacher reviewing before release. It was evaluated in a twelve-week quasi-experiment in a professional core course, Micro-lecture Design and Production, at a Chinese vocational college. Two intact classes (n = 45, 44) were taught by the same instructor, who designed the framework; assignment was not random. Covariance-adjusted posttest means differed by 5.6 points (82.3 versus 76.7, $p < .001$, partial $\eta^2 = .24$), and the framework-supported class reported higher motivation and satisfaction; self-directed learning differed modestly. Preparation time per unit fell by half, though the work became more concentrated and the knowledge base carried a large one-off cost. Because the design cannot separate the framework from the additional resources it produced, the results are framed as associations rather than effects.

Keywords: *generative artificial intelligence; personalised learning; instructional resource design; micro-lecture production; teacher education*

1. Introduction

Large language models now draft explanations, summarise readings, compose practice questions and rewrite passages for different reading levels, a capability documented from the transformer architecture [1] to the scale at which instruction-following behaviour emerges [2].

Differentiated instructional material is among the most labour-intensive parts of teaching and is now one of the tasks generative AI supports most directly [3,4].

Institutional responses have been cautious but constructive. UNESCO treats generative AI as a tool whose value depends on the governance and pedagogy built around it rather than on the model [5], and reviews reach a similar conclusion about the potential and the risks [6]. What determines whether such a tool helps is not what it generates but whether its output fits the curriculum, the learners and the assessment already in place [7].

The empirical literature has not kept pace. Reviews keep finding the same profile — small perception surveys, a concentration on writing and language learning, general-education settings [8,9] — with three recurring weaknesses named in recent scoping work [10]. Many studies report what students think about generative AI rather than what they learn with it; where outcomes are measured, the intervention is usually the model's general capability rather than a designed workflow another teacher could reproduce; and courses assessed on what students make rather than what they recall are almost absent.

The gap is most consequential in teacher education, where using these tools well sits at the junction of technology, pedagogy and subject matter [11], and studies of pre-service teachers report positive attitudes but seldom test what those teachers can actually design [12]. Micro-lectures make the question concrete: short videos are routine in blended, vocational and distance teaching [13], and production quality rests on design judgements — how far to segment a topic, how narration should sit against imagery — that experienced teachers apply tacitly and that govern how long students watch and how much they retain [14].

We designed and evaluated a generative-AI-supported framework for personalised instructional resource development (PIRD) in the professional core course Micro-lecture Design and Production, taken by second-year Modern Educational Technology students at a Chinese higher vocational college. The evaluation is a twelve-week quasi-experiment in two intact classes combining an examination with validated scales, platform logs and interviews; it asked whether achievement differs in a course whose outcomes are design and production skill, how the affective outcomes compare, and what changes for the teacher. Two boundaries apply from the outset: classes were not randomly assigned and the instructor was also a researcher, so the design supports statements about association rather than effect, and the framework was deployed as a package, so no difference can be attributed to a single component (Section 5.2).

2. Related Work

2.1. Generative AI in Education: From Capability to Application

Research on generative AI in education has grown quickly, with an uneven centre of gravity. A systematic review catalogued largely pedagogical opportunities and challenges — adaptive pacing, timely feedback, reduced administrative burden — along with the persistent difficulty of aligning automated output with learning outcomes [4], while a companion review of higher education found empirical work clustered around writing support, language learning and student perceptions, with little attention to how instructors build the technology into a course [8]. Recent scoping work adds a warning: without deliberate design, LLM output favours fluent but shallow text over the disciplinary reasoning courses exist to develop [10].

Studies that do look at instruction tend to examine how students use a general-purpose chatbot, not how a purpose-built system performs. Surveys across four universities found benefits framed as efficiency and quality of explanation, and concerns framed as accuracy and over-dependence [15]; a study in science education found strong performance on conceptual questions and unreliable performance on procedural ones [16]. What remains unsettled is not whether a language model can produce instructional text, but how to organise that capability so the text is accurate, well pitched and tied to assessment.

2.2. Personalised and Adaptive Learning: The Older Evidence Base

Personalised instruction is an old ambition. Bloom framed the problem four decades ago: one-to-one tutoring produces outcomes about two standard deviations above conventional group instruction [17]. Reviews of adaptive learning document steady growth alongside persistent methodological problems [18], and a multi-year implementation study found modest average effects that varied across sites, with implementation strength predicting outcomes better than the technology chosen [19].

A synthesis of this literature distinguishes what gets personalised: pace, content, sequence or modality [20]. Most adaptive systems personalise pace and sequence through item-response estimation, which helps little when the difficulty is representational — when a student cannot connect a step she can already perform to the reason it works. Generative AI is attractive exactly here, because it can produce alternative explanations and representations of the same concept on demand. Two theories constrain how those alternatives should be graded: cognitive load theory holds that worked examples lose their advantage as expertise grows, so the useful form of a resource depends on the learner's current state rather than a fixed template [21], while the

zone of proximal development adds that support should sit just beyond what the learner can already do alone [22]. The framework operationalises both crudely, as a difficulty band with a target success probability of .60 to .85 on the practice items attached to a resource (Section 4.5); a three-level band is not a continuous estimate of what a student can do with help.

2.3. Resource Design and Assessment in Teacher-Education Courses

Instructional design for complex professional skills has its own tradition. The four-component model organises learning tasks by whole-task progression [23], and Merrill's first principles make the same point from the learner's side: new knowledge is best demonstrated in the context of a task the student will perform [24].

Courses that produce instructional media add a second body of design knowledge. Multimedia learning research has established that the way verbal and pictorial information is combined determines how much of it learners can process, and that a small set of principles — segmenting a presentation into learner-paced parts, presenting narration and imagery together instead of duplicating one in the other, and signalling what matters — has consistent and substantial effects [25]. These principles specify what a script, a storyboard and a screen layout should look like, so a resource can be judged against criteria that do not depend on taste; Section 4.5 audits all 36 released resources against five of them. Assessment research points the same way: formative assessment produces gains when it gives students information they can act on [26], the effect depends on the quality of feedback rather than its quantity [27], and students benefit most from feedback that helps them judge their own work against a standard [28]. For generative AI this is also a hazard, since the model that supplies an immediate explanation can also supply a fluent but wrong one: hallucination is a documented behaviour of language generation rather than an occasional defect [29], which is why grounding output in a retrieved corpus has become standard wherever factual reliability matters [30].

Teacher knowledge is the third element, and it is where this study sits. TPACK separates what a tool can supply from what only a teacher can decide [11]: the framework supplies content knowledge (a concept graph and an item bank), technological content knowledge (tiered explanations and annotated storyboards) and technological pedagogical knowledge (automated assessment and the infrastructure for interaction), but no module supplies pedagogical content knowledge — the judgement about what counts as a good micro-lecture on this topic, for these students. That judgement sits at the review stage, which is why the stage is obligatory rather than optional, and why the study measured student outcomes and teacher time rather than the teacher's TPACK.

3. Methodology

3.1. Research Design

The study used a quasi-experimental pretest-posttest design with two intact classes; random assignment was not possible within the institution's timetable. The class that received framework-generated resources is referred to as the experimental class and the other as the control class. A pretest was administered in week one, the intervention ran for twelve teaching weeks, and a posttest followed in week thirteen, with platform logs, interviews and teacher-side time logs as supplementary data.

Three features of the design bound what can be concluded. Both classes were taught by the same instructor, who had designed the framework and knew the group assignment, so instructional behaviour cannot be assumed identical across classes. With 45 and 44 students the study had 80% power to detect a difference of about d = 0.60, so smaller effects, the common case in education research, would have gone undetected. The intervention was deployed as a complete package, so the analysis estimates the association between the package and the outcomes rather than the contribution of its components.

3.2. Setting, Participants and Ethics

The study was conducted at a public higher vocational college in Hunan Province, China, in the professional core course Micro-lecture Design and Production, which runs for sixteen weeks at four class hours per week in a digital media production workshop and covers the full production cycle, from script writing to narration recording, editing and publishing. Two second-year classes took part, giving 89 students after three were excluded for missing more than two of the twelve weekly quizzes; one instructor taught both, with 45 and 44 students respectively, using the same syllabus, textbook, workshop sequence and examination paper.

Students were told that declining or withdrawing would not affect their grades, that the two classes would be taught with different supporting materials, and that the comparison would be disclosed to both classes at the end of the semester; written consent was obtained from all 89. Platform data were anonymised and students were referenced by pseudonymous identifiers.

Two measures were taken against the most obvious sources of bias. The examination was marked against a scheme finalised before the intervention began, and its four design questions were scored independently by two raters from a different department who had no teaching contact with either class, were not told which class was which, and rated de-identified submissions (agreement, ICC(2,2) = .81). Neither addresses instructional bias: the author

designed the framework and taught both classes, and its screening rules were fixed before the semester rather than adjusted during it (Section 5.4).

Because the materials supplemented rather than replaced instruction, the control class received the full curriculum; delayed provision was judged the least restrictive way to permit a comparison, and the package was released to control-class students after the posttest (Section 4.7).

3.3. The PIRD Framework

The design rests on one principle: generative AI should be embedded in a defined instructional function, with a human decision at the point where material reaches students. It has four modules supported by a shared service layer and is bounded above by a teacher review stage that no resource bypasses (Figure 1); Table 1 summarises what each module does and which teacher-side task it replaces.

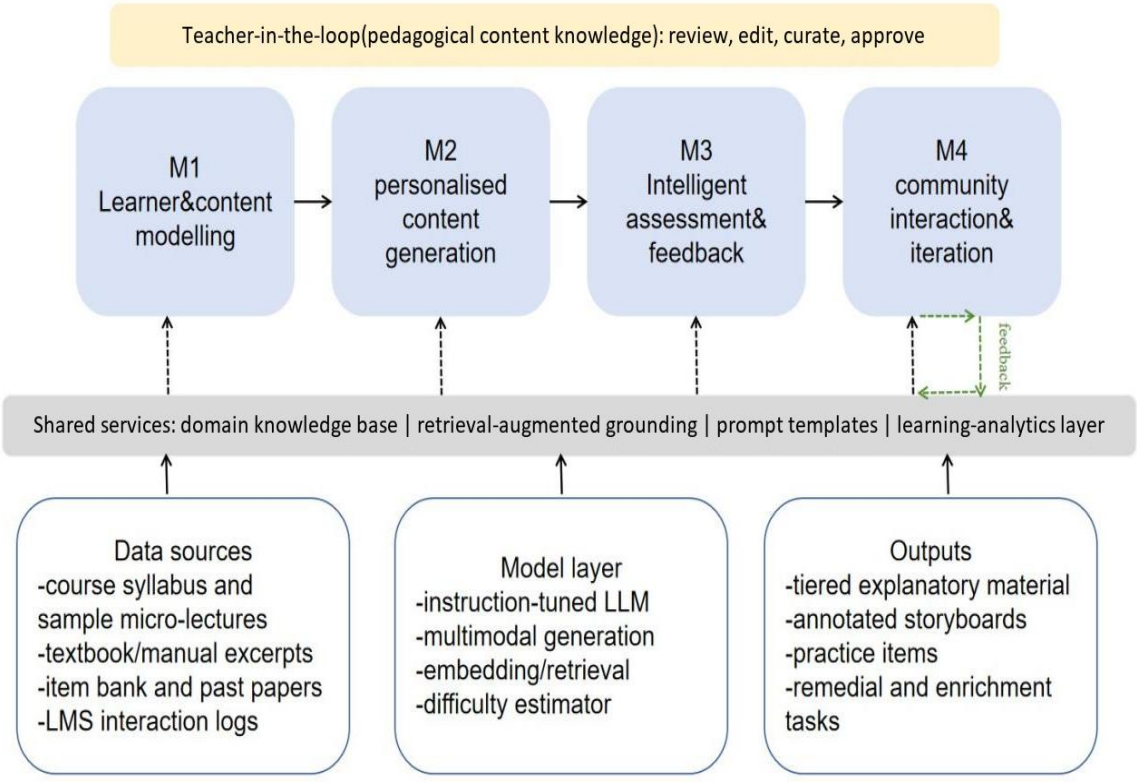


Figure 1. Architecture of the proposed PIRD framework, showing the four modules, the shared service layer, the feedback path from assessment back to learner modelling, and the teacher review stage at which pedagogical content knowledge is exercised.

Two parts of the knowledge base were built rather than collected, and their provenance bears on transfer. The concept graph was constructed by the instructor together with experienced

teachers of the same course and revised over two rounds of independent review, in which a node was retained only where the reviewers judged it a prerequisite for a topic the course assesses and relations were kept only where they agreed on direction; the resulting graph was then checked against the previous cohort's error patterns on the weekly quizzes. The item bank was calibrated before release rather than written to a target in the abstract. Items inherited from earlier cohorts carry difficulty estimates derived from their p-values in those cohorts' quizzes, and items that were newly generated were reviewed by hand before entering the bank, so that every item is stored with the topic and the difficulty band it was assigned. This is what allows a generation request to specify a target difficulty band instead of leaving that judgement to the model.

Table 1. Functional design of the four framework modules.

Module	Primary function	Key technologies	Teacher-side task replaced
M1. Learner and content modelling	Per-student mastery and representation profiles; course concept graph	Diagnostic scoring; log analysis; prerequisite mapping	Manual review of quiz and homework records
M2. Personalised content generation	Tiered explanations, worked examples, guided practice items	Instruction-tuned LLM; retrieval-augmented generation; storyboard rendering	Drafting of differentiated handouts and exercise sets
M3. Intelligent assessment	Calibrated items, automated scoring, error profiles	Difficulty estimation; structured marking schemes; concept tagging	Item writing; marking of objective items; error analysis
M4. Community interaction and iteration	Annotated discussion; approval-gated replies; resource regeneration	Forum integration; request logging; weekly analytics	Handling repetitive questions; revising materials between cohorts

The system runs on an instruction-tuned LLM accessed through a commercial API, with an off-the-shelf embedding model for retrieval [2]. No model was fine-tuned: discipline-specific behaviour came from retrieval grounding and prompt design, which keeps the system portable across courses and puts the burden of accuracy on the knowledge base. Module 3 reuses the same embedding model for tagging instead of training a separate classifier.

3.4. Knowledge Grounding and Prompt Design

Generation follows a fixed seven-step workflow (Figure 2): define the objective, the difficulty band and the output format; retrieve grounded passages; assemble a structured prompt from a versioned template; produce three candidate resources; screen them automatically; obtain the teacher's review and release; and collect performance data, which updates the learner

profile and closes the loop.

The screening in step 5 is where an unchecked model could release a plausible error. Each candidate is checked on alignment with the objective, on unsupported factual claims relative to the retrieved passages, on readability, and on consistency between the declared difficulty band and the estimated item difficulty. From week six, after the audits reported in Section 4.5 showed that design-level properties were being missed, signalling markers and temporal contiguity between narration and imagery were added as further checks. A candidate that fails is regenerated once, and a second failure flags it for manual attention.

Prompts were reusable templates with four slots — role, task, constraint and exemplar. One constraint-slot template asks the model to explain to second-year pre-service teachers how narration and on-screen imagery should be synchronised in a micro-lecture, in at most 400 characters and drawing only on the retrieved passages, while the exemplar slot holds one instructor-approved resource that fixes tone and structure. Structured prompts of this kind, and explicit step-by-step reasoning cues, change output quality substantially [31]. The templates used here are the fourth revision, iterated through a two-week pilot.

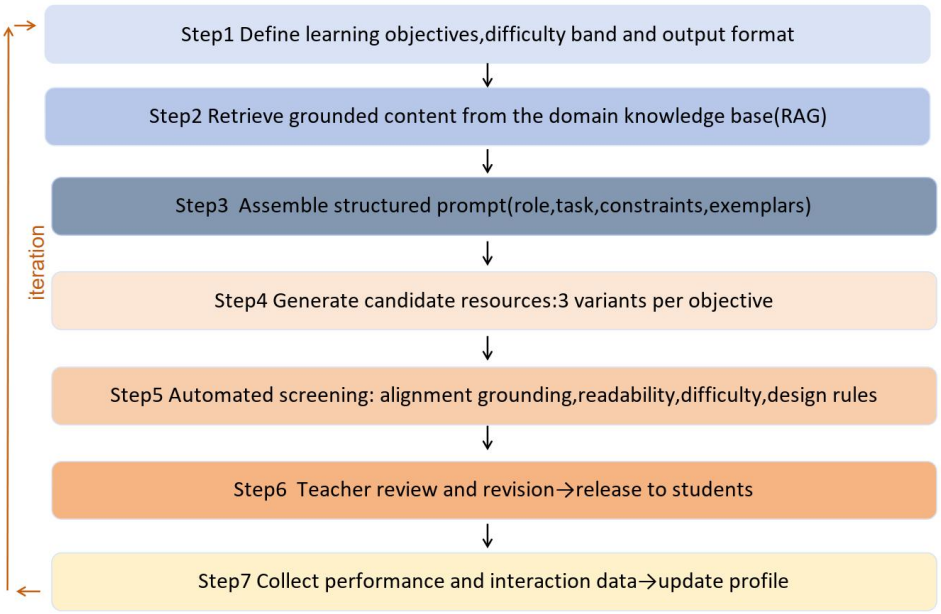


Figure 2. Seven-step resource generation workflow with automated screening and teacher review.

Three design decisions shaped the results. Every request carries an explicit difficulty band instead of leaving that judgement to the model. The grounding requirement is enforced in the prompt and then checked automatically, which reduces unsupported claims without eliminating them. And the framework offers several variants at once, on the reasoning that choosing between options is easier than specifying the answer in advance — a decision that had a

consequence for students we had not anticipated (Section 4.7).

Retrieval is fixed rather than tuned. The corpus described in Section 3.3 is chunked and indexed offline, and a request is answered from a bounded set of the highest-ranked passages for that query rather than from the whole corpus; the grounding check in step 5 is then carried out against exactly those passages, so the retrieved set places a hard ceiling on what can be released. The same retriever serves generation in Module 2 and concept tagging in Module 3, so a released resource and the error profile derived from it rest on the same representation of the course content. Two properties of this layer matter for replication. It is a replaceable component rather than part of the intervention: the embedding model is a general-purpose, off-the-shelf one and, like the generator, was not fine-tuned, so the framework's behaviour was pinned to the knowledge base and the prompt templates rather than to a particular vector space, and a re-implementation can substitute another embedding model without altering the framework's logic. And no retrieval setting was tuned on outcome data: the values in force during the semester were those configured before it began and were left unchanged while it ran.

3.5. Instruments

Four instruments were used. Achievement was measured by a 100-point examination with 40 objective items worth 40 marks and four design questions worth 60 marks, written by the teaching team and piloted on a previous cohort. Learning motivation used an adapted 14-item MSLQ subscale covering intrinsic goal orientation and task value [32]; satisfaction used a 10-item scale adapted from earlier work on technology-supported courses; and self-directed learning used a 12-item scale covering goal setting, strategy use and self-monitoring [28].

All self-report scales used a five-point Likert format, and reliability was estimated separately at pretest and posttest (Table 2); the motivation measure was the weakest at pretest, and results involving it are read with that in mind. Semi-structured interviews were held with 12 students from the experimental class, chosen to span the achievement distribution, lasting between 22 and 38 minutes; intercoder agreement reached Cohen's $\kappa = .74$.

Because the three self-report scales are adaptations rather than new instruments, their adaptation is described here. Each was translated into Chinese and then back-translated to check that the items retained their meaning, with discrepancies resolved by discussion rather than by dropping items, so that the scales can be read as the instruments they are adapted from. Two of the three also inherit a validated factor structure: the motivation measure is the MSLQ subscale whose intrinsic goal orientation and task value dimensions were established in the original

validation [32], and the self-directed learning measure follows the goal setting, strategy use and self-monitoring structure reported by Zimmerman [28] and used since. The present study therefore treats the factor structure of the scales as inherited rather than re-established, and reports reliability rather than a new factor solution. With 89 respondents and 36 items across the three scales, the case-to-item ratio is roughly 2.5 to 1, well below the 5 to 1 usually advised for exploratory factor analysis and far below what a confirmatory model would need; a solution estimated on this sample would be unstable, and none is reported. Scale scores are used as observed composites, and the coefficients in Table 2 should be read alongside the original validations rather than in place of them.

Table 2. Instruments, constructs, and reliability coefficients.

Instrument	Construct measured	Items	Response format	Source	α (pretest) / α (posttest)
Achievement test	Course knowledge and micro-lecture design reasoning	40 objective + 4 design	Score 0-100	Written by the teaching team, externally reviewed	n/a / .87
Motivation scale	Intrinsic goal orientation and task value	14	5-point Likert	Adapted from Pintrich et al. [32]	.79 / .83
Satisfaction scale	Perceived usefulness, material clarity, continuance intention	10	5-point Likert	Adapted from prior technology-supported course studies	.84 / .86
Self-directed learning scale	Goal setting, strategy use, self-monitoring	12	5-point Likert	Adapted from Zimmerman [28]	.81 / .85

3.6. Intervention Procedure

The intervention ran for twelve consecutive teaching weeks. In the experimental class, each topic was supported by three framework-generated resources released at the start of the week alongside the standard textbook reading, and students could request alternative explanations and extra practice items through the platform, with every request logged. The control class covered the same topics with the same syllabus and workshop sequence, using the textbook, the instructor's slides and a fixed exercise set from previous cohorts, with no additional materials and no access to the assistant. The asymmetry in material volume is treated as a confound rather than a design feature: experimental students received roughly three times as many resources and, through the request mechanism, substantially more practice and feedback

(Section 5.2).

Several features were held constant: the same contact hours, workshop exercises and weekly ten-item quiz. The quizzes differed in origin but not in coverage, since experimental quizzes came from the approved pool of generated items and control quizzes from the historical bank, with each item matched on topic and estimated difficulty beforehand. What could not be held constant is how much additional material each student worked through, since time on task was not recorded for the control class, whose practice material was distributed on paper.

3.7. Data Analysis

Analysis was carried out in SPSS 26. Baseline equivalence was tested with independent-samples t tests and χ^2 tests, and we report 95% confidence intervals for the differences rather than relying on the significance test alone, since at this sample size a non-significant difference is not evidence of equivalence.

The primary outcome, posttest achievement, was analysed with analysis of covariance using pretest score as the covariate. Four assumptions were checked and are reported with the result: normality of the residuals within groups (Shapiro-Wilk), homogeneity of error variance (Levene's test), a linear relationship between covariate and outcome within each group, and homogeneity of regression slopes. The between-group difference is reported as an adjusted mean difference with a 95% confidence interval, next to the raw difference.

Affective and skill outcomes were analysed with a mixed analysis of variance, group between subjects and time within. The group × time interaction is the confirmatory test; where it was significant, the posttest between-group comparison is reported as a difference with a 95% confidence interval and as a standardised mean difference with its own interval. Because four confirmatory outcomes were tested, the corresponding p values are also reported after Holm correction [33] within that family. Interview transcripts were analysed thematically following Braun and Clarke's six-phase procedure [34], with two coders working independently and disagreements resolved by discussion.

4. Results

4.1. Baseline Equivalence

The two classes did not differ significantly on any baseline measure. As Table 3 shows, they were comparable in gender composition, age, prior grade point average and prior experience with AI tools, and their pretest achievement scores were statistically indistinguishable (68.4

versus 67.2, t(87) = 0.58, p = .56). Self-reported motivation, satisfaction and self-directed learning at pretest also showed no significant differences, with all p values above .60.

The intervals tell a more careful story than the tests. The 95% confidence interval for the pretest achievement difference was [-2.9, 5.3] points, so a real baseline gap of several points cannot be excluded, and the same holds for the other continuous measures. With 89 students, equivalence is not something this study can demonstrate, so baseline balance is a reason for caution rather than a licence to attribute the posttest difference to the intervention.

Table 3. Baseline characteristics of the two intact classes.

Characteristic	Experimental (n = 45)	Control (n = 44)	Test statistic	p	Difference [95% CI]
Female, n (%)	15 (33.3)	14 (31.8)	$\chi^2 = 0.02$	.88	
Age, M (SD)	19.4 (0.8)	19.5 (0.8)	t = -0.58	.56	-0.1 [-0.5, 0.3]
Prior GPA, M (SD)	76.8 (6.4)	77.3 (6.1)	t = -0.38	.71	-0.5 [-3.1, 2.1]
Prior use of AI tools, n (%)	27 (60.0)	24 (54.5)	$\chi^2 = 0.27$	.60	
Pretest achievement, M (SD)	68.4 (9.6)	67.2 (10.1)	t = 0.58	.56	1.2 [-2.9, 5.3]
Pretest motivation, M (SD)	3.71 (0.62)	3.66 (0.59)	t = 0.39	.70	0.05 [-0.20, 0.30]
Pretest satisfaction, M (SD)	3.55 (0.60)	3.58 (0.63)	t = -0.23	.82	-0.03 [-0.28, 0.22]
Pretest self-directed learning, M (SD)	3.42 (0.58)	3.45 (0.61)	t = -0.24	.81	-0.03 [-0.28, 0.22]

Note. CI = confidence interval. All tests are two-tailed with df = 87. Dashes indicate categorical variables, for which a difference in means is not defined.

4.2. Academic Achievement

The experimental class scored higher on the end-of-term examination. Raw posttest means were 82.6 (SD = 8.1) and 76.4 (SD = 9.3), a difference of 6.2 points, 95% CI [2.6, 9.8], d = 0.71 [0.28, 1.14]. Analysis of covariance with pretest score as the covariate gave adjusted means of 82.3 and 76.7, an adjusted difference of 5.6 points, 95% CI [2.9, 8.3], F(1, 86) = 27.4, p < .001, partial $\eta^2 = .24$. All four assumption checks were acceptable, including homogeneity of regression slopes. The full test statistics are given in Table 4.

The distribution of scores says something the means do not: 31 of 45 experimental students (68.9%) scored above 80, against 18 of 44 (40.9%) in the control class, $\chi^2(1) = 7.03$, p = .008, and 4 against 11 scored below 70, so the difference was not confined to the top of the distribution.

Table 4. Achievement test results with analysis of covariance.

Group	n	Pretest M (SD)	Posttest M (SD)	Adjusted posttest M (SE)	95% CI for adjusted mean
Experimental	45	68.4 (9.6)	82.6 (8.1)	82.3 (0.95)	[80.4, 84.2]
Control	44	67.2 (10.1)	76.4 (9.3)	76.7 (0.96)	[74.8, 78.6]

Note. Adjusted difference = 5.6 [2.9, 8.3]; $F(1, 86) = 27.4$, $p < .001$, partial $\eta^2 = .24$; raw difference = 6.2 [2.6, 9.8], Cohen's d = 0.71 [0.28, 1.14]. Assumption checks: Shapiro-Wilk $W = .98$, $p = .31$ (experimental) and $W = .97$, $p = .24$ (control); Levene $F(1, 87) = 1.42$, $p = .24$; covariate-outcome $r = .62$ and .58 within groups; covariate-by-group $F(1, 85) = 0.71$, $p = .40$. CI = confidence interval.

The two halves of the examination moved differently: on the 40 objective items the classes scored 34.1 and 32.4, a difference of 1.7 marks, and on the four design questions 48.5 and 44.0, a difference of 4.5 marks. The proportional gap is larger on the design questions, where a well-grounded worked example should help most, but a sample of this size cannot establish that the two gaps differ reliably, so the pattern is suggestive rather than evidence that the framework acted specifically on design reasoning.

Weekly quizzes tracked the same pattern (Figure 3): the classes began close together, at 71.2 and 70.4, and the gap widened during weeks four to seven, which cover script writing and storyboarding, then held at about six points. That timing is not accidental, since those are the two topics with the densest prerequisite structure in the concept graph, and they generated 214 of the 583 requests logged across the semester.

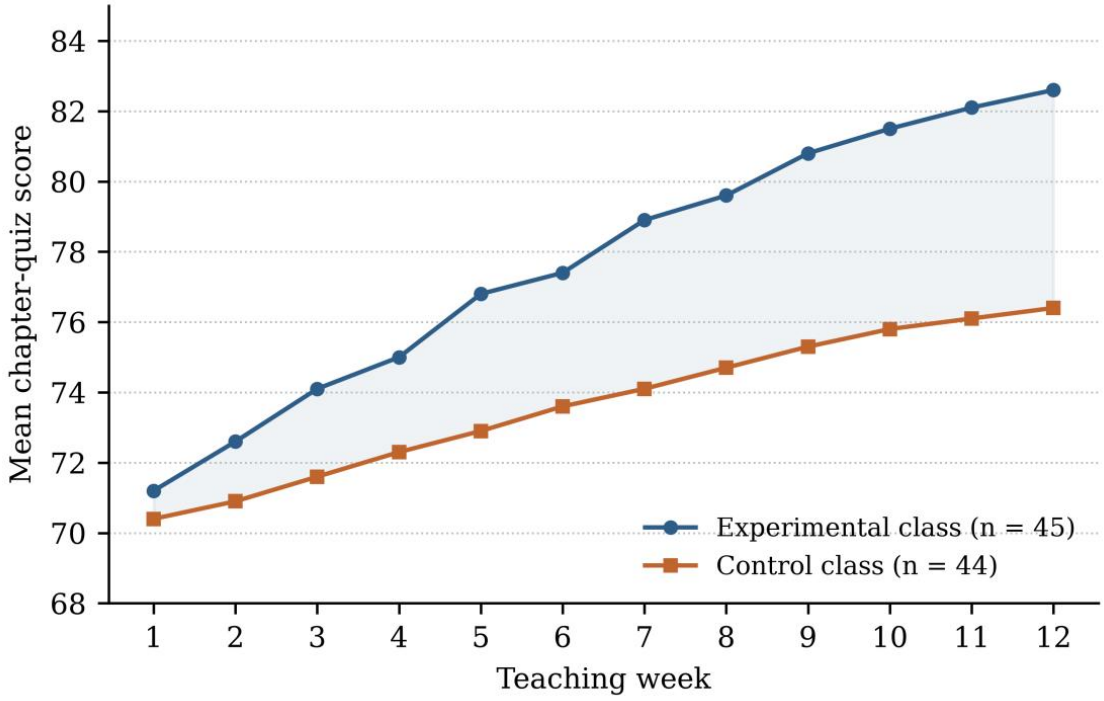


Figure 3. Mean weekly quiz scores for the two classes across twelve teaching weeks.

One further check bears on the volume confound discussed in Section 5.2. If the difference were driven mainly by students working through more material, platform use should track performance within the experimental class. It does not: request count correlated weakly with posttest score, r(43) = .14, p = .36, and students in the top quartile of requests scored 83.1 against 82.4 for the rest, t(43) = 0.36, p = .72. With 45 students this is a weak test, and a null association cannot exclude a volume effect that operated equally across the class.

4.3. Motivation and Satisfaction

Both classes became more motivated, and the framework-supported class did so to a greater degree: the group × time interaction was F(1, 87) = 11.6, p = .001, partial η^2 = .12, with posttest means of 4.19 (SD = 0.52) against 3.84 (SD = 0.57), a difference of 0.35, 95% CI [0.12, 0.58], d = 0.64 [0.21, 1.07]. Learning satisfaction followed a similar and slightly stronger pattern: 4.33 (SD = 0.48) against 3.97 (SD = 0.55), a difference of 0.36, 95% CI [0.15, 0.57], interaction F(1, 87) = 18.9, p < .001, partial η^2 = .18, d = 0.70 [0.27, 1.13]. Descriptive statistics appear in Table 5 and Figure 4; the posttest comparisons appear in Table 6.

The two kinds of statistic are not interchangeable: in Table 5 the interaction F and partial η^2 describe change over time, whereas the standardised differences in Table 6 describe the gap between classes at posttest. The multiplicity correction changes no conclusion, since every Holm-adjusted p value leads to the same decision as the uncorrected one.

Table 5. Pretest and posttest scores on the affective and skill outcomes.

Variable	Exp. pretest M (SD)	Exp. posttest M (SD)	Ctrl. pretest M (SD)	Ctrl. posttest M (SD)	F (group × time)	p	partial η^2
Learning motivation	3.71 (0.62)	4.19 (0.52)	3.66 (0.59)	3.84 (0.57)	11.61	.001	.12
Learning satisfaction	3.55 (0.60)	4.33 (0.48)	3.58 (0.63)	3.97 (0.55)	18.90	<.001	.18
Self-directed learning	3.42 (0.58)	4.05 (0.51)	3.45 (0.61)	3.71 (0.58)	5.41	.022	.06

Note. All F tests in this table are the mixed-analysis-of-variance group × time interaction, with df = 1, 87. The partial η^2 values describe that interaction; posttest between-group differences, including the covariate-adjusted achievement comparison (F(1, 86)), are reported in Table 6.

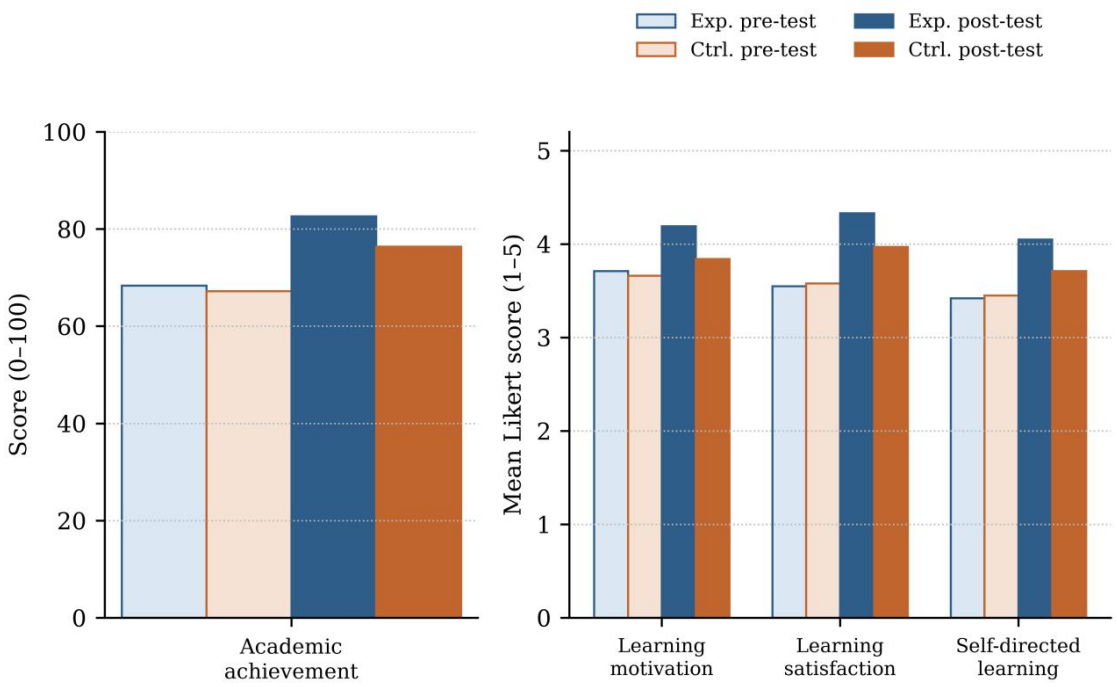


Figure 4. Pretest and posttest scores on achievement and on the three affective and skill measures.

Table 6. Posttest between-group comparisons with confidence intervals and multiplicity correction.

Variable	Difference [95% CI]	Cohen's d [95% CI]	p	p (Holm)
Academic achievement (adjusted)	5.6 [2.9, 8.3]	0.71 [0.28, 1.14]	<.001	<.001
Learning motivation	0.35 [0.12, 0.58]	0.64 [0.21, 1.07]	.001	.003
Learning satisfaction	0.36 [0.15, 0.57]	0.70 [0.27, 1.13]	<.001	<.001
Self-directed learning	0.34 [0.11, 0.57]	0.62 [0.19, 1.05]	.022	.044

Note. Achievement is the covariate-adjusted difference in examination marks; the remaining rows are posttest differences in scale points. Holm = Holm-Bonferroni correction within the family of four confirmatory comparisons. CI = confidence interval.

4.4. Self-Directed Learning and Creative Design Performance

Self-directed learning followed the affective measures with a smaller margin: posttest means of 4.05 (SD = 0.51) and 3.71 (SD = 0.58), a difference of 0.34, 95% CI [0.11, 0.57], d = 0.62 [0.19, 1.05], interaction $F(1, 87) = 5.4$, $p = .022$, partial $\eta^2 = .06$. This is the weakest of the significant comparisons and the one that would change if a stricter correction were used; Holm correction within the family leaves it significant at .044, and we read it as tentative rather than settled.

A fifth measure is noted only for completeness. A five-dimension creative design rubric

produced a nonsignificant difference of 0.07 scale points, 95% CI [-0.22, 0.36], p = .25, with a ceiling on three of the five dimensions. Because it measured a transfer the framework was not built to produce and could not discriminate among competent work, the result is uninformative in either direction and is not treated as a finding.

4.5. Resource Quality: Multimedia Design Principles and Difficulty Calibration

The principles cited in Section 2.3 are the criteria against which the micro-lectures students were asked to produce would be judged, so a framework that models them should satisfy them. All 36 released resources were audited against five of these principles by two raters, with agreement of Cohen's κ = .71 (Table 7).

Table 7. Audit of released resources against five multimedia design principles.

Multimedia principle	Operational check applied to the released resource	Resources passing, n = 36 (%)
Segmenting	Content divided into learner-paced parts of 120 seconds or less	33 (92)
Signalling	Key elements explicitly marked in the script or in the framing	30 (83)
Redundancy	No duplication of spoken narration as on-screen text	35 (97)
Temporal contiguity	Narration and the imagery it describes presented together	28 (78)
Modality	Spoken narration used in place of on-screen prose alongside imagery	34 (94)

Note. Rated by two raters from scripts and storyboards; inter-rater agreement, Cohen's κ = .71. From week six the screening step added checks for signalling and temporal contiguity, after which compliance with contiguity rose from 64% to 89% and with signalling from 72% to 91%.

Compliance was high on three principles and weak on the other two, and the weak two were exactly the properties the screening step did not check before week six: screening on textual properties did not transfer to design properties visible only when narration and image are considered together.

Difficulty calibration can be assessed in the same way, since Module 3 scores the practice items that accompany each resource. Mean realised accuracy was .72, inside the .60 to .85 target in 29 of 36 resources (81%), and misses were concentrated in the first three weeks, when 5 of 12 resources fell outside the band against 2 of 24 thereafter. Because realised accuracy is a post-hoc quantity while the band was set in advance, this checks calibration rather than tests it.

4.6. Resource Development Efficiency

The teacher-side effects rest on logs rather than self-report and are the most directly interpretable figures in the study. Table 8 compares the conventional workflow used with the previous cohort against the framework-supported workflow used here, per teaching unit of four class hours, across the 16 units of the course.

Table 8. Resource development efficiency per teaching unit of four class hours.

Indicator	Conventional workflow	Framework-supported workflow	Change
Drafting time per unit, h	4.5	1.1	-76%
Review and editing time per unit, h	1.0	1.7	+70%
Other preparation time per unit, h	1.3	0.4	-69%
Total preparation time per unit, h	6.8	3.2	-53%
Resources released per unit	1	3	+200%
Preparation time per released resource, h	6.8	1.07	-84%
Share of preparation time requiring disciplinary judgement, %	31	74	+43
Disciplinary-judgement time per unit, h	2.1	2.4	+14%
Global workload rating per unit (1-5), M (SD)	4.31 (0.28)	3.14 (0.41)	-1.17
First-cohort cost per unit including knowledge base, h	6.8	20.7	+204%

Note. Conventional figures come from the previous cohort of the same course (n = 47), taught by the same instructor. Unit = four class hours; with 16 units, paired comparisons have df = 15. Knowledge-base construction is a one-off cost of about 280 hours, excluded from the steady-state rows and included in the first-cohort row; the approach reaches break-even after roughly 78 teaching units of reuse.

Total preparation time fell from 6.8 to 3.2 hours per unit, a paired difference across the 16 units of -3.6 hours, $t(15) = -11.4$, $p < .001$, and its composition changed: drafting fell from 4.5 to 1.1 hours while review and editing rose from 1.0 to 1.7 hours. The task shifted from authoring to reviewing.

The saving needs three qualifications, and they matter more than the headline. Released resources per unit tripled, so on a per-resource basis the comparison is not 3.2 against 6.8 hours but 1.07 against 6.8: part of the gain was reinvested in more varied material rather than in

reduced workload. The work also became more concentrated. The instructor judged that 74% of the framework-workflow time required disciplinary judgement, against 31% before, so although total preparation time roughly halved, the absolute time spent on disciplinary verification changed little, from about 2.1 to 2.4 hours per unit. The global workload rating fell from 4.31 (SD = 0.28) to 3.14 (SD = 0.41), a paired difference of -1.17, t(15) = -9.7, p < .001, consistent with less total effort but a heavier demand per minute. Finally, these figures describe steady-state operation and exclude the cost of building the knowledge base: charged to the first cohort alone, that cost raises the figure from 6.8 to about 20.7 hours per unit, and the approach breaks even only after roughly 78 teaching units, or five cohorts.

Review effort rose by 70%: 68% of released resources were accepted without edit, 22% needed minor revision and 10% were discarded, and the discard rate fell from 19% in the first three weeks to 4% by week eight, suggesting that the templates and the knowledge base improved through use.

4.7. Student Perceptions

Thematic analysis of the twelve interviews produced four themes. Perceived relevance came up most often: students valued the fit between the generated material and the specific teaching points they were working on, and several contrasted it with general-purpose chatbots, which produced answers that sounded right but did not match the workshop's approach. Cognitive load was the second theme: the tiered resources made it easier to start a topic, but a smaller group said the abundance of material created a new problem, namely which of three variants to read first. Trust was the third, and it pointed in two directions at once: students trusted the material because the instructor had approved it, distrusted general chatbots for the mirror-image reason, and yet credited the discussion that followed a generated explanation, rather than the explanation itself, for their improvement. That suggests the teacher-in-the-loop stage works as a credibility signal and not only as a quality filter.

The fourth theme, raised by six students, was fairness. Two in the experimental class worried that access to the assistant gave them an advantage they had not earned, and four asked at the disclosure session whether the control class had been disadvantaged. Two control-class students had asked during the semester why the classes were taught differently and were told only that the difference would be explained later, which protected the comparison but left a legitimate question unanswered. After disclosure, 29 of 44 control-class students (66%) downloaded at least one unit's set within two weeks and nine attended an additional tutorial; delayed provision did not resolve the equity question, because half the cohort still spent a semester without a

support that the study suggests was useful.

5. Discussion

5.1. Interpreting the Achievement Difference

The central result is a difference of about 5.6 points between two intact classes in a course that would seem to resist automated support. Micro-lecture design is procedural, judgement-heavy and tool-specific: quality depends on decisions experienced teachers rarely put into words, and generic models produce scripts that read fluently while ignoring the principles that make a short video learnable. Because assignment was not random, and both classes were taught by the instructor who built the framework, the difference is an association and not an effect, and three features of the design are plausible candidates for it, none separable from the others.

Grounding is the first: restricting generation to passages retrieved from a course-specific knowledge base constrains the model to the terminology, design conventions and evaluation criteria students meet in the workshop, which addresses the mismatch they themselves described. Difficulty calibration is the second: because every request carried an explicit difficulty band and a representation preference, material was pitched at the level the learner profile indicated, and the check reported in Section 4.5 shows the target was met for most resources. This connects to the personalisation literature, addressing the representation and explanation dimensions that item-response systems handle poorly [20], and operationalising, crudely, the graded fading of support that cognitive load theory recommends [21]. Teacher review is the third: it did not remove the instructor but moved the effort from composing to verifying and, in TPACK terms, left pedagogical content knowledge with the person who had it, so the difference reflects generated resources, expert review and the additional material released together, which this study cannot separate.

5.2. What This Design Cannot Separate

Three rival explanations are worth naming, and one is more plausible than the framework-based account. The first is the bundle itself: the framework combined a knowledge base, retrieval grounding, difficulty calibration, teacher review, three variants per objective, automated screening and a request mechanism, and the study reports a single estimate for all of them together. A difference attributed to "the framework" is therefore a statement about a package that could be more or less effective than any of its parts.

The second explanation is the most plausible of the three and the least exotic: the

experimental class received more of everything — three resources per unit instead of one, a mechanism for requesting further explanations and further practice, and a teacher who was reviewing material rather than writing it. Increasing the quantity of practice with feedback is one of the better-established ways to raise achievement, and this study did not hold it constant. The platform logs offer a partial answer: request frequency was unrelated to posttest score within the experimental class, which argues against a simple dosage account for the more engaged students but cannot rule out an effect operating for everyone. Until the comparison is made, the conservative reading is that this study evaluated a resource-rich package, not an AI component within a fixed quantity of instruction.

The third is differential attention. A teacher who designed a framework and knows which class is receiving it is not a neutral instrument, and expectancy effects are hard to exclude when the same person teaches both groups. Measurement was constrained by a fixed marking scheme and blind scoring, but delivery was not, so the estimate is better read as an upper bound. A design that could decompose the package needs at least three arms beyond a control: the full framework, a grounded-only arm with calibration and the variant mechanism disabled, and a volume-matched arm in which the control class receives three conventional resources and an equivalent practice set with no generative component. The third arm matters most, because it separates the technology question from the quantity question.

5.3. Implications for Practice

The knowledge base is the binding constraint. The efficiency gain came from keeping generation on grounded content, so the quality of that grounding decides whether the output is useful or merely fluent. An institution considering a similar approach should budget for knowledge-base construction as the main cost, not for model access, which is cheap by comparison.

The cost structure is not what the headline number suggests. Preparation time per unit halved, but per released resource it fell much further, and the absolute time spent on disciplinary verification barely moved; the pattern to expect is a redistribution towards less drafting, more verification, more volume and more concentration. In TPACK terms a framework of this kind can supply content, representational and assessment knowledge, but the judgement about what counts as good work in a given course stays with the teacher, which is why the review stage is not a transitional safeguard that better models will make unnecessary.

There is a student-facing implication too: some students handled three variants well and others did not know where to start, so a short orientation in the first week, and possibly a

recommended reading order in the interface, would fix a problem the framework created rather than solved. Finally, on these figures the approach is more expensive than the conventional workflow for at least the first cohort and becomes cheaper only on reuse — a normal profile for an investment in reusable material, but not what "preparation time fell by half" conveys on its own.

5.4. Limitations

The most serious limitation is the non-random assignment of classes. The groups were equivalent on every baseline measure collected, but the confidence intervals in Table 3 show how little that establishes at this sample size, unmeasured differences between intact classes cannot be ruled out, and the design does not support causal claims. The instructor also designed the framework, taught both classes and knew the group assignment, so expectations cannot be entirely excluded even though the examination scheme was fixed in advance and the design questions were scored blind. Replication with instructors who did not build the system is needed before the results travel.

The second limitation is the bundle described in Section 5.2: the study cannot attribute the difference to grounding, calibration, teacher review or added volume, and the volume explanation is the most obvious rival account. Time on task was not measured for the control class, and the study could detect only effects of about $d = 0.60$ or larger, so the self-directed learning result would not have been detected had it been smaller. The scope is narrow for the same reason: one course, one institution, 89 students, twelve weeks, with the posttest collected immediately after the intervention, so nothing here speaks to retention.

The self-report scales carry a further limitation of the same kind: they were translated and back-translated but not re-validated in this cohort, so their construct validity rests on the original instrument validations rather than on evidence collected here.

Finally, the ethics of unequal access deserve to be stated rather than acknowledged in passing. Half the cohort spent a semester without a support that appears to have helped their peers, a deliberate design choice made to permit a comparison. Releasing the materials after the posttest mitigates but does not undo the asymmetry, and in an institutional deployment the question of who receives AI-supported instruction, and on what basis, would have to be settled as policy rather than as an experimental condition.

6. Conclusion

This study asked whether generative AI can support instruction in a course whose outcomes

are made rather than recalled, when it is organised around a defined framework instead of used as a general-purpose assistant. The evidence is consistent with a moderate positive association: after adjusting for pretest achievement, the framework-supported class scored 5.6 points higher, reported higher motivation and satisfaction, and showed a smaller gain in self-directed learning, though the design cannot separate the framework from the extra resources that came with it.

Two conclusions survive those qualifications. Grounding and calibration are not optional refinements: general models produce fluent material about teaching and learning, and fluent material that is subtly wrong is worse than nothing for students who cannot yet tell the difference, so restricting generation to a curated knowledge base and fixing the difficulty of output against a learner profile are what made generation usable as instruction here. And the teacher stays in the loop, which is where the pedagogy lives. The most defensible claim the study supports is not that generative AI teaches, but that a constrained generative workflow changes what a teacher can produce in the time available, at the price of a substantial upfront investment.

Three questions follow. Does the association hold when the volume of material is held constant, as in the ablation study described in Section 5.2? Does it persist beyond the semester and outside the course for which the knowledge base was built? And how should the benefits be distributed when access cannot be universal and a comparison requires that it is not?

Fund Project

This paper is the research result of the general project of the 2024 Scientific Research Project of the Hunan Provincial Department of Education, "Research on Personalized Teaching Resource Design Based on Generative AI Technology" (Project No.: 24C1265).

References

[1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998-6008. https://arxiv.org/abs/1706.03762

[2] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., & Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877-1901. https://arxiv.org/abs/2005.14165

[3] Kasneci, E., Sessler, K., Kuechemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Guennemann, S., Huellermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., Stadler, M., Weller, J., Kuhn, J., & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. Learning and Individual Differences, 103, 102274. https://doi.org/10.1016/j.lindif.2023.102274

[4] Chiu, T. K. F., Xia, Q., Zhou, X., Chai, C. S., & Cheng, M. (2023). Systematic literature review on opportunities, challenges, and future research recommendations of artificial intelligence in education. Computers and Education: Artificial Intelligence, 4, 100118. https://doi.org/10.1016/j.caeai.2022.100118

[5] UNESCO. (2023). Guidance for generative AI in education and research. Paris: UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000386693

[6] Holmes, W., Bialik, M., & Fadel, C. (2019). Artificial intelligence in education: Promises and implications for teaching and learning. Boston, MA: Center for Curriculum Redesign. https://curriculumredesign.org/our-work/artificial-intelligence-in-education/

[7] Celik, I., Dindar, M., Muukkonen, H., & Jaervelae, S. (2022). The promises and challenges of artificial intelligence for teachers: A systematic review of research. TechTrends, 66(4), 616-630. https://doi.org/10.1007/s11528-022-00715-y

[8] Crompton, H., & Burke, D. (2023). Artificial intelligence in higher education: The state of the field. International Journal of Educational Technology in Higher Education, 20(1), 22. https://doi.org/10.1186/s41239-023-00392-8

[9] Zawacki-Richter, O., Marin, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education: Where are the educators? International Journal of Educational Technology in Higher Education, 16(1), 39. https://doi.org/10.1186/s41239-019-0171-0

[10] Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., Li, X., Jin, Y., & Gasevic, 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

[11] Mishra, P., & Koehler, M. J. (2006). Technological pedagogical content knowledge: A framework for teacher knowledge. Teachers College Record, 108(6), 1017-1054. https://doi.org/10.1111/j.1467-9620.2006.00684.x

[12] Celik, I. (2023). Towards intelligent-TPACK: An empirical study on teachers' professional knowledge to ethically integrate artificial intelligence (AI)-based tools into education. Computers in Human Behavior, 138, 107468. https://doi.org/10.1016/j.chb.2022.107468

[13] Shieh, D. (2009). These lectures are gone in 60 seconds. The Chronicle of Higher Education, 55(26), A1, A13.

[14] Guo, P. J., Kim, J., & Rubin, R. (2014). How video production affects student engagement: An empirical study of MOOC videos. In Proceedings of the First ACM Conference on Learning at Scale (pp. 41-50). New York, NY: ACM. https://doi.org/10.1145/2556325.2566239

[15] 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), 43. https://doi.org/10.1186/s41239-023-00411-8

[16] Cooper, G. (2023). Examining science education in ChatGPT: An exploratory study of generative artificial intelligence. Journal of Science Education and Technology, 32(3), 444-452. https://doi.org/10.1007/s10956-023-10039-y

[17] Bloom, B. S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. Educational Researcher, 13(6), 4-16. https://doi.org/10.3102/0013189X013006004

[18] Xie, H., Chu, H.-C., Hwang, G.-J., & Wang, C.-C. (2019). Trends and development in technology-enhanced adaptive/personalized learning: A systematic review of journal publications from 2007 to 2017. Computers & Education, 140, 103599. https://doi.org/10.1016/j.compedu.2019.103599

[19] Pane, J. F., Steiner, E. D., Baird, M. D., & Hamilton, L. S. (2017). Informing progress:

Insights on personalized learning implementation and effects. Santa Monica, CA: RAND Corporation. https://www.rand.org/pubs/research_reports/RR2042.html
[20] Zhang, L., Basham, J. D., & Yang, S. (2020). Understanding the implementation of personalized learning: A research synthesis. Educational Research Review, 31, 100339. https://doi.org/10.1016/j.edurev.2020.100339
[21] Sweller, J., van Merrienboer, J. J. G., & Paas, F. (2019). Cognitive architecture and instructional design: 20 years later. Educational Psychology Review, 31(2), 261-292. https://doi.org/10.1007/s10648-019-09465-5
[22] Vygotsky, L. S. (1978). Mind in society: The development of higher psychological processes. Cambridge, MA: Harvard University Press.
[23] van Merrienboer, J. J. G., & Kirschner, P. A. (2018). Ten steps to complex learning: A systematic approach to four-component instructional design (3rd ed.). New York, NY: Routledge. https://doi.org/10.4324/9781315113210
[24] Merrill, M. D. (2002). First principles of instruction. Educational Technology Research and Development, 50(3), 43-59. https://doi.org/10.1007/BF02505024
[25] Mayer, R. E., & Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia learning. Educational Psychologist, 38(1), 43-52. https://doi.org/10.1207/S15326985EP3801_6
[26] Black, P., & Wiliam, D. (1998). Assessment and classroom learning. Assessment in Education: Principles, Policy & Practice, 5(1), 7-74. https://doi.org/10.1080/0969595980050102
[27] Hattie, J., & Timperley, H. (2007). The power of feedback. Review of Educational Research, 77(1), 81-112. https://doi.org/10.3102/003465430298487
[28] 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
[29] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), 1-38. https://doi.org/10.1145/3571730
[30] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Kuttler, H., Lewis, M., Yih, W., Rocktaeschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33, 9459-9474. https://arxiv.org/abs/2005.11401
[31] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems, 35, 24824-24837. https://arxiv.org/abs/2201.11903
[32] Pintrich, P. R., Smith, D. A. F., Garcia, T., & McKeachie, W. J. (1993). Reliability and predictive validity of the Motivated Strategies for Learning Questionnaire (MSLQ). Educational and Psychological Measurement, 53(3), 801-813. https://doi.org/10.1177/0013164493053003024
[33] Holm, S. (1979). A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics, 6(2), 65-70.
[34] Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. Qualitative Research in Psychology, 3(2), 77-102. https://doi.org/10.1191/1478088706qp063oa