Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

The division of epistemic work in learning ecologies with Generative AI

Jiaying Xiao
Minzu University of China
Yi Dai
yidai@cityu.edu.mo
City University of Macau
Yuanyuan Feng
East China Normal University
Fu Xu
Minzu University of China
Huiyong Li
Kyushu University

Article

Keywords:

Posted Date: November 20th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11296160/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

The division of epistemic work in learning ecologies with Generative AI

Jiaying Xiao[1], Yi Dai[2*], Yuanyuan Feng[3], Fu Xu[1] and Huiyong Li[4]

[1] School of Education, Minzu University of China, Beijing 100000, China

[2] School of Education, City University of Macau, Macau SAR, China

[3] School of Foreign Languages, East China Normal University, Shanghai, China

[4] Research Institute for Information Technology, Kyushu University, Fukuoka, Japan

* Correspondence: yidai@cityu.edu.mo

Abstract

Generative AI redistributes the work of producing and evaluating knowledge in teacher-designed learning tasks, but how students retain direction while judging its outputs remains unclear. We examined how teacher instructions and materials enter student–GenAI dialogue, how reasoning display and web search relate to interaction time, what students evaluate and verify, and how groups coordinate systems. We analysed 81 screen recordings from 29 graduate student groups completing educational design tasks, coding 1,149 events and 275 prompts. Students relayed task requirements, steered successive requests and transferred outputs across platforms. Task directives persisted across turns, whereas references to course materials declined and elaboration requests increased. Reasoning display and web search were each associated with approximately 65% longer machine turns and a 19-percentage-point larger machine share of coded interaction time. Across 150 turns displaying reasoning and 86 with web retrieval, the recordings showed no reopening of a reasoning trace or inspection of a cited source. Students corrected pedagogical fit while asking the machine to assure factual accuracy. These findings distinguish control over the dialogue from responsibility for establishing credibility. Learning tasks can make that responsibility visible by requiring students to justify a consequential claim against its source or course material before using it.

Introduction

When students complete a teacher-designed assignment with generative artificial intelligence (GenAI), deciding what to ask, which resources to admit and whether to accept an answer becomes part of the learning activity. Contemporary systems generate explanations, retrieve information and sustain sequences of requests, drawing on capabilities developed in language-model and agent research[1–3]. Students can also change what those systems make visible: enabling reasoning display exposes intermediate textual steps, and enabling web search brings external material into the response. Because these settings alter what can be inspected without altering what must be produced, they make the division of knowledge work observable within a single assignment.

That division matters because assisted performance and learning can diverge. A meta-analysis reports benefits of ChatGPT across several learning-related outcomes[4], yet a school mathematics experiment shows that improved performance during supported practice can coexist with poorer subsequent unaided performance[5]. Process studies identify lower mental effort and shallower inquiry[6] and shifts in self-regulation described as metacognitive laziness[7]. These results give substance to concerns about over-reliance and classroom adoption[8–10], but they share a limitation: an acceptable answer does not reveal which judgments the student made. Identifying those judgments requires tracing how assistance is used inside an assigned activity, and bringing the teacher's design into the analysis, which addresses the limited attention to educators identified in earlier research[11].

Existing accounts of technology in the didactical relation explain how cognitive activity is mediated and distributed. Rezat and Sträßer's socio-didactical tetrahedron treats artefacts as constituents that mediate relations among teacher, student and content[12], and Salomon and colleagues distinguish performance achieved with a technology from cognitive effects that persist after its removal[13]. Distributed cognition spreads cognitive function across people and

artefacts[14,15], and instrumental genesis explains how users turn an available artefact into an instrument for their purposes[16]. Within this distributed activity, producing a claim and establishing grounds for accepting it are distinct functions, and a distributed system can allocate them to different components or leave one of them unassigned. This distinction organises our analysis.

Accounts of machine agency and hybrid intelligence make the allocation of control explicit but leave open which function that control governs. Machine agency is theorised through both system capabilities and users' responses to them[17], and analyses of learner–GenAI relationships show why an analogy with a conventional tool is insufficient[18]. Hybrid-intelligence frameworks distribute control and regulation among learners, teachers and AI[19–21]. Within one exchange, selecting the next prompt, generating an explanation and judging whether a claim is credible may fall to different actors. Control over a dialogue therefore does not specify who is answerable for its content.

Teaching enters such a dialogue through the students who operate it. Classroom orchestration covers activity design and the monitoring that accompanies its enactment[22,23], and teacher-awareness systems make ongoing activity available for intervention[24]. In a chatbot dialogue, instructions and course resources usually reach the system because students put them there. Brousseau's didactic contract directs attention to the expectations that organise responsibility for knowledge[25], and those expectations are enacted through what students choose to forward and what contribution they request. We call this transmission relay and distinguish two routes: directive relay carries a task statement or its required components, whereas epistemic relay supplies course materials as grounds for the answer. A requirement specifies what must be produced; a reading can also supply criteria for judging what is produced. Gradual release of responsibility concerns learners' increasing responsibility for a task[26], so a student's growing

control over prompts need not displace the teacher's agenda, and need not transfer responsibility for evaluating the result.

The two features offer different resources for that evaluation. Chain-of-thought prompting elicits intermediate steps[27] and reasoning models generate extended traces[28], while search-enabled systems retrieve external material through approaches related to retrieval-augmented generation[29]. A trace presents a sequence a student can examine; a citation presents a source against which a claim can be checked. Worked-example research explains why an interpretable sequence of steps can support learning[30], and these features provide opportunities to examine a generated sequence and consult its supporting sources. Whether they do depends on the cost of doing so: requiring deliberate engagement with AI recommendations reduces over-reliance[31], and explanations reduce over-reliance when they lower the relative cost of verification[32]. The stakes are set by the material itself, since generated reasoning can omit the influences on an answer[33] and cited material may not support the claims a generative search engine makes[34]. Lateral reading offers one way to investigate a source beyond the page presenting a claim[35], and research separating information access from memory and perceived understanding[36,37], together with analyses of illusions of understanding in AI-assisted research[38], explains why access alone is an incomplete indicator of epistemic engagement.

These features also redistribute time. Reasoning and retrieval lengthen the interval between a request and a completed response, altering when students read, discuss or compose. We use interactional floor to describe the temporal distribution of coded machine and student activity. ICAP links learning to the cognitive activities learners undertake[39], so a longer machine turn acquires educational meaning only through what students do within and after it; cognitive offloading[40] and work on critical thinking and metacognitive demands in GenAI use[41,42] identify delegation, monitoring and verification as the processes at stake. Studies combining screen

activity with other process data[43,44] and reviews of self-regulation traces[45] show how such processes become visible in sequences of action.

Interpreting those actions requires specifying what is being evaluated. Epistemic dependence is unavoidable when people use knowledge they cannot establish themselves[46], so the question is not whether students depend on GenAI but whether that dependence is warranted. Epistemic-cognition theory distinguishes aims for knowing from the sources, justifications and processes used to pursue them[47], and trust-calibration research asks whether reliance corresponds to system capability[48,49]. In an educational design task, students may hold substantial knowledge of their intended learners while depending on GenAI for factual material, so a correction aimed at the suitability of a teaching plan can coexist with acceptance of the factual claims supporting it. Asking for a plan better suited to adult learners applies a contextual criterion; asking for authentic references states a criterion of accuracy; opening and examining a reference applies that criterion to evidence. Coding corrections alongside verification requests and source-inspection actions therefore separates expressed standards from exercised ones.

Responsibility extends beyond a single dialogue when students use several systems. Multi-agent research examines coordination among systems with differentiated roles[2,3], and in classroom use students may supply that coordination themselves by repeating requests, comparing answers and transferring material between platforms, developing practices around the artefacts available to them[16]. Once an answer has moved, its origin and evidential basis become harder to follow, which makes coordination an epistemic act as well as a procedural one.

We examined these processes through 81 screen recordings of graduate student groups working on educational design tasks. Students selected their own combinations of reasoning display and web search and could change them during a task, which preserved their ordinary workflow and allowed four observed configurations to be compared. Figure 1 traces the classroom workflow

from teacher-designed tasks through student–GenAI dialogue to recording, configuration classification and analysis. Event and prompt coding addressed four questions. RQ1: How do teacher instructions and materials enter student–GenAI dialogue, and how does relay relate to student-authored steering across turns? RQ2: How is feature configuration associated with the allocation of coded interaction time? RQ3: What aspects of an answer do students evaluate, and what verification do they request from the machine? RQ4: How do groups coordinate platforms and configurations within a task?

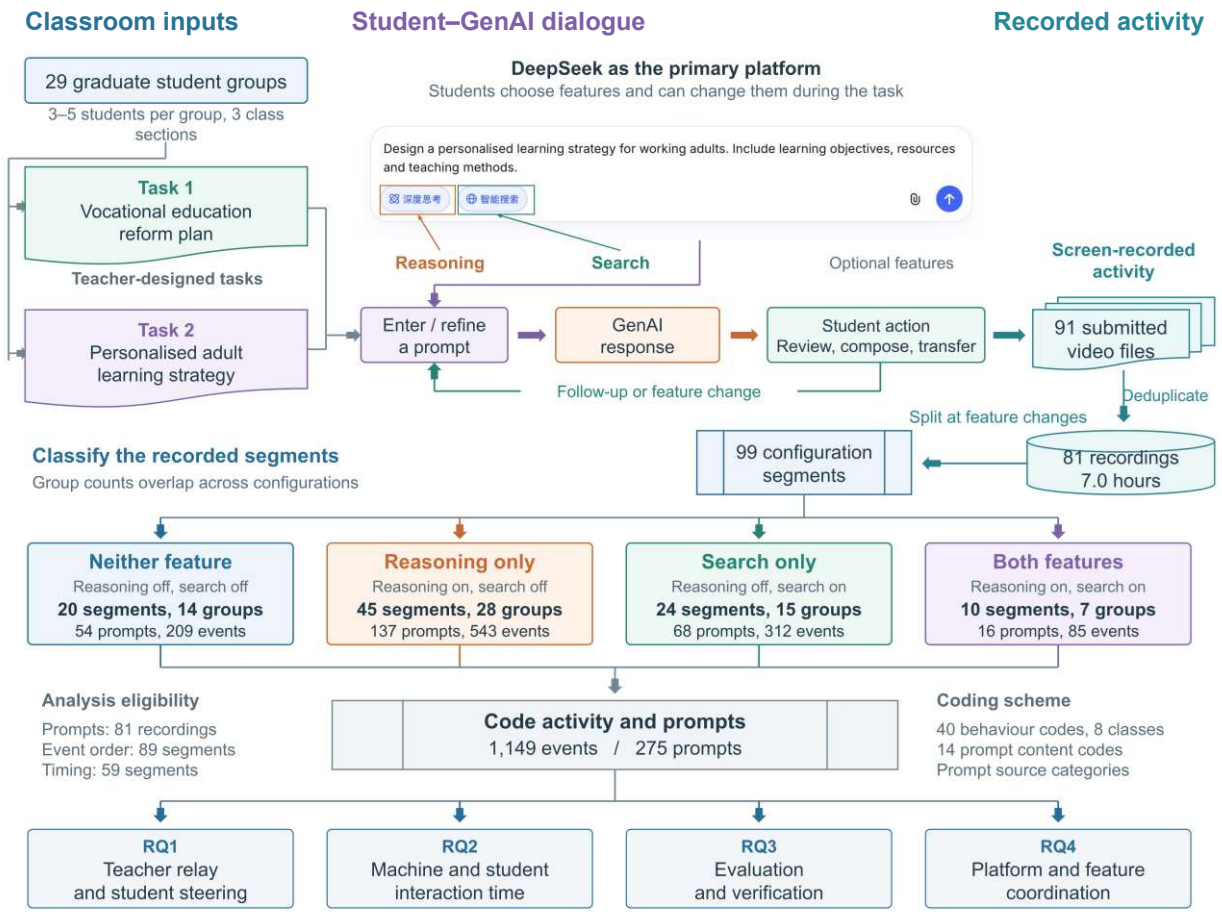

Fig. 1 | Classroom interaction and the path from recordings to analysis. Twenty-nine groups of three to five graduate students in three class sections worked on teacher-designed educational tasks. DeepSeek was the primary platform; other GenAI systems were permitted. Students chose reasoning display and web search and could change these settings during a task. The dialogue loop represents follow-up requests and feature changes; review, composition and transfer are alternative student actions. The DeepSeek interface screenshot contains a demonstration prompt entered in October 2026, separate from the study recordings. Deduplication reduced 91 submitted files to 81 recordings totalling 7.0 hours. Feature changes defined 99 segments, classified into four observed configurations. Coloured branches report segments, contributing groups, prompts and events; group counts overlap across configurations. Coding yielded 1,149 events and 275 prompts. Prompt analysis used all 81 recordings, event-order analysis used 89 segments and timing analysis used 59 segments. The four research questions connect these records to teacher relay, interaction time, evaluation and coordination.

Results

Sample and data

The dataset comprised 81 recordings totalling 7.0 hours from 29 groups in three class sections at a university in Macao. Nineteen groups submitted recordings for one task and ten for both. Task 1 required a vocational-education reform plan addressing specified components. Task 2 required a personalised learning strategy for a chosen adult learner group, drawing on three readings and course principles. DeepSeek was the primary platform for all groups, which were permitted to use other GenAI systems. Changes in students' selected configurations divided the recordings into 99 segments (Table 1).

Table 1 | Analytic samples by feature configuration

Configuration	Groups	Segments	Prompts	Coded events	Timed segments	Post-answer cycles	Timed cycles
No feature	14	20	54	209	14	43	36
Reasoning display only	28	45	137	543	26	89	86
Web search only	15	24	68	312	14	52	51
Reasoning display + web search	7	10	16	85	5	10	8
Total	29	99	275	1,149	59	194	181

A segment is a continuous sequence under one feature configuration; groups can contribute to multiple configurations. Prompt analysis includes all 81 recordings. Event-order analyses exclude 10 retrospective recordings, leaving 89 segments. Timing analyses use 50 complete recordings. The 194 post-answer cycles contain an observed student action after the machine response; the 181 timed cycles have usable machine-duration data. The paired duration plot uses 124 cycles bounded by successive send events within the same segment.

Teacher relay and student steering

Teacher content entered prompts through two routes. In Task 1, 45.0% of prompts contained the teacher's task statement, 30.8% had a teacher-derived source and 10.0% came from question lists prepared by groups. In Task 2, 0.6% of prompts reproduced the task statement and 32.9% requested grounding in course materials, usually readings or lecture slides supplied as files or screenshots (odds ratio (OR) for material grounding in Task 2 versus Task 1 = 2.17, 95% confidence interval (CI) 1.08–4.33). Figure 2a compares task-statement content across tasks.

The two routes diverged across turns (Fig. 2b). In Task 2, material grounding appeared in 48.9% of first prompts and 14.3% of prompts at turn 4 or later (OR per additional turn = 0.65, 95% CI

0.49–0.87, P = 0.003). In Task 1, directive anchoring appeared in 47.2% and 48.4%, respectively (OR per turn = 0.94, 95% CI 0.71–1.25, P = 0.69). Groups divided the teacher's specification into successive requests for reform goals, implementation steps, expected effects and challenges, so the task remained a running agenda while students determined the sequence of requests.

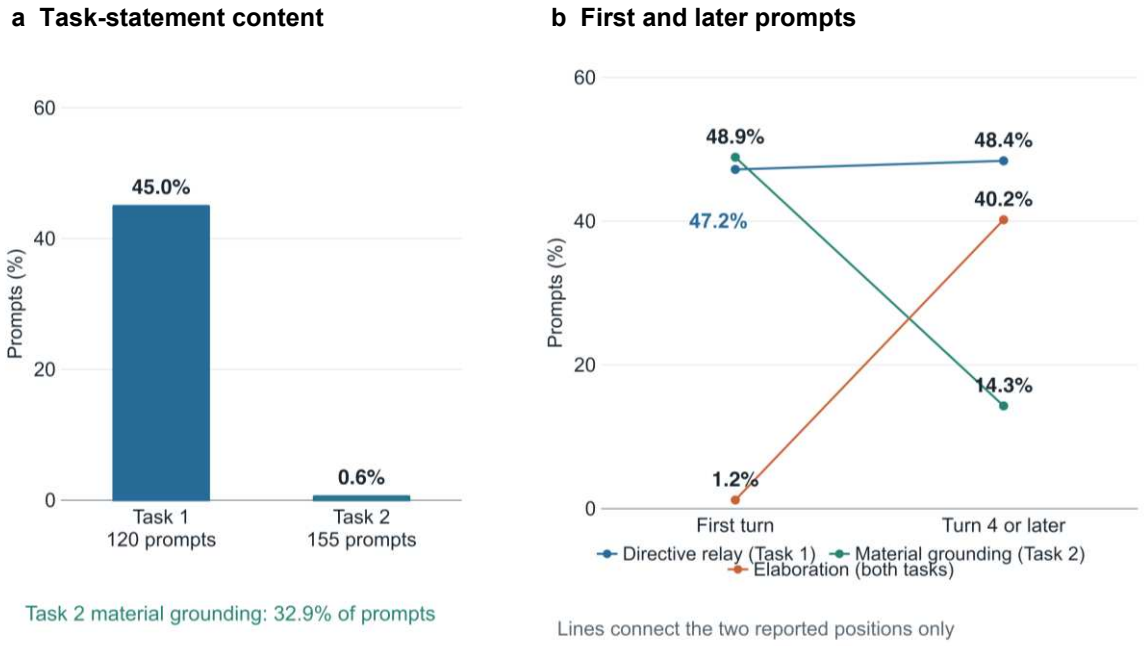

Fig. 2 | Teacher relay and student steering. a, Prompts containing the task statement (content code Q1), shown separately for Task 1 (120 prompts) and Task 2 (155 prompts). b, Percentages at the first turn and at turn 4 or later. Directive relay uses the teacher-anchor indicator in Task 1, epistemic relay uses course-material grounding in Task 2, and elaboration is pooled across tasks. Lines connect the two displayed positions. The teacher-anchor indicator combines Q1 with teacher-derived source; panel a uses Q1 alone.

Elaboration requests increased from 1.2% of first prompts to 40.2% at turn 4 or later (OR per turn = 1.55, 95% CI 1.31–1.84, P < 0.001). Corrections and redirections also increased (OR = 1.22, 95% CI 1.00–1.47, P = 0.045). Students assigned the machine a role in 32.1% of first prompts, commonly an adult-education expert or teacher, and five prompts (1.8%) used machine-suggested follow-ups. Web search was associated with more information-seeking prompts (OR = 3.38, 95% CI 1.17–9.77) and less course-material grounding (OR = 0.44, 95% CI 0.25–0.77; Table 2). In one recording, DeepSeek prevented web search while files were attached, so course material and external evidence could not be combined in a single request.

Allocation of interaction time

Mean machine shares of coded interaction time were 49.1% without either feature, 65.8% with reasoning display, 60.8% with web search and 86.7% with both (59 fully timed segments; Fig. 3a). A linear mixed model with task adjustment and a group random intercept estimated increases of 18.7 percentage points for reasoning display (95% CI 8.6–28.8) and 19.2 points for search (95% CI 8.5–29.9), both $P < 0.001$, with an interaction test of $P = 0.81$. Group fixed-effects estimates from 48 segments in 14 groups using multiple configurations were 19.5 points for reasoning (95% CI 4.1–35.0) and 22.2 points for search (95% CI 6.4–37.9). DeepSeek-only estimates were 24.8 and 16.5 points (Table 2).

Across the 181 timed cycles, median machine-turn duration was 35.5 s without features, 60.5 s with reasoning display, 56.0 s with search and 68.5 s with both. Each feature was associated with a duration ratio of approximately 1.65 (reasoning 95% CI 1.25–2.17; search 1.25–2.19; both $P < 0.001$), with similar within-group estimates. In 22 turns with an interface-reported reasoning duration, the mean was 20.3 s (range 15–25 s); where a retrieved-page count was available, the mean was 37 (range 3–50). In the paired sample (Fig. 3b), median post-answer student-event durations were 32, 32, 35 and 9 s, and models for this outcome yielded ratios of 0.73 (95% CI 0.43–1.27, $P = 0.27$) and 0.69 (95% CI 0.39–1.21, $P = 0.20$) on the shifted time scale defined in Methods. Review share was lower with reasoning display (−8.8 percentage points, 95% CI −16.2 to −1.4), with a search estimate of −6.4 points (95% CI −14.2 to 1.4); compose-share estimates were −7.6 and −8.7 points (both $P < 0.05$). The clearest temporal difference was in machine activity; estimates for post-answer student-event duration did not establish a corresponding difference between configurations.

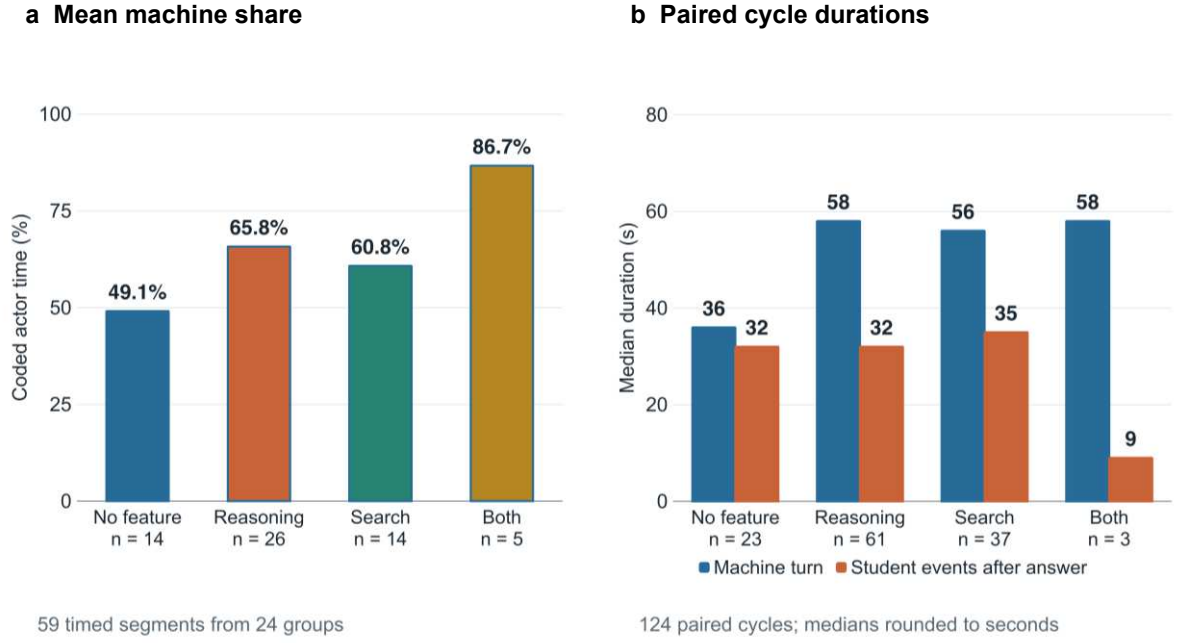

Fig. 3 | Machine and student time across configurations. a, Mean machine share of coded actor time in 59 fully timed segments from 24 groups. Configuration sample sizes are 14, 26, 14 and 5. b, Median machine duration and summed post-answer student-event duration in 124 paired cycles bounded by successive send events within the same segment (23, 61, 37 and 3 cycles). Medians are rounded to whole seconds. This paired subset is drawn from the 181 timed cycles used for the main machine-duration analysis. Model estimates appear in Table 2.

Table 2 | Associations of reasoning display and web search with interaction measures

Outcome	Model	Units (groups)	Reasoning display	Web search
Interaction time (primary outcomes)				
Machine share of coded actor time, percentage points	LMM	59 segments (24)	+18.7 (8.6, 28.8)***	+19.2 (8.5, 29.9)***
within groups	Group FE	48 segments (14)	+19.5 (4.1, 35.0)*	+22.2 (6.4, 37.9)**
DeepSeek segments only	LMM	34 segments (19)	+24.8 (11.8, 37.8)***	+16.5 (3.4, 29.7)*
cluster bootstrap interval	OLS	59 segments (24)	(8.5, 27.1)	(3.4, 27.4)
Machine turn length, ratio	LMM, log	181 cycles (24)	1.65 (1.25, 2.17)***	1.65 (1.25, 2.19)***
within groups	Group FE, log	181 cycles (24)	1.73 (1.09, 2.74)*	1.69 (1.24, 2.31)**
Recorded student activity				
Review share of interaction time, percentage points	LMM	59 segments (24)	−8.8 (−16.2, −1.4)*	−6.4 (−14.2, 1.4)
Compose share of interaction time, percentage points	LMM	59 segments (24)	−7.6 (−15.0, −0.1)*	−8.7 (−16.3, −1.0)*
Post-answer student time, ratio of (1 + seconds)	LMM, log	124 cycles (22)	0.73 (0.43, 1.27)	0.69 (0.39, 1.21)
Review as first action after answer, OR	GEE	194 cycles (29)	0.69 (0.27, 1.76)	0.88 (0.30, 2.59)
Next prompt as first action after answer, OR	GEE	194 cycles (29)	1.26 (0.56, 2.81)	1.84 (0.41, 8.32)
Prompt content				
Elaboration (Q7), OR	GEE	275 prompts (29)	0.91 (0.48, 1.73)	1.05 (0.49, 2.25)
Evaluative prompt (Q8, Q10 or Q12), OR	GEE	275 prompts (29)	0.93 (0.30, 2.88)	1.10 (0.30, 4.04)
Information seeking (Q3), OR	GEE	275 prompts (29)	1.20 (0.48, 2.98)	3.38 (1.17, 9.77)*
Course-material grounding (Q13), OR	GEE	275 prompts (29)	0.97 (0.42, 2.23)	0.44 (0.25, 0.77)**

Estimates and 95% confidence intervals. Models include both features and task; prompt models also include turn position. LMM, linear mixed model with a group random intercept; Group FE, ordinary least squares with group fixed effects and cluster-robust standard errors; GEE, generalised estimating equations with a logit link and exchangeable within-group correlation; OR, odds ratio. Machine-duration models use log(maximum of duration in seconds and 1). Student-duration models use log(1 + seconds), so exponentiated coefficients describe ratios on this shifted scale. The no-association value is 1 for ratios and ORs and 0 for percentage-point differences. Bootstrap intervals use 2,000 group resamples. *$P < 0.05$; **$P < 0.01$; ***$P < 0.001$.

Objects of evaluation and verification

Review was the first observed student action after a machine response in 69.8% of eligible cycles without features, 55.1% with reasoning, 57.7% with search and 70.0% with both (43, 89, 52 and 10 cycles). GEE estimates were imprecise (reasoning OR = 0.69, 95% CI 0.27–1.76; search OR = 0.88, 95% CI 0.30–2.59), and the segment-label permutation test for any feature versus none yielded $P = 0.32$. Event-level analysis of the coded state immediately following answer generation gave Answer-to-Review probabilities of 0.68, 0.52, 0.43 and 0.33 (Fig. 4). Answer-to-Compose transitions met the descriptive residual threshold with reasoning (0.23, $z = 2.18$) and search (0.30, $z = 2.73$) but not without features (0.16). Reasoning-only sequences also included Answer-to-Material transitions (0.11) and Answer-to-Process transitions ($z = 4.07$).

The coded records contained no reopening of a reasoning trace after an answer, no inspection of a cited source and no independent student web search (Table 3). Reasoning display occurred in 150 turns and web retrieval in 86 turns, and citation display accounted for 12 events. Regeneration was absent from the recorded events, although version markers in five recordings, four of them retrospective, indicated activity outside the captured sequence. In search-only sequences, Cite-to-Review had a conditional probability of 0.71 ($z = 5.10$): students continued reading answer text after a citation appeared rather than opening it.

Of 25 correction or redirection prompts, 13 concerned fit to learners or setting, seven re-anchored the answer to the task or course materials and five addressed scope or format (Table 3). G14 asked the system to adapt a curriculum to vocational students, extend an industry analysis beyond Hubei and revise an age range for adult education. G29 progressively specified a

training plan for newly arrived teachers from outside Tibet. G23 requested revision using "the useful theoretical frameworks in the attached readings". None of the 25 corrections disputed a factual claim or a source. Seven verification prompts expressed three distinct requests: "list the real web addresses for reference" (G06), "the materials must not be fabricated" (G19, five recorded instances) and "the resources must be authentic" (G25). All were addressed to the machine, so students articulated standards of accuracy while directing their own corrective judgments towards pedagogical suitability.

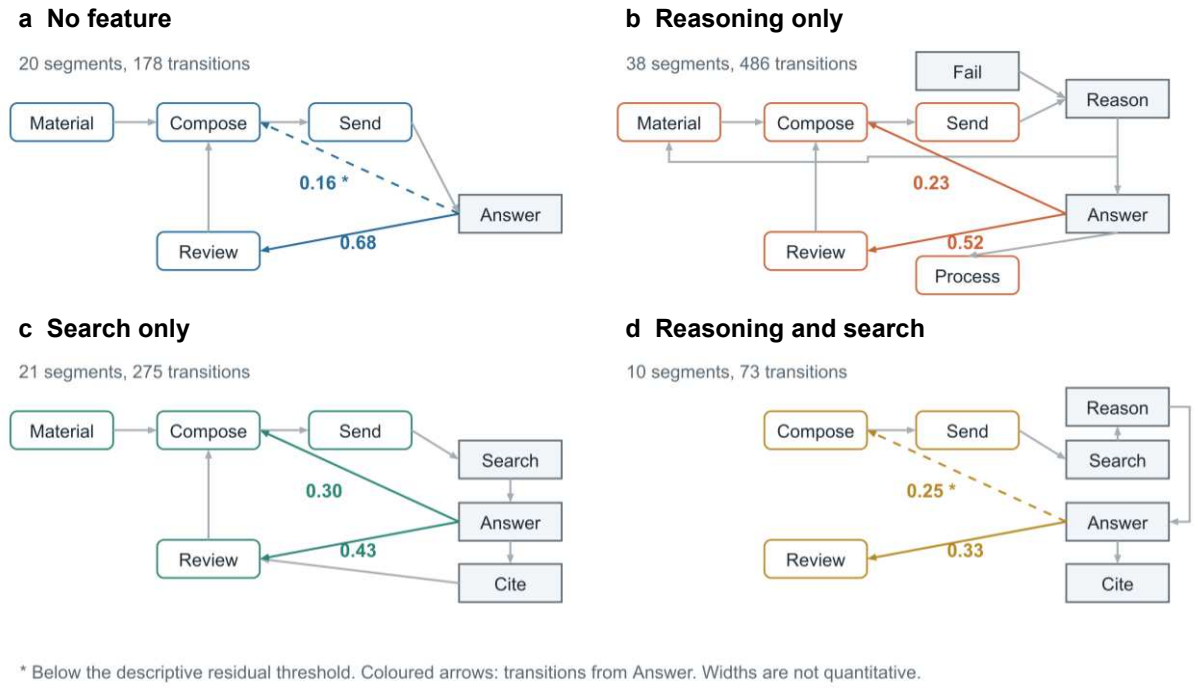

Fig. 4 | Selected lag-1 transitions. Diagrams show selected transitions from 89 segments with usable event order. Grey arrows mark transitions meeting the descriptive criterion of adjusted residual $z > 1.96$ and at least three occurrences. Coloured arrows show Answer-to-Review and Answer-to-Compose probabilities; dashed lines mark transitions below the criterion. All arrows have equal width. Rounded nodes denote student actions and square nodes machine states. Material combines file upload and consultation; Process combines copying, document editing, export and related output handling.

Table 3 | Recorded verification actions and evaluative prompts

Measure	n	Recorded check or note
Available material and recorded checking		
Reasoning-display turns	150	Reasoning traces reopened after answer: 0
Web-retrieval turns	86	Cited sources inspected: 0
Citation-display events	12	Cited sources inspected: 0
All prompts	275	Independent web searches: 0

Measure	n	Recorded check or note
Targets of correction or redirection prompts (Q8; 25 prompts)		
Learners or setting	13	
Task or materials	7	
Scope or format	5	
Verification and comparison prompts		
Verification requests (Q12)	7	Three distinct request intents
Comparison or evaluation requests (Q10)	2	

Counts retain their respective units (turns, events or prompts). The 25 correction prompts were coded Q8; verification requests were coded Q12; and comparison or evaluation requests were coded Q10. Prompt codes can co-occur.

Coordination across platforms and configurations

Twenty-one of 29 groups (72%) used at least two configurations and 19 (66%) used at least two platform families. Fourteen groups (48%) repeated prompt prefixes across configurations, based on matches within the first 40 characters. Groups also transferred outputs between systems: G12 exported plans from three configurations as PDFs and uploaded them to another agent for integration, and G29 extracted keywords from readings with one agent and used them to prompt another. Sequences prepared in notes or shared documents accounted for 33 prompts (12.0%). Two prompts explicitly requested comparison or evaluation, and the criteria for selecting final outputs were not visible in the recordings.

Discussion

RQ1 Teacher relay and student steering

Teacher requirements entered dialogue through copied task directives in Task 1 and course-material grounding in Task 2, while students increasingly steered successive requests.

The teacher's design persisted through this arrangement without governing it. Directive relay held steady in Task 1 while elaboration requests rose, so students took over the conduct of the task while continuing to reproduce its requirements: the didactic contract[25] was not dissolved but redistributed, with its productive clauses passing to the students and the machine while responsibility for factual evaluation remained implicit in the observed dialogue. Gradual release of responsibility[26] consequently needs to separate control over how a task is pursued from

responsibility for judging what it yields. Dividing a reform plan into successive requests evidences the former; explaining and defending the resulting plan would evidence the latter.

The decline in explicit references to course readings raises a question about their function across turns. Material grounding fell from 48.9% to 14.3% of prompts across turns, and a reading made available to GenAI may remain operative in the conversation even when no longer cited. The socio-didactical account of artefacts[12] locates the relevant change: the same document can supply content for generation or criteria for judging that content, and prompt-level mentions alone do not distinguish these uses. That distinction, not the frequency of citation, is what would need to be traced from a passage through a generated answer to a group's final decision.

RQ2 Allocation of interaction time

Reasoning display and web search each increased machine-turn duration and the machine share of coded interaction time, without a clearly established change in post-answer student duration. A long machine response creates an interval in which students may read, discuss or wait, and the larger machine share coincided with longer machine turns; estimates for post-answer student duration did not establish a clear feature-related difference. ICAP[39] and cognitive-offloading theory[40] both imply that elapsed machine time carries no educational meaning on its own, so the interval is an opportunity rather than a loss. A group might state, before the response completes, what would make a proposed strategy appropriate for its learners, and afterwards test one recommendation against that criterion, which would connect the next prompt to an articulated judgment[41,42].

RQ3 Evaluation and verification

Students mainly evaluated pedagogical fit, task alignment and format. Recorded requests for factual assurance went to the machine; the recordings showed no student inspection of cited sources or reopening of reasoning traces after an answer.

Directing GenAI and establishing grounds for accepting its output came apart. Students made consequential decisions about pedagogical fit, sustained the teacher's agenda across turns and

coordinated work across systems, yet addressed their requests for factual assurance to the machine and showed no recorded opening of a cited source or reopening of a reasoning trace after an answer. Extensive interaction with a system can therefore coexist with the absence of visible verification, because the two concern different functions rather than different amounts of effort.

This specifies what accounts of over-reliance and metacognitive laziness[7,8] describe in aggregate. Students revised scope and suitability and rejected recommendations that did not fit their learners. Their evaluative effort concentrated on criteria they could apply from knowledge of the learning context, while assurance about factual content remained delegated. The asymmetry identifies a more precise instructional target than reducing reliance: connecting students' existing capacity to reject an unsuitable recommendation to the evidence needed to justify accepting a suitable one.

The same asymmetry refines hybrid-intelligence accounts of shared control[19–21]. Choosing a tool or directing a prompt gives students influence over the activity, but influence does not determine who establishes credibility. A group can control the request and evaluate contextual fit while delegating both information production and factual assurance, an arrangement we call asymmetric complementarity. Control is therefore specified not by its extent but by the responsibility it governs, which turns an abstract question about shared control into an observable question about the judgments attached to particular claims.

Instrumental genesis offers a reading of this arrangement. Students used recurring strategies to turn GenAI into an instrument of production: decomposing a specification into requests, assigning the system a role, re-anchoring an answer to attached readings and routing material between platforms. A corresponding routine for inspecting reasoning or sources was absent from the coded records, despite reasoning display in 150 turns and web retrieval in 86. Instrument use

concentrated on producing and adapting a task response. In distributed-cognition terms[14,15], production was allocated across students and artefacts, whereas factual validation had no visible counterpart in students' recorded actions. Salomon and colleagues' distinction between effects with and effects of a technology[13] therefore remains relevant: active use alone does not establish what students can subsequently explain or justify themselves.

Verification by instruction and verification by inspection proved to be different acts. Asking that materials "must not be fabricated" states a standard and leaves the machine to satisfy it; opening a citation applies the standard to evidence. Epistemic-cognition theory distinguishes aims for knowing from the justifications and processes used to pursue them[47], and the gap observed here falls exactly between them. Hardwig's account of epistemic dependence[46] clarifies what is at issue: students need not establish every fact themselves, and a group may reasonably take a factual premise from an external source while contributing its own knowledge of a learner population. The recordings show requests for factual assurance without a corresponding recorded inspection of evidence; asking a system to guarantee its own output does not independently establish a claim's credibility. Trust-calibration research[48,49] frames the corresponding requirement as reliance matched to capability rather than reliance reduced.

RQ4 Coordination across platforms and configurations

Groups coordinated platforms and feature configurations by repeating prompts, transferring outputs and asking one system to integrate another's work; the evidential basis of transferred claims was not visibly checked.

Cross-platform coordination adds a further judgment. Groups routed material between systems and asked one to integrate another's output, performing the coordination functions discussed in multi-agent research[2,3]. Agreement between two outputs supports comparison but does not establish that their factual support is independent, and transfer makes provenance harder to recover. Retaining, for a selected recommendation, the originating platform, the supporting source or course passage and the reason for preferring that version would make selection and

integration decisions available for feedback, connecting AI-literacy work on learner agency and opaque systems[52,53] to a classroom practice. The corresponding opportunity for system design is to preserve the link between a claim and the student's judgment about it, and to make that record available to the teacher, extending the logic of classroom-awareness tools[24,54] from activity to the evaluation of generated content.

Implications

Both features were associated with longer machine turns and a larger machine share of coded interaction, yet visible checking was absent. This pattern is compatible with evidence that the effort required for verification influences whether available material is examined[31,32]. In a task oriented towards producing a plan, asking the system to improve an answer integrates into the workflow more readily than investigating its support. Research integrating peer assessment with GenAI-supported instructional design illustrates how evaluation can be built into a production task[50]. Hu, Zhu, Pei and Gu also found that LLM-based teaching simulation and reflection improved teaching plans[51]. Together, these approaches suggest ways to make revision and evaluation explicit in GenAI-supported design. One extension is to ask each group to check one claim that matters for its recommendation before using it, opening the cited source where search is enabled, or examining an assumption about target learners against course material where reasoning is displayed. Because generated reasoning and citations can themselves be incomplete or misleading[33,34], lateral reading[35] rather than the system's assurance supplies the standard. Prioritising a recommendation's central premise keeps the requirement proportionate, and an assessment rubric can credit the quality of the evidential decision, including a decision to retain the original answer for a well-supported reason.

Two propositions follow. First, additional material should be more likely to prompt evaluation when the task assigns students responsibility for a decision that depends on that material, which ties resource availability to epistemic responsibility rather than to visibility. Second, students

can increase their control over a dialogue while continuing to reproduce a teacher-defined agenda, which makes instructional direction and conversational control distinct dimensions rather than ends of one scale.

Limitations and Future Work

The evidence comes from graduate education students in one course, with feature use self-selected and platform, task sequence and group intentions uncontrolled, so differences between configurations may partly reflect the purposes for which students chose them; the sample with both features enabled was small. Screen recordings captured visible actions, leaving audio, gaze, learning outcomes and product quality outside the analysis, and teacher presence was identifiable only through task requirements and course materials rather than through planning records or interviews. Some keyframe-based event boundaries had a resolution of approximately 5 s, and reliability was assessed as human–LLM agreement on pre-segmented events. Zheng and colleagues' Cognitive Echo framework uses LLM-based simulated students to enrich think-aloud protocols[55]. Such simulations offer a complementary methodological resource; participants' own verbal accounts would be needed to investigate the judgments behind the screen actions observed here. A study comparing ordinary feature selection with a structured verification requirement, measuring the quality of the final plan alongside subsequent unaided explanation and recording group discussion, would test whether making evaluative responsibility explicit changes how students use the material that reasoning and search supply.

Methods

Context and participants

The study took place from January 2025 to May 2025 in Adult Teaching Strategies, a graduate education course at a university in Macao SAR, China. Participants were first-year master's students in three class sections. Students formed their own groups of three to five and shared one

computer, tablet or phone during the recorded activity (Fig. 5). The analysis followed each group's activity on its shared screen, treating the group as the unit of repeated observation.

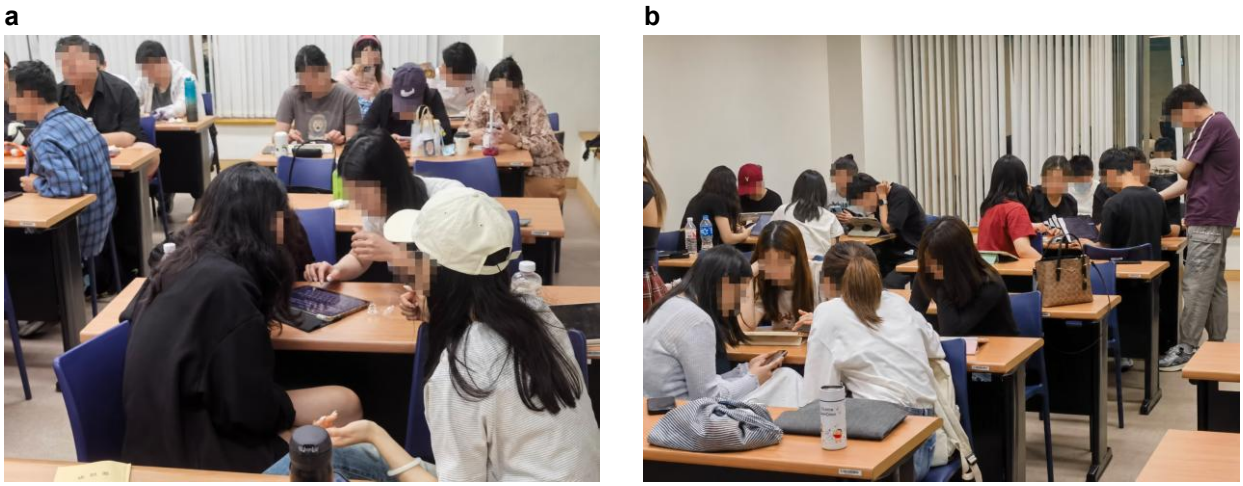

Fig. 5 | Classroom setting for group work. a, Students collaborating around a shared tablet. b, Several groups working concurrently in the classroom. These photographs illustrate the physical setting; the analysis followed activity on each group's shared screen.

Tasks and GenAI configurations

Task 1 presented three cases from Chinese vocational education and required a reform plan covering goals, the current situation, implementation, expected effects, and challenges with countermeasures. Implementation addressed curriculum, teaching methods, assessment and teacher training. Task 2 required a personalised learning strategy or pathway for a chosen adult learner group, including objectives, resources and methods, using three course readings and principles discussed in class. DeepSeek was the primary platform for all groups, which could use other GenAI platforms and submitted screen recordings of their activity.

Students chose whether to enable reasoning display and web search and could change these settings during a task, to minimise researcher control over how groups adapted their use of GenAI to the assignment. Configuration was coded from the recorded interface: reasoning display was enabled when its toggle was active or a visible thinking phase preceded the answer, and search was enabled when its toggle was active or the platform reported retrieval. Each configuration change began a new segment, and segments were classified as neither feature,

reasoning only, search only or both (Fig. 1). The same group could contribute segments under several configurations.

Data processing

Groups submitted 91 video files; removing byte-identical duplicates and repeated submissions of the same recording left 81 recordings. Nineteen groups provided recordings for one task and ten for both, as attendance varied across task sessions. Frames were sampled at one-second intervals, with keyframes retained at screen changes. Coding used keyframe mosaics, enlarged regions and optical character recognition, and student-exported documents supplemented answer text. Ten retrospective recordings showed completed conversations being scrolled and contributed prompts but no event sequences. Twenty-one edited, accelerated or late-starting recordings contributed usable event order but no durations. Fifty complete recordings contributed timing data.

Coding

Event sampling recorded observable actions and their start and end times, with concurrent machine events such as generation and citation display occupying separate rows. The coding scheme contained 40 behaviour codes in eight classes (Table 4). Prompts received any applicable content codes from a set of 14, plus a source category: teacher-derived, group-scripted, student-authored during the dialogue, re-sent, machine-suggested or unidentifiable. Source classification used the recordings and available prompt files.

Prompt source identifies where the wording or request originated and prompt content identifies what the group asked the system to do, so a teacher-derived prompt could also request elaboration or material grounding and content labels were allowed to overlap. Behaviour codes describe visible actions around those prompts, connecting requests with the ensuing sequence of machine and student activity.

Table 4 | Behaviour and prompt coding scheme

Class	Codes	Behaviours coded
Behaviour codes (event sampling, 40 codes in eight classes)		
Prompt construction	PR1–PR7	Typing, pasting, editing before sending, sending, editing a sent prompt, dictating, clicking a suggested follow-up
GenAI response	AI1–AI6	Reasoning phase, web retrieval, answer generation, citation display, failure or interruption, file or image output
Reading and monitoring	RD1–RD5	Reading the reasoning trace after the answer, expanding or collapsing it, reading the answer, opening a cited source, scrolling back to earlier turns
Result processing	PC1–PC7	Copying, pasting into a document, editing a document, regenerating, stopping generation, exporting or screenshotting, producing a final product
External resources	EX1–EX3	Independent web search, consulting local or course files, switching to another GenAI platform
Platform operation	OP1, OP2, OP8, OP9	Opening the platform, starting a new conversation, uploading files or images, browsing conversation history
Feature regulation	OP3–OP7, OP10	Switching reasoning display or web search on or off, changing model, a blocked attempt to enable a feature
Other	OT1–OT2	Off-task or idle; unreadable screen
Prompt content codes (multiple codes per prompt)		
Relay	Q1, Q13	Teacher's task statement (Q1); answer to be based on uploaded or cited course materials (Q13)
Framing	Q2, Q6	Role or context setting (Q2); format or structure constraint (Q6)
Requests	Q3–Q5, Q9	Information seeking (Q3); analysis (Q4); plan generation (Q5); example request (Q9)
Steering	Q7, Q8, Q11	Elaboration of part of the previous answer (Q7); correction or redirection (Q8); integration of earlier turns (Q11)
Evaluation	Q10, Q12	Comparison or evaluation request (Q10); source or accuracy verification request (Q12)
Social	Q14	Thanks, confirmation
Prompt source		
Source	6 categories	Teacher-derived; group-scripted; student-authored during the dialogue; re-sent; machine-suggested; unidentifiable

The coding manual defines each behaviour and its event boundaries. Codes RD1, RD2, RD4, EX1, PC4 and PC5 were retained in the scheme with zero observed frequency.

Directive relay was measured using a teacher-anchor indicator combining task-statement content (Q1) with teacher-derived prompt source, and Q1 and source were also analysed separately. Epistemic relay was coded as course-material grounding (Q13). The first author classified the 25 correction prompts (Q8) by target: learner or contextual fit, task or material alignment, and scope or format. Verification prompts (Q12) were grouped by distinct request intent. Repeated prompting across configurations was assessed by matching the first 40 characters within each group.

Reliability and use of a large language model

A large language model served as an independent second coder, classifying 14 randomly selected recordings containing 231 events. It received the coding manual, event boundaries and keyframes, without the first coder's labels or notes, and produced only behaviour-code classifications. Human–LLM agreement was $\kappa = 0.81$ for the 40 codes and $\kappa = 0.85$ for the eight classes, interpreted using the categories of Landis and Koch[56]. Four recordings initially segmented from a template were re-segmented frame by frame before this assessment, for which pre-correction agreement was $\kappa = 0.40$. Agreement was assessed for classification given the event boundaries.

Measures

For each eligible timed segment, machine time M was the union of machine-event intervals and student time S the union of student-event intervals, with machine share calculated as $M/(M + S)$ after merging overlapping intervals separately for each actor. The denominator represents summed actor time, including concurrent machine and student activity. Review and compose shares used the corresponding coded-event durations divided by $M + S$. Machine-turn duration extended from a send event to the end of the last machine event in the segment-bounded cycle. Cycles without a subsequent coded student action were classified as End and excluded from first-action models, and first actions were grouped as review, next prompt, orchestration or idle. The paired duration plot required a subsequent send within the same segment, and post-answer student time summed the durations of coded student events between the last machine event and that send.

Sequence analysis

Lag sequential analysis[57] used 14 observed analytic states derived from behaviour codes. Consecutive repetitions were collapsed, and transitions were counted within segments and pooled by configuration. Adjusted residuals compared observed counts with those expected

from row and column marginals, and transitions were displayed at $z > 1.96$ with at least three occurrences. This descriptive threshold was unadjusted for multiple comparisons, and cycle-level regression provided comparisons across configurations that accounted for group clustering.

Statistical analysis

Regression models accounted for repeated observations within groups. Segment proportions and transformed durations were analysed with linear mixed models using a group random intercept and restricted maximum likelihood estimation. Binary outcomes used generalised estimating equations[58] with a logit link, exchangeable within-group working correlation and robust standard errors. Feature models included binary reasoning and search indicators and task, and prompt-content models also included unpooled turn position, with plots pooling turn 4 and later. Machine duration was transformed as log(maximum of seconds and 1) and student duration as log(1 + seconds). Machine share and machine duration were designated primary outcomes. Sensitivity analyses used group fixed effects, a DeepSeek-only subset and 2,000 group-bootstrap resamples. An exploratory first-action permutation test reassigned segment configuration labels within task over 1,999 permutations, preserving task strata. Secondary analyses were exploratory. Analyses used Python with pandas, statsmodels and SciPy, with fixed random seeds.

Ethics approval and consent

The study was conducted in accordance with the Declaration of Helsinki. The research protocol was reviewed and approved by the Ethical Committee of the City University of Macau (Reference No. 2025-RE-19) in January 2025, before data collection began. Informed consent was obtained from all participants prior to data collection, covering the study's purpose, the voluntary nature of participation and data-privacy measures. The analysis used desktop screen recordings collected during regular class periods, which posed minimal risk comparable to routine educational activity. Screen recordings were analysed using numeric video identifiers.

Data availability

The original contributions presented in the study are included in the article. Further inquiries can be directed to the corresponding author.

Code availability

Analysis and figure-generation scripts are available from the corresponding author on request.

Acknowledgements

The research was funded by the Macao Science and Technology Development Fund (FDCT) (Grant number 0071/2023/RIB3), the Joint Research Funding Program between the Macau Science and Technology Development Fund (FDCT) and the Department of Science and Technology of Guangdong Province (2024) (FDCT-GDST) (Grant number 0003-2024-AGJ), the Macao Science and Technology Development Fund (FDCT) (Grant number 0065/2024/RIB2), and the Teaching Reform Project of Colleges in Guangdong Province (Grant number 2024-9-802).

Author contributions

Conceptualization, J.X. and Y.D.; methodology, J.X., Y.D. and F.X.; software, H.L.; validation, J.X. and F.X.; writing—original draft preparation, J.X. and Y.D.; writing—review and editing, Y.D., J.X., Y.F. and F.X.; supervision, Y.D.; funding acquisition, Y.D. All authors have read and agreed to the submitted version of the manuscript.

Competing interests

The authors declare no competing interests.

ORCID iDs

Jiaying Xiao https://orcid.org/0009-0009-6985-6636

Yi Dai https://orcid.org/0009-0003-6382-6790

Yuanyuan Feng https://orcid.org/0000-0002-5638-2058

Fu Xu https://orcid.org/0009-0001-7683-3063

Huiyong Li https://orcid.org/0000-0001-9916-7908

References

1. Wang, L. et al. A survey on large language model based autonomous agents. *Front. Comput. Sci.* **18**, 186345 (2024).

2. Guo, T. et al. Large language model based multi-agents: a survey of progress and challenges. In *Proc. 33rd International Joint Conference on Artificial Intelligence (IJCAI-24)* 8048–8057 (IJCAI, 2024).

3. Park, J. S. et al. Generative agents: interactive simulacra of human behavior. In *Proc. 36th Annual ACM Symposium on User Interface Software and Technology* Article 2 (ACM, 2023).

4. Deng, R., Jiang, M., Yu, X., Lu, Y. & Liu, S. Does ChatGPT enhance student learning? A systematic review and meta-analysis of experimental studies. *Comput. Educ.* **227**, 105224 (2025).

5. Bastani, H. et al. Generative AI without guardrails can harm learning: evidence from high school mathematics. *Proc. Natl Acad. Sci. USA* **122**, e2422633122 (2025).

6. Stadler, M., Bannert, M. & Sailer, M. Cognitive ease at a cost: LLMs reduce mental effort but compromise depth in student scientific inquiry. *Comput. Hum. Behav.* **160**, 108386 (2024).

7. Fan, Y. et al. Beware of metacognitive laziness: effects of generative artificial intelligence on learning motivation, processes, and performance. *Br. J. Educ. Technol.* **56**, 489–530 (2025).

8. Zhai, C., Wibowo, S. & Li, L. D. The effects of over-reliance on AI dialogue systems on students' cognitive abilities: a systematic review. *Smart Learn. Environ.* **11**, 28 (2024).

9. Yan, L. et al. Practical and ethical challenges of large language models in education: a systematic scoping review. *Br. J. Educ. Technol.* **55**, 90–112 (2024).

10. Kasneci, E. et al. ChatGPT for good? On opportunities and challenges of large language models for education. *Learn. Individ. Differ.* **103**, 102274 (2023).

11. Zawacki-Richter, O., Marín, V. I., Bond, M. & Gouverneur, F. Systematic review of research on artificial intelligence applications in higher education – where are the educators? *Int. J. Educ. Technol. High. Educ.* **16**, 39 (2019).

12. Rezat, S. & Sträßer, R. From the didactical triangle to the socio-didactical tetrahedron: artifacts as fundamental constituents of the didactical situation. *ZDM* **44**, 641–651 (2012).

13. Salomon, G., Perkins, D. N. & Globerson, T. Partners in cognition: extending human intelligence with intelligent technologies. *Educ. Res.* **20**, 2–9 (1991).

14. Hutchins, E. *Cognition in the Wild* (MIT Press, 1995).

15. Pea, R. D. Practices of distributed intelligence and designs for education. In *Distributed Cognitions: Psychological and Educational Considerations* (ed. Salomon, G.) 47–87 (Cambridge Univ. Press, 1993).

16. Vérillon, P. & Rabardel, P. Cognition and artifacts: a contribution to the study of thought in relation to instrumented activity. *Eur. J. Psychol. Educ.* **10**, 77–101 (1995).

17. Sundar, S. S. Rise of machine agency: a framework for studying the psychology of human–AI interaction (HAII). *J. Comput.-Mediat. Commun.* **25**, 74–88 (2020).

18. Lodge, J. M., Yang, S., Furze, L. & Dawson, P. It's not like a calculator, so what is the relationship between learners and generative artificial intelligence? *Learn. Res. Pract.* **9**, 117–124 (2023).

19. Molenaar, I. Towards hybrid human-AI learning technologies. *Eur. J. Educ.* **57**, 632–645 (2022).

20. Cukurova, M. The interplay of learning, analytics and artificial intelligence in education: a vision for hybrid intelligence. *Br. J. Educ. Technol.* **56**, 469–488 (2025).

21. Järvelä, S., Nguyen, A. & Hadwin, A. Human and artificial intelligence collaboration for socially shared regulation in learning. *Br. J. Educ. Technol.* **54**, 1057–1076 (2023).

22. Dillenbourg, P. Design for classroom orchestration. *Comput. Educ.* **69**, 485–492 (2013).

23. Kali, Y., McKenney, S. & Sagy, O. Teachers as designers of technology enhanced learning. *Instr. Sci.* **43**, 173–179 (2015).

24. Holstein, K., McLaren, B. M. & Aleven, V. Co-designing a real-time classroom orchestration tool to support teacher–AI complementarity. *J. Learn. Anal.* **6**, 27–52 (2019).

25. Brousseau, G. *Theory of Didactical Situations in Mathematics* (Kluwer, 1997).

26. Pearson, P. D. & Gallagher, M. C. The instruction of reading comprehension. *Contemp. Educ. Psychol.* **8**, 317–344 (1983).

27. Wei, J. et al. Chain-of-thought prompting elicits reasoning in large language models. *Adv. Neural Inf. Process. Syst.* **35**, 24824–24837 (2022).

28. Guo, D. et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature* **645**, 633–638 (2025).

29. Lewis, P. et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Adv. Neural Inf. Process. Syst.* **33**, 9459–9474 (2020).

30. Sweller, J., van Merriënboer, J. J. G. & Paas, F. Cognitive architecture and instructional design: 20 years later. *Educ. Psychol. Rev.* **31**, 261–292 (2019).

31. Buçinca, Z., Malaya, M. B. & Gajos, K. Z. To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proc. ACM Hum.-Comput. Interact.* **5**, 188 (2021).

32. Vasconcelos, H. et al. Explanations can reduce overreliance on AI systems during decision-making. *Proc. ACM Hum.-Comput. Interact.* **7**, 129 (2023).

33. Turpin, M., Michael, J., Perez, E. & Bowman, S. R. Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting. *Adv. Neural Inf. Process. Syst.* **36**, 74952–74965 (2023).

34. Liu, N. F., Zhang, T. & Liang, P. Evaluating verifiability in generative search engines. In *Findings of the Association for Computational Linguistics: EMNLP 2023* 7001–7025 (ACL, 2023).

35. Wineburg, S. & McGrew, S. Lateral reading and the nature of expertise: reading less and learning more when evaluating digital information. *Teach. Coll. Rec.* **121**, 1–40 (2019).

36. Sparrow, B., Liu, J. & Wegner, D. M. Google effects on memory: cognitive consequences of having information at our fingertips. *Science* **333**, 776–778 (2011).

37. Fisher, M., Goddu, M. K. & Keil, F. C. Searching for explanations: how the Internet inflates estimates of internal knowledge. *J. Exp. Psychol. Gen.* **144**, 674–687 (2015).

38. Messeri, L. & Crockett, M. J. Artificial intelligence and illusions of understanding in scientific research. *Nature* **627**, 49–58 (2024).

39. Chi, M. T. H. & Wylie, R. The ICAP framework: linking cognitive engagement to active learning outcomes. *Educ. Psychol.* **49**, 219–243 (2014).

40. Risko, E. F. & Gilbert, S. J. Cognitive offloading. *Trends Cogn. Sci.* **20**, 676–688 (2016).

41. Lee, H.-P. et al. The impact of generative AI on critical thinking: self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers. In *Proc. 2025 CHI Conference on Human Factors in Computing Systems* Article 1121 (ACM, 2025).
42. Tankelevitch, L. et al. The metacognitive demands and opportunities of generative AI. In *Proc. 2024 CHI Conference on Human Factors in Computing Systems* Article 680 (ACM, 2024).
43. Kim, J., Lee, S.-S., Detrick, R., Wang, J. & Li, N. Students–generative AI interaction patterns and its impact on academic writing. *J. Comput. High. Educ.* 38, 504–525 (2026).
44. Nguyen, A., Hong, Y., Dang, B. & Huang, X. Human–AI collaboration patterns in AI-assisted academic writing. *Stud. High. Educ.* **49**, 847–864 (2024).
45. Du, J., Hew, K. F. & Liu, L. What can online traces tell us about students' self-regulated learning? A systematic review of online trace data analysis. *Comput. Educ.* **201**, 104828 (2023).
46. Hardwig, J. Epistemic dependence. *J. Philos.* **82**, 335–349 (1985).
47. Chinn, C. A., Buckland, L. A. & Samarapungavan, A. Expanding the dimensions of epistemic cognition: arguments from philosophy and psychology. *Educ. Psychol.* **46**, 141–167 (2011).
48. Lee, J. D. & See, K. A. Trust in automation: designing for appropriate reliance. *Hum. Factors* **46**, 50–80 (2004).
49. Parasuraman, R. & Manzey, D. H. Complacency and bias in human use of automation: an attentional integration. *Hum. Factors* **52**, 381–410 (2010).
50. Liu, C.-C. et al. Facilitating pre-service teachers' instructional design and higher-order thinking with generative AI: an integrated approach with the peer assessment and concept map. *J. Res. Technol. Educ.* **58**, 880–905 (2026).
51. Hu, B., Zhu, J., Pei, Y. & Gu, X. Exploring the potential of LLM to enhance teaching plans through teaching simulation. *npj Sci. Learn.* **10**, 7 (2025).
52. Gašević, D., Siemens, G. & Sadiq, S. Empowering learners for the age of artificial intelligence. *Comput. Educ. Artif. Intell.* **4**, 100130 (2023).
53. Bearman, M. & Ajjawi, R. Learning to work with the black box: pedagogy for a world with artificial intelligence. *Br. J. Educ. Technol.* **54**, 1160–1173 (2023).
54. Cai, H., Han, B., Lu, L., Wong, L. H. & Gu, X. When AI joins the conversation: the impact of AI-powered teacher dashboard on pre-service teachers' collaborative interaction in the video-based professional learning communities. *Interact. Learn. Environ.* **34**, 2547–2559 (2026).
55. Zheng, L., He, A., Qi, C., Zhang, H. & Gu, X. Cognitive Echo: Enhancing think-aloud protocols with LLM-based simulated students. *Br. J. Educ. Technol.* **56**, 2019–2042 (2025).
56. Landis, J. R. & Koch, G. G. The measurement of observer agreement for categorical data. *Biometrics* **33**, 159–174 (1977).
57. Bakeman, R. & Quera, V. *Sequential Analysis and Observational Methods for the Behavioral Sciences* (Cambridge Univ. Press, 2011).
58. Liang, K.-Y. & Zeger, S. L. Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22 (1986).