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.

Medical English writing, Human-machine symbiosis, Instructional experiment, Self-efficacy, Collaboration

Shuang Du
Sichuan International Studies University
Kaiwen Cheng
kevincheng@sisu.edu.cn

Sichuan International Studies University

Research Article

Keywords: Medical English writing, Human-machine symbiosis, Instructional experiment, Self-efficacy, Collaboration

Posted Date: July 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9915798/v1

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

Additional Declarations: No competing interests reported.

Abstract

Background

Academic English writing is an essential competency for medical students to conduct scholarly research and engage in international academic exchanges. Nevertheless, most medical students suffer from deficiencies in linguistic expression and low writing self-confidence. Although AI tools (e.g., ChatGPT) have innovated writing instruction and peer feedback has proven effective in improving Chinese medical students' writing ability, two prominent research gaps remain. Existing systems lack cognitively-aware AI agents for scalable, interactive engagement, and no established framework systematically allocates cognitive responsibilities between human learners and AI agents throughout the academic writing process.

Methods

This study developed a Human–Machine Symbiotic Cognition (HMSC) framework that integrated AI-based feedback with collaborative peer writing to improve medical students' writing performance, self-efficacy, and teamwork satisfaction. A quasi-randomized instructional experiment was conducted with 41 undergraduate medical students from a Chinese university. Over two weeks, the intervention group engaged with the HMSC framework via a web-based platform where four dedicated AI agents (Information Provider, Feedback Coach, Co-creator, and Co-decider) supported each writing stage, while the control group received conventional instruction. Assessments were conducted at baseline (T0), post-intervention (T1), and one-month follow-up (T2), measuring writing performance, self-efficacy, and teamwork satisfaction. Data were analyzed using linear mixed-effects models.

Results

The intervention group showed significant improvements in overall writing scores from T0 to T1, with an average increase 8.4 points ($p < 0.001$), whereas the control group showed no significant change ($p > 0.1$). Significant improvements were also found in grammatical accuracy and content relevance (both $p < 0.01$). For writing self-efficacy, the intervention group achieved prominent gain in Ideation and Self-Regulation subscales (both $p < 0.01$), and reported significantly higher team collaboration satisfaction than the control group ($p < 0.05$). These findings indicate that the HMSC framework substantially enhanced medical students' academic writing performance.

Conclusion

The HMSC framework yielded notable improvements in medical students' English academic writing performance, self-efficacy and teamwork satisfaction in the short term. The findings lend support to

distributed cognition, ICAP-based engagement, and social regulation in a human-machine symbiotic environment. Further research with longer intervention periods is warranted to strengthen learners' long-term writing competence across various medical education contexts.

Introduction

The ability to write clear, well-structured academic English is no longer an ancillary skill in undergraduate medical education — it is a core competency that shapes students' capacity to engage with the international biomedical literature, communicate clinical findings, and participate in global scholarly discourse [1, 2]. For Chinese medical undergraduates, developing this competency presents distinctive challenges. Beyond general English proficiency, these students must master discipline-specific academic vocabulary, adhere to the rhetorical and genre conventions of biomedical writing, and manage the cognitive demands of synthesizing evidence into coherent arguments — all while navigating well-documented affective barriers, particularly writing anxiety [3, 4]. These challenges are compounded by structural constraints in medical English instruction: large class sizes, limited contact hours, and a persistent shortage of instructors with dual expertise in language pedagogy and medical content [5]. The result is a systematic mismatch between the intensive, individualized writing support that medical students require and what conventional classroom-based instruction can feasibly deliver.

Recent empirical work has systematically documented the scope of these difficulties. Ren et al. [1], in a cross-sectional study of Chinese medical students, identified vocabulary limitations, grammatical inaccuracy, and writing anxiety as the most prevalent barriers to L2 academic writing, with a majority of respondents reporting that prior English coursework had left them underprepared. Wen et al. [5] corroborated these findings through a comprehensive review delineating five interconnected categories of challenge — linguistic, cultural-rhetorical, literature-synthesis, psychological, and supervisory — with 65% of students reporting moderate-to-severe anxiety during thesis writing. Intervention studies have demonstrated that structured pedagogical approaches can yield measurable gains: peer feedback protocols produced sustained improvements across multiple IELTS writing dimensions [2], while flipped-classroom designs enhanced SCI writing skills among medical postgraduates [4]. Yet these interventions remain constrained by the same structural factors they seek to address — most critically, the limited capacity of instructors to deliver the timely, granular, and individualized feedback that writing development demands.

Computer-Supported Collaborative Learning (CSCL) offers a theoretically grounded pathway for addressing these constraints through technology-mediated, socially constructed knowledge building [6, 7]. Within this tradition, the ICAP framework [8] predicts learning gains as a function of cognitive engagement depth, with interactive modes — those involving substantive peer dialogue and co-construction — consistently yielding the largest effect sizes. Collaborative writing interventions aligned with ICAP principles have produced significant improvements in medical students' writing coherence, lexical resource, and grammatical accuracy [2, 9]. However, implementing genuinely interactive instruction at scale confronts persistent obstacles: large-enrollment medical English courses lack the

facilitation capacity that interactive modes require [10], and peer feedback quality deteriorates when learners possess weak prior linguistic knowledge — limiting the cognitive gains achievable through peer interaction alone [11]. These well-documented bottlenecks motivate the search for a scalable scaffolding mechanism that can sustain interactive engagement without depending entirely on instructor facilitation or advanced peer competence.

Recent advances in large language models (LLMs) provide a structurally novel response to this challenge. LLM-based agent systems can engage learners through natural language interaction, generate real-time personalized scaffolding, and deliver multimodal feedback grounded in semantic analysis — capabilities that directly target the feedback timeliness, individualization, and regulation deficits that constrain conventional collaborative learning [12]. Within medical education, an accumulating evidence base supports the efficacy of AI-assisted writing instruction: studies document significant improvements in writing quality, vocabulary, and organization following ChatGPT-mediated revision [13, 14], sustained gains on EAP writing assessments among EFL medical undergraduates using generative AI tools [15], and enhanced outcomes when GenAI chatbots are integrated with peer feedback in hybrid instructional models [16]. Cross-sectional evidence further indicates that medical students hold broadly positive attitudes toward AI-augmented learning [17, 18], establishing a receptive user base for well-designed AI-mediated interventions.

However, these studies conceptualize AI primarily as an external writing tool or feedback provider — a resource that students consult during revision — rather than as a structured cognitive partner embedded within a theoretically grounded pedagogical framework [19, 20]. Two gaps remain unaddressed. First, existing CSCL-mediated writing interventions, while grounded in ICAP principles, have not incorporated cognitively-aware AI agents capable of sustaining interactive engagement at scale without intensive instructor facilitation — the systematic consideration of learner cognition (e.g., cognitive load regulation, metacognitive scaffolding) has been largely absent from AI-writing integration efforts. Second, no existing framework, to our knowledge, systematically distributes cognitive responsibilities between human learners and AI agents across the complete academic writing cycle — from domain-specific information gathering and comprehension, through evidence-based argument construction, to multi-criteria revision decisions. These two gaps converge on a single unmet need in medical education: a pedagogical framework that embeds AI agents as structured cognitive partners with principled decision-authority allocation, complementing rather than substituting human capabilities in contextualized understanding, value judgment, and creative expression [21, 22].

To address these two gaps, we propose the Human–Machine Symbiotic Cognition (HMSC) framework. By operationalizing academic writing as an observable, auditable interaction protocol with stage-specific artifacts, HMSC reconceptualizes AI from a passive writing aid into a structured cognitive partner — complementing rather than supplanting human capabilities in contextualized understanding, value judgment, and creative expression [12, 19]. To empirically evaluate the proposed framework, we developed an HMSC-based collaborative writing platform and conducted a two-week quasi-randomized

experiment with 41 undergraduate medical students at a university in Chongqing, China, guided by the following research questions:

RQ1: Compared with the conventional ICAP-based instructional model, does the HMSC framework significantly enhance medical students' English academic writing performance in terms of language accuracy, content relevance, and textual coherence?

RQ2: Does the HMSC framework effectively improve medical students' writing self-efficacy, specifically in the dimensions of ideation, language conventions, and self-regulation?

RQ3: In the agent-mediated collaborative writing environment, do medical students demonstrate significantly higher perceived teamwork quality—including task coordination, idea integration, and conflict resolution—compared with those in conventional peer collaboration?

Methodology

HMSC Framework

Theoretical Foundations of the HMSC Framework

The HMSC framework is grounded in three complementary theoretical traditions that collectively motivate the distribution of cognitive labor across human and artificial agents. First, distributed cognition theory posits that cognitive processes extend beyond individual minds to encompass artifacts, tools, and collaborators within a functionally integrated system [21]. HMSC operationalizes this principle by treating LLM-based agents not as external writing aids but as components of a distributed cognitive architecture, in which information retrieval, pattern recognition, and text generation are allocated to machines, while contextualized interpretation, value judgment, and creative synthesis remain under human purview. Second, self-regulated learning (SRL) theory characterizes learning as a cyclical process of forethought, performance monitoring, and reflection [23]; the four HMSC layers are explicitly aligned with these SRL phases, embedding metacognitive checkpoints at each transition to promote learners' awareness of their own writing processes. Third, the ICAP framework establishes that learning outcomes are a function of cognitive engagement depth, with interactive modes consistently yielding the largest effect sizes [8]. HMSC extends ICAP by introducing an artificial agent as a structured interaction partner capable of sustaining interactive engagement even when human peer interaction is constrained by proficiency asymmetries, large class sizes, or limited instructor availability [10, 11]. These three traditions converge on a design premise that defines the framework's core contribution: an optimal cognitive architecture for academic writing instruction distributes processing across human and machine agents in a principled, traceable, and pedagogically intentional manner, rather than treating AI as an undifferentiated productivity tool.

The four layers of the HMSC framework correspond to distinct phases of the academic writing process and are operationalized through dedicated AI agent roles (Fig. 1). The Perception layer addresses the

information-gathering challenges that writers face when approaching an unfamiliar academic topic. Table 1 details the constitutive functions at each layer.

Table 1
Hierarchical Structure of the HMSC Framework

Level	Human (H)	Machine (M)	Human–Machine Symbiosis	Typical Activities / Outputs
Perception (Passive)	Engages in selective attention and initial encoding; identifies topic, genre, and task constraints.	Generates a briefing pack tailored to the set topic and proficiency: 120–180-word synopsis; terminology cards; key facts with cited sources; a mini reading plus three comprehension questions. Ensures provenance and basic de-noising.	Predominantly one-way transmission with emphasis on input quality and boundary control.	Materials list; terminology cards; key-fact table with sources; mini reading and comprehension questions.
Comprehension (Active)	Poses questions around concepts and rules; paraphrases, contrasts, and explains cause–effect; identifies points of difficulty.	Acts as a Socratic coach: provides concise answers plus one follow-up probe; detects and corrects misconceptions; offers rule-of-thumb guidance and brief exemplars.	Bidirectional Q&A; triadic alignment of Question–Explanation–Evidence; corrections are made explicit.	Q&A transcript; misconception log; concept-contrast table; "I-understand" checklist.
Prediction (Constructive)	Proposes 2–3 competing claims or writing plans; organizes preliminary arguments and an outline.	Serves as co-creator: builds an Evidence Matrix; runs scenario/corpus simulations; identifies boundary conditions and constraints; provides revision suggestions.	Co-creation coupling; at least one counterexample or boundary condition per claim; versioned evolution is documented.	Evidence Matrix; A/B/C plan cards; revision notes; version-difference log.
Decision (Interactive)	Sets evaluation weights (Accuracy, Coherence, Feasibility, Fairness); makes the final choice and assumes responsibility and ethical accountability.	Provides multi-criteria scoring, sensitivity analysis, risk and bias alerts; outputs a justificatory rationale chain and an action checklist.	Joint adjudication; upon "Lock & Submit," the agent is frozen, producing an auditable final version.	Weighted score sheet; recommendation with rationale chain; risk note; action checklist; lock record.

From the perspective of human–machine interaction, the four-tiered HMSC framework defines role specifications, task units, and decision-right allocations at each hierarchical layer across three distinct dimensions: Human (H), Machine-based Agent (M), and Human–Machine Interaction (H–M). This triadic specification serves two core pedagogical functions. First, it operationalizes a participation-upgrade trajectory aligned with the ICAP taxonomy: learners progress from passive information reception (Perception) through constructive dialogue (Comprehension) and interactive co-creation (Prediction) to agentic decision endorsement (Decision), with the agent providing calibrated support at each stage. Second, it makes explicit the distribution of decision authority, which is configured in three modes depending on task characteristics: (a) machine-led with human review, applied to information retrieval and integration tasks where AI efficiency advantages are maximal and the risk of error propagation is mitigated by human gatekeeping [20, 24]; (b) human-led with machine support, applied to value-laden judgments and ethical trade-offs where human endorsement is non-negotiable and AI inputs serve to broaden the solution space [25]; and (c) joint adjudication, applied to option appraisal and plan revision where the integration of data-driven analytics and domain expertise yields more robust decisions than either party could produce alone [26, 27]. This graduated, task-sensitive allocation of decision authority is a defining feature of the HMSC framework: it differentiates symbiotic cognition from both fully automated writing tools and purely human collaborative models, occupying a theoretically motivated intermediate position that preserves human agency while exploiting AI capabilities.

The Setup of HMSC-Based Writing Collaborative Platform

To operationalize the HMSC framework, we developed a web-based collaborative writing platform that implements the four-stage human-AI workflow described above. The platform presents learners with a structured interface in which each stage—topic comprehension (Perception), problem clarification (Comprehension), argument construction (Prediction), and option selection (Decision)—is visually demarcated and associated with a specific AI agent role, with all interactions logged to produce an auditable process trail. Figure 2 illustrates the platform interface; Appendix A provides a complete interaction flow for a representative topic, and Appendix B documents the full set of GPT-4o-assisted prompts used at each stage.

Technical Implementation and Data Privacy Measures

To support effective interaction with AI agents, the study employed OpenAI's GPT-4o model (gpt-4o-2024-08-06) accessed via the official API. Role configuration was implemented through structured system prompts specifying each agent's pedagogical function, input-output format, interaction protocol, and behavioral constraints (e.g., factual accuracy requirements, citation obligations, and prohibitions against generating complete essays on behalf of students). Prompt templates were iteratively refined through pilot testing with a separate cohort of 15 medical students prior to the main experiment, during which agent response quality was evaluated on dimensions of accuracy, pedagogical appropriateness, language level alignment, and cultural sensitivity. All API calls were made with temperature set to 0.3 to

balance response coherence with instructional variety, and a maximum output token limit of 2,048 was imposed to maintain manageable response lengths. All sensitive data—including user profiles, interaction logs, and submitted writing drafts—were stored exclusively in local browser databases (IndexedDB) on institution-managed devices, with no data transmitted to external servers beyond the API calls to OpenAI. API communication was conducted over encrypted HTTPS connections in strict compliance with OpenAI's API data usage policies, which stipulate that data submitted through the API is not used for model training. Student participants were assigned anonymized identifiers, and all interaction logs were stripped of personally identifiable information before analysis. The study protocol, including data handling procedures, was reviewed and approved by the institutional ethics committee (Approval No. 20250021).

Participants of the Instructional Experiment

We recruited 41 undergraduate medical students from a university in Chongqing, China (no major writing disabilities). All participants provided informed written consent, and procedures conformed to the Declaration of Helsinki. The students were randomly assigned to either the experimental group (n = 21; M_age = 22.45, SD = 0.23) or the control group (n = 20; 16 female; M_age = 22.72, SD = 0.81) using a random number table. Baseline characteristics (age, gender) showed no significant differences between groups. Only those who completed the course and all three assessments were included in the final analysis. A priori power analysis was conducted using G*Power [28], with parameters set to f = 0.17, α = 0.05, power = 0.95, and inter-measurement correlation r = 0.80. The minimum required sample size was determined to be 38 participants.

Experimental Procedure

As illustrated in Fig. 3, the experimental phase lasted two weeks. Both groups participated in four instructional sessions delivered by the same instructor, covering identical writing strategies. The only distinction was the instructional model: the experimental group followed the HMSC-based instructional strategy, while the control group used conventional ICAP-based collaborative writing instructional model (identical writing strategy content to the HMSC group): each session began with the instructor delivering explicit instruction on writing strategies and language points, followed by individual writing exercises completed independently by students. This model corresponds to the Passive-to-Active range of the ICAP framework, in which learners receive information and apply it individually without structured peer interaction or technology-mediated feedback. Classes were conducted in multimedia computer labs. The HMSC group accessed a GPT-4–enabled web interface for real-time interaction with AI agents, whereas the control group received traditional instruction consisting of lectures and quizzes. Assessments were administered at three time points: baseline (T0), immediate post-test (T1), and one-month follow-up (T2). All assessments were completed online and scored by an independent team blinded to the group allocation and intervention details.

Instruments

Writing Performance Assessment

The writing tasks were aligned with participants' English proficiency level (approximately CEFR B1-B2) and the course learning objectives, which emphasized the ability to produce coherent short academic texts with appropriate vocabulary, clear organization, and adherence to basic academic genre conventions. Each task required students to compose a short academic essay of at least 70 words on a designated health-science topic (e.g., 'The role of technology in modern healthcare,' 'Strategies for promoting healthy lifestyle behaviors') within 30–40 minutes. This word-count threshold, though modest relative to authentic academic writing, was calibrated to the participants' English proficiency and the 40-minute instructional session constraint, ensuring that all students could produce complete texts amenable to rubric-based evaluation within the available time. Writing performance was assessed using Hyland's (2003) analytic scoring rubric, which was selected for its theoretical grounding in genre-based pedagogy and its demonstrated sensitivity to the linguistic and organizational dimensions most relevant to L2 academic writing. The rubric covers three dimensions: Sentence Structure and Vocabulary (SSV, maximum 40 points), Organization and Coherence (OC, maximum 20 points), and Format and Content (FC, maximum 40 points), yielding a composite score of 100 points. Two experienced associate professors served as primary raters, and an additional two senior professors independently double-rated each essay after completing a standardized calibration training session in which they applied the rubric to a set of 10 training essays and reached consensus on scoring criteria. The average of the two independent ratings was used as the final score for each dimension. Inter-rater reliability, assessed via Spearman's rank-order correlation, was 0.87, indicating strong consistency between raters. Discrepancies exceeding 5 points on the composite score were resolved through discussion and re-rating.

Self-Efficacy for Writing Scale

To measure students' confidence in their writing abilities, the study adopted a culturally adapted Chinese version of the Self-Efficacy for Writing Scale (SEWS), originally developed and validated by Bruning et al. [29] and subsequently refined by Ramos-Villagrasa et al. [30]. The adaptation process followed established cross-cultural instrument translation guidelines: the original English items were independently translated into Chinese by two bilingual researchers, reconciled through expert panel discussion, back-translated into English by a third translator to verify semantic equivalence, and piloted with a sample of 25 medical students not participating in the main experiment. The resulting instrument consists of 16 items distributed across three dimensions: Ideation (5 items, e.g., 'I can generate ideas for writing on a given topic'), Language Conventions (6 items, e.g., 'I can use appropriate academic vocabulary in my writing'), and Self-Regulation (5 items, e.g., 'I can monitor and adjust my writing process to meet task requirements'). Each item is rated on a 5-point Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree). Reliability analyses based on pilot testing yielded Cronbach's α values of 0.91 (Ideation), 0.89 (Language Conventions), and 0.92 (Self-Regulation), and 0.94 for the total scale,

indicating excellent internal consistency. Confirmatory factor analysis conducted on pilot data confirmed the three-factor structure with satisfactory fit indices (CFI = 0.93, RMSEA = 0.06, SRMR = 0.05).

Teamwork Satisfaction Scale

Students' perceived quality of collaboration during the human-agent co-writing process was assessed using a culturally adapted version of the Teamwork Satisfaction Scale (TSS) developed by Ku et al. [31]. The original TSS was designed to evaluate collaborative learning experiences in technology-mediated environments and was adapted for the present study through the same translation and pilot-testing procedure described above for the SEWS. The adapted 10-item instrument employs a 5-point Likert scale (1 = strongly disagree, 5 = strongly agree) and captures five facets of teamwork: Feedback Mechanisms (e.g., 'Team members provided constructive feedback on my contributions'), Knowledge Sharing (e.g., 'The team effectively shared domain knowledge and writing strategies'), Learning Motivation (e.g., 'Collaborating with team members increased my motivation to improve my writing'), Problem-Solving (e.g., 'The team worked together effectively to resolve writing difficulties'), and Perceived Output Quality (e.g., 'The final writing product was better than what I could have produced alone'). The scale demonstrated strong internal reliability in the present sample (Cronbach's α = 0.92). Higher total scores indicated greater satisfaction with the collaborative process. In conjunction with the writing performance and self-efficacy measures, this instrument provided a multidimensional assessment of the interpersonal and motivational dynamics underpinning collaborative writing in an AI-augmented environment.

Statistical Analysis

Descriptive statistics, including means and standard deviations, were computed for all study variables using Mplus. The Shapiro-Wilk test and Q-Q plots were used to evaluate normality, while Levene’s test assessed homogeneity of variance. Independent-samples t-tests were conducted on pre-intervention (T0) scores to assess group equivalence. Non-significant results across key variables confirmed the comparability of the HMSC and ICAP groups prior to the intervention.

Linear Mixed Effects Modeling (LMM) was used to analyze both main and interaction effects for all outcome variables. Group and Time were included as fixed effects, while participant ID was modeled as a random intercept. Restricted Maximum Likelihood (REML) estimation was used to determine variance components, with degrees of freedom adjusted via the Satterthwaite method. For significant interactions, simple effects analyses were conducted to explore group differences over time.

Results

Baseline comparisons using linear mixed effects modeling revealed no significant differences between the HMSC and ICAP groups in writing self-efficacy, teamwork satisfaction, or writing performance (all $p > .05$), confirming initial equivalence and ruling out pre-existing differences as confounding factors. Visual inspection of post-intervention outcomes indicated a general upward trajectory for the HMSC group across timepoints, particularly immediately after the intervention. While the control group also

demonstrated moderate gains, the improvements were notably smaller. Figure 4 illustrates the estimated mean trends over time for both groups, reflecting the overarching pattern.

Note

1. The writing assessment comprises three components: SSV, OC, and FC, referring to Sentence Structures and Vocabulary, Organization and Coherence, and Format and Content, respectively. 2. The Self-Efficacy for Writing Scale (SEWS) consists of three subscales: Ideation, Conventions, and Self-Regulation. 3. TSS stands for the Teamwork Satisfaction Scale.

Outcomes of Writing Performance

Table 2 presents descriptive statistics and time effects for all writing performance measures. The HMSC group showed a significant increase in overall writing scores from T0 to T1 (mean increase 8.4 points, $p < .001$), whereas the control group's change was non-significant ($p > .1$). Among the three sub-dimensions, Sentence Structure and Vocabulary (SSV) exhibited the largest between-group difference at T1, with the HMSC group outperforming controls by 6.36 points ($p < .01$). Organization and Coherence (OC) also showed a significant group × time interaction, with the HMSC group gaining 2.93 points more than controls from T0 to T1 ($p < .05$). However, for Format and Content (FC), the HMSC group's improvement was modest and did not reach statistical significance. At T2, the advantages in SSV and OC had attenuated, with neither reaching statistical significance, suggesting partial fade-out of writing gains one month after the intervention.

Table 2
Descriptive statistics and time effects analysis.

	Measures, Mean (*SD*)			*F*	*p*
	T0	**T1**	**T2**		
SSV					
ICAP	29.76 (3.00)	30.32 (1.81)	30.07 (3.53)	0.134	0.718
HSMC	29.88 (2.88)	36.25 (1.90)	33.49 (2.95)	30.006	**< 0.001**
OC					
ICAP	14.70 (1.41)	14.80 (0.92)	14.90 (2.26)	0.235	0.633
HSMC	14.62 (1.33)	17.55 (0.71)	16.23 (2.20)	13.962	**0.001**
FC					
ICAP	30.32 (2.79)	29.90 (1.89)	28.66 (4.10)	4.65	**0.043**
HSMC	30.46 (2.62)	35.53 (1.35)	32.85 (3.64)	7.885	**0.011**
Ideation					
ICAP	4.01 (0.11)	4.10 (0.10)	3.85 (0.50)	2.228	0.151
HSMC	4.03 (0.08)	5.06 (0.07)	4.65 (0.46)	38.081	**< 0.001**
Conventions					
ICAP	4.02 (0.10)	4.06 (0.07)	3.83 (0.43)	3.903	0.62
HSMC	4.01 (0.06)	5.09 (0.08)	4.83 (0.60)	38.02	**< 0.001**
Self-Regulation					
ICAP	3.80 (0.10)	3.92 (0.09)	3.68 (0.37)	2.271	0.147
HSMC	3.81 (0.07)	4.90 (0.07)	4.69 (0.48)	57.57	**< 0.001**
TSS					
ICAP	3.03 (0.01)	3.07 (0.01)	2.86 (0.52)	2.187	0.155
HSMC	3.04 (0.01)	4.05 (0.01)	3.88 (0.45)	68.352	**< 0.001**

Note: Boldface entries indicate statistical significance ($p < 0.05$). SD, standard deviation.

Changes in Writing Self-Efficacy

Table 2 also reports self-efficacy outcomes across the three SEWS subscales. The HMSC group demonstrated significant immediate gains in Ideation and Language Conventions at T1 (both $p < .01$), with the Ideation subscale showing the larger effect. Self-Regulation scores improved significantly from

T0 to T1 in the HMSC group ($p < .01$) and, notably, this gain was sustained at T2 ($p < .05$), whereas the control group showed no significant change in any self-efficacy dimension across time points. The differential persistence pattern—immediate gains in Ideation and Conventions that partially attenuated, alongside durable improvement in Self-Regulation—suggests that the HMSC framework's structured cognitive layering may cultivate metacognitive monitoring skills that outlast the specific writing techniques acquired during the intervention.

Outcomes of Team Collaboration

Team collaboration satisfaction was higher in the intervention group (mean 4.2/5 at T1) than in the control group (3.6/5, $p < 0.01$), and remained significantly higher at T2. The intervention group showed significantly higher teamwork satisfaction than the control group at both T1 and T2, with an increase of 1.01 points from T0 to T1 ($p = .043$) and a sustained gain of 0.84 points at T2 ($p = .031$). As shown in Table 3, these results suggest the intervention improved not only the immediate quality of collaborative experience but also its longer-term perceived value.

Table 3
Key outcomes from linear mixed model analysis

	Effect	Estimate	SE	95% CI	*t*	*p*	R^2 Marg	R^2 Cond
SSV	Intercept	29.88	0.62	(28.67, 31.09)	48.55	**< .001**	0.653	0.943
	Group Time [T0.T1]	6.36	0.62	(5.16, 7.58)	10.32	0.575		
	Group Time [T0.T2]	3.60	0.62	(2.40, 4.82)	5.85	0.496		
OC	Intercept	14.6	0.35	(13.92, 15.31)	41.26	**< .001**	0.588	0.931
	Group Time [T0.T1]	2.93	0.35	(2.25, 3.62)	8.39	**0.039**		
	Group Time [T0.T2]	1.61	0.35	(0.93, 2.30)	4.62	0.385		
FC	Intercept	30.46	0.65	(29.20, 31.73)	47.08	**< .001**	0.584	0.935
	Group Time [T0.T1]	5.06	0.68	(3.74, 6.40)	7.45	**0.091**		
	Group Time [T0.T2]	2.38	0.68	(1.05, 3.71)	3.5	0.469		
Ideation	Intercept	4.025	0.06	(3.90, 4.15)	62.96	**< .001**	0.702	1.003
	Group Time [T0.T1]	1.04	0.09	(0.87, 1.21)	11.71	**0.015**		
	Group Time [T0.T2]	0.62	0.09	(0.45, 0.80)	7.02	0.223		
Conventions	Intercept	4.02	0.07	(3.88, 4.15)	58.34	**< .001**	0.711	1.007
	Group Time [T0.T1]	1.07	0.1	(0.89, 1.26)	11.16	**0.065**		

	Group[T.HMSC]:Time [T0.T2]	0.82	0.11	(0.63, 1.01)	8.51	0.171		
Self-Regulation	Intercept	3.81	0.06	(3.70, 3.93)	66.66	**< .001**	0.784	0.997
	Group[T.HMSC]:Time [T0.T1]	1.08	0.08	(0.93, 1.23)	13.79	**0.002**		
	Group[T.HMSC]:Time [T0.T2]	0.87	0.08	(0.72, 1.03)	11.2	**0.045**		
TSS	Intercept	3.03	0.06	(2.91, 3.16)	48.07	**< .001**	0.736	1.011
	Group[T.HMSC]:Time [T0.T1]	1.01	0.09	(0.84, 1.19)	11.4	**0.043**		
	Group[T.HMSC]:Time [T0.T2]	0.84	0.09	(0.67, 1.02)	9.46	**0.031**		

R^2 Marg, $R^2_{Marginal}$; R^2 Cond, $R^2_{Conditional}$; T0, Baseline; T1, Immediate Post-test; T2, One-month Follow-up. Boldface entries indicate statistical significance ($p < 0.05$).

The Group× Time [T0.T1] interaction assesses the differential effect of being in the HMSC group compared to the ICAP control group from T0 to T1. Similarly, the Group× Time [T0-T2] interaction evaluates the differential effect from T0 to T2.

Discussion

The findings of this study carry several actionable implications for medical English education. First, the significant short-term gains observed in Sentence Structure and Vocabulary (SSV) and Organization and Coherence (OC) within the HMSC group indicate that AI-augmented collaborative writing instruction is especially well-suited to addressing the linguistic accuracy and structural coherence deficits that have been widely documented among Chinese medical students [1, 3, 4, 2]. Medical academic writing imposes layered cognitive demands—terminological precision and collocational accuracy at the lexical level, and logical progression of evidence at the discourse level—and the HMSC framework's four-layer architecture (Perception → Comprehension → Prediction → Decision) maps directly onto these demands. In the present study, the Information Provider and Feedback Coach agents appear to have functioned as continuously available linguistic resources, partially compensating for the instructor bandwidth constraints that represent a well-documented systemic barrier in medical English education [5]. This finding resonates with broader evidence that medical students are receptive to AI-mediated

instructional support [17, 18], suggesting that the HMSC model enters a pedagogical context that is already favorably disposed toward thoughtfully designed AI integration.

Second, the sustained improvement in writing self-efficacy—particularly in the self-regulation dimension—carries specific significance for medical education. Writing anxiety is a prevalent and well-documented phenomenon among Chinese medical students, one that has been shown to suppress willingness to engage in writing practice and to perpetuate a self-reinforcing cycle of avoidance and skill stagnation [3, 1]. The HMSC framework's design—in which students interact with AI agents at each cognitive layer and receive immediate, non-judgmental, and task-specific feedback—may disrupt this cycle by providing a psychologically safer entry point into academic writing than peer-only or instructor-only feedback configurations. When students can privately test alternative lexical choices, syntactic structures, and rhetorical moves with an agent before presenting their work to human peers, the perceived risk of negative evaluation diminishes, potentially lowering the affective barriers that disproportionately constrain medical students with weaker English foundations. The durability of self-regulation gains at follow-up further suggests that HMSC-style interactions may foster metacognitive habits that persist beyond the immediate intervention window.

Third, the agent's role as a neutral third party in team collaboration addresses a persistent challenge in medical English classrooms: ensuring equitable contribution and constructive peer dynamics when students exhibit wide variation in English proficiency. In conventional group writing tasks, more proficient students tend to dominate interaction while less proficient peers remain passive—a participation asymmetry that exacerbates rather than narrows pre-existing skill gaps [10]. By automatically documenting revision histories and furnishing objective, process-oriented feedback, the HMSC agents in this study shifted the evidentiary basis of peer evaluation from interpersonal impression to data-grounded discussion of specific writing choices. This mechanism may hold particular value in medical education, where collaborative competence is increasingly recognized as a core professional competency—enshrined in frameworks such as the Accreditation Council for Graduate Medical Education (ACGME) core competencies—yet instructional tools for systematically cultivating it within language classrooms remain underdeveloped [32, 33]. Post-intervention interviews corroborated this interpretation: students described assigning diverse functional roles to agents—as virtual mentors providing draft feedback or as cognitive partners coordinating structural revisions—enabling more effective resource allocation and a higher degree of role complementarity within teams.

It is important to acknowledge, however, that the attenuation of writing performance gains at the one-month follow-up—particularly the non-significant result for Format and Content (FC)—indicates that a two-week intervention, while sufficient to generate statistically and practically significant short-term improvements, is insufficient to consolidate durable writing competence. This pattern of partial fade-out is consistent with findings from the broader CSCL literature on the challenge of sustaining intervention effects once structured support is withdrawn [34]. The implication is clear: integrating AI-assisted collaborative writing into medical English curricula should be conceptualized not as a discrete, time-limited intervention but as an ongoing pedagogical infrastructure. A distributed implementation model, in

which HMSC-style agent support is embedded across multiple courses and academic semesters, appears necessary to transform short-term performance gains into stable, transferable writing proficiency. Future iterations of the HMSC platform might also explore discipline-adapted agent configurations—for instance, agents pre-trained on biomedical corpora—to strengthen support for the content-level dimensions of medical academic writing that proved less responsive to the current intervention.

Collectively, these findings provide preliminary empirical support for the three theoretical premises underpinning the HMSC framework. The differential gains in SSV and OC are consistent with the distributed cognition premise: when information retrieval and pattern recognition are allocated to agents, learners can redirect cognitive resources toward higher-order processes such as argument structuring and rhetorical planning. The durable self-regulation gains align with the SRL premise: the structured transition through four cognitive layers, each with embedded metacognitive checkpoints, appears to cultivate monitoring habits that persist beyond the intervention window. The collaboration satisfaction findings support the ICAP-extension premise: introducing an AI agent as a structured interaction partner can sustain interactive engagement in settings where conventional facilitation is infeasible, providing evidence that HMSC can bridge the gap between the theoretical predictions of ICAP and the practical realities of large-enrollment medical English instruction.

Conclusion

This study developed and empirically evaluated the HMSC framework, a theoretically grounded instructional model that integrates LLM-based intelligent agents with collaborative peer learning to support medical students' academic English writing. First, the HMSC framework significantly improved medical students' English academic writing performance in the short term, particularly in Sentence Structure and Vocabulary and in Organization and Coherence. These two dimensions correspond directly to the linguistic accuracy and structural clarity demands that prior research has identified as major barriers for Chinese medical students transitioning from general English to academic English writing. Second, the deeper interactive processes embedded in the HMSC workflow—whereby students engaged with agents across the Perception, Comprehension, Prediction, and Decision layers—appeared to be internalized by learners, producing durable improvements in writing self-efficacy, particularly in the capacity for self-regulation. This finding is notable given the high levels of writing anxiety that characterize this learner population. Third, students in the HMSC group reported significantly higher satisfaction with team collaboration, attributable in part to the agent's role as a complementary and neutral team member that facilitated constructive dialogue and reduced interpersonal friction.

Limitations

Several limitations should be noted. First, the two-week intervention period, while sufficient to detect immediate effects, was too brief to assess the durability of writing performance gains; the observed attenuation at one-month follow-up suggests that longer intervention durations and sustained exposure

to the HMSC framework may be required to establish stable improvements, particularly in content-level writing dimensions. Second, the present study focused exclusively on student–agent interaction during the learning phase and did not incorporate the teacher's role in the assessment and feedback loop. A more comprehensive implementation would systematically integrate pedagogical delivery, the learning process, and evaluation mechanisms within a unified symbiotic framework. Third, the novelty of interacting with AI agents may produce temporary Hawthorne effects that inflate short-term outcomes; longitudinal studies with extended exposure are needed to distinguish durable learning gains from initial engagement effects. Finally, future research should investigate whether gains in writing self-efficacy and collaborative skills transfer to autonomous writing contexts where AI agent scaffolding is withdrawn.

Abbreviations

HMSC	Human-Machine Symbiotic Cognition
ICAP	Interactive, Constructive, Active, Passive
LLM	Large Language Model
CSCL	Computer-Supported Collaborative Learning
SEWS	Self-Efficacy for Writing Scale
TSS	Teamwork Satisfaction Scale
LMM	Linear Mixed Effects Modeling
REML	Restricted Maximum Likelihood
SSV	Sentence Structures and Vocabulary
OC	Organization and Coherence
FC	Format and Content
API	Application Programming Interface
CEFR	Common European Framework of Reference for Languages
GPT	Generative Pre-trained Transformer

Declarations

Ethics approval and consent to participate

This study was approved by the Ethics Committee of Sichuan International Studies University (Approval No. 20250021). All procedures conformed to the Declaration of Helsinki. All participants provided informed written consent prior to participation.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Funding

This research was supported by Ministry of Education of the People's Republic of China (Grant No. DCA230458).

Author Contribution

S.D. was involved in designing the research, analyzing the data, and writing the manuscript. K.C. was involved in writing the proposal, and editing the manuscript. Both authors have reviewed and approved the final version of the manuscript.

Acknowledgement

We are thankful to our students who voluntarily participated in the study and submitted their essays.

Data Availability

The datasets used and analyzed during the current study are available from the corresponding author on reasonable request.

References

1. Ren Y, Gao R, Zhang S, et al. AI-assisted academic writing in medical postgraduate education: a cross-sectional study of L2 challenges, affective barriers, and instructional implications. BMC Med Educ. 2026. https://doi.org/10.1186/s12909-026-09067-0.
2. Wu C, Zhang YW, Li AW. Peer feedback and Chinese medical students' English academic writing development: a longitudinal intervention study. BMC Med Educ. 2023;23:578. https://doi.org/10.1186/s12909-023-04578-w.

3. Cui Y, Wang X, Kaur S. Writing anxiety and self-regulation strategies among medical English students. Int J Engl Lang Lit Stud. 2024;13(3):126–40. https://doi.org/10.55493/5019.v13i3.5102.
4. Wen X, Chen Y, Yan L. The challenges and strategies of Chinese postgraduate students in academic writing. Cogent Educ. 2025;12(1):2488544. https://doi.org/10.1080/2331186X.2025.2488544.
5. Wang C. On improving the English language ability of Science Citation Index papers for medical postgraduates through flipped classroom teaching. Korean J Med Educ. 2023;35(1):55–69. https://doi.org/10.3946/kjme.2023.249.
6. Jeong H, Hmelo-Silver CE. Seven Affordances of Computer-Supported Collaborative Learning: How to Support Collaborative Learning? How Can Technologies Help? Educ Psychol. 2016;51(2):247–65. https://doi.org/10.1080/00461520.2016.1158654.
7. Dashkina A, Kobicheva A, Lazovskaya T, Tokareva E, Tarkhov D, Guselnikova I. Examining the Effectiveness of Computer-Supported Collaborative Learning for Language Proficiency Purposes. Sustainability. 2022. https://doi.org/10.3390/su14105908.
8. Chi MT, Wylie R. The ICAP framework: Linking cognitive engagement to active learning outcomes. Educ Psychol. 2014;49(4):219–43.
9. Rahimi M, Fathi J. Exploring the impact of wiki-mediated collaborative writing on EFL students' writing performance, writing self-regulation, and writing self-efficacy: A mixed methods study. Comput Assist Lang Learn. 2021;35:2627–74. https://doi.org/10.1080/09588221.2021.1888753.
10. Kreijns K, Kirschner PA, Jochems W. Identifying the pitfalls for social interaction in computer-supported collaborative learning environments: A review of the research. Comput Hum Behav. 2003;19(3):335–53.
11. Cress U, Kimmerle J. A systematic and cognitive view on collaborative knowledge building with wikis. Int J Comput Support Collab Learn. 2008;3(2):105–22.
12. Molenaar I. Towards hybrid human-AI learning technologies. Eur J Educ. 2022;57(4):632–45.
13. Li J, Zong H, Wu E, Wu R, Peng Z, Li J. Exploring the potential of artificial intelligence to enhance the writing of English academic papers by non-native English-speaking medical students—the educational application of ChatGPT. BMC Med Educ. 2024;24:736. https://doi.org/10.1186/s12909-024-05736-y.
14. Shahsavar Z, Kafipour R, Khojasteh L, Pakdel F. Is artificial intelligence for everyone? Analyzing the role of ChatGPT as a writing assistant for medical students. Front Educ. 2024;9:1457744. https://doi.org/10.3389/feduc.2024.1457744.
15. Alshakhi A. Effects of the use of generative AI tools on EAP writing development: A case study with medicine undergraduates. Int J Learn Teach Educ Res. 2024;24(5):1–20.
16. Chen Y, Liu X, Zhang H. Leveraging generative AI-powered rubric-indexed feedback as a formative assessment strategy for enhancing medical English education. Educ Inf Technol. 2025. https://doi.org/10.1007/s10639-025-13412-z.
17. Hu N, Jiang XQ, Wang YD, et al. Status and perceptions of ChatGPT utilization among medical students: a survey-based study. BMC Med Educ. 2025;25:831. https://doi.org/10.1186/s12909-025-

07438-0.

18. Alzoubi H, Ba-shammakh SA, Khasawneh S, et al. Exploring the use of AI methods and chatbots in medical education among medical students in Jordan: a cross-sectional study. BMC Med Educ. 2026;26:518. https://doi.org/10.1186/s12909-026-05518-0.
19. Darvishi A, Khosravi H, Sadiq S, Gašević D, Siemens G. Impact of AI assistance on student agency. Comput Educ. 2024;210:104967.
20. Barrot JS. Using ChatGPT for second language writing: Pitfalls and potentials. Assess Writ. 2023;57:1–6. https://doi.org/10.1016/j.asw.2023.100745.
21. Inga J, Ruess M, Robens JH, Nelius T, Rothfuß S, Kille S, et al. Human-machine symbiosis: A multivariate perspective for physically coupled human-machine systems. Int J Hum Comput Stud. 2023;170:102926.
22. Alfredo R, Echeverria V, Jin Y, Yan L, Swiecki Z, Gašević D et al. Human-centred learning analytics and AI in education: A systematic literature review. Comput Educ Artif Intell. 2024;100215.
23. Winne PH. Learning strategies, study skills, and self-regulated learning in postsecondary education. Higher Education: Handbook of Theory and Research: Volume 28. Netherlands: Springer; 2013. pp. 377–403.
24. Bender SM. Awareness of Artificial Intelligence as an Essential Digital Literacy: ChatGPT and Gen-AI in the Classroom. Chang Engl. 2024;31(2):161–74. https://doi.org/10.1080/1358684X.2024.2309995.
25. Stenalt MH, Lassesen B. Does student agency benefit student learning? A systematic review of higher education research. Assess Eval High Educ. 2021;47(5):653–69. https://doi.org/10.1080/02602938.2021.1967874.
26. Yan D. Impact of ChatGPT on learners in a L2 writing practicum: An exploratory investigation. Educ Inf Technol. 2023;28(11):13943–67. https://doi.org/10.1007/s10639-023-11742-4.
27. Yan D. Feedback seeking abilities of L2 writers using ChatGPT: A mixed method multiple case study. Kybernetes. 2024. https://doi.org/10.1108/K-09-2023-1933.
28. Faul F, Erdfelder E, Buchner A, Lang AG. Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. Behav Res Methods. 2009;41(4):1149–60. https://doi.org/10.3758/BRM.41.4.1149.
29. Bruning R, Dempsey M, Kauffman DF, McKim C, Zumbrunn S. Examining dimensions of self-efficacy for writing. J Educ Psychol. 2013;105(1):25–38. https://doi.org/10.1037/a0029692.
30. Ramos-Villagrasa PJ, Sánchez-Iglesias I, Grande-de-Prado M, Oliván-Blázquez B, Martín-Peña J, Cancer-Lizaga P. Spanish version of Self-Efficacy for Writing Scale (SEWS). Anales de Psicología. 2017;34(1):86. https://doi.org/10.6018/analesps.34.1.264931.
31. Ku HY, Tseng HW, Akarasriworn C. Collaboration factors, teamwork satisfaction, and student attitudes toward online collaborative learning. Comput Hum Behav. 2013;29(3):922–9. https://doi.org/10.1016/j.chb.2012.12.019.

32. Mageira K, Pittou D, Papasalouros A, Kotis K, Zangogianni P, Daradoumis A. Educational AI Chatbots for Content and Language Integrated Learning. Appl Sci. 2022. https://doi.org/10.3390/app12073239.
33. Xue B, Green R, Zhang M. Artificial Intelligence in New Zealand: Applications and innovation. J R Soc N Z. 2023;53:1–5. https://doi.org/10.1080/03036758.2023.2170165.
34. Zheng B, Warschauer M, Farkas G. Digital writing and diversity: The effects of school laptop programs on literacy processes and outcomes. J Educ Comput Res. 2013;48(3):267–99.

Figures

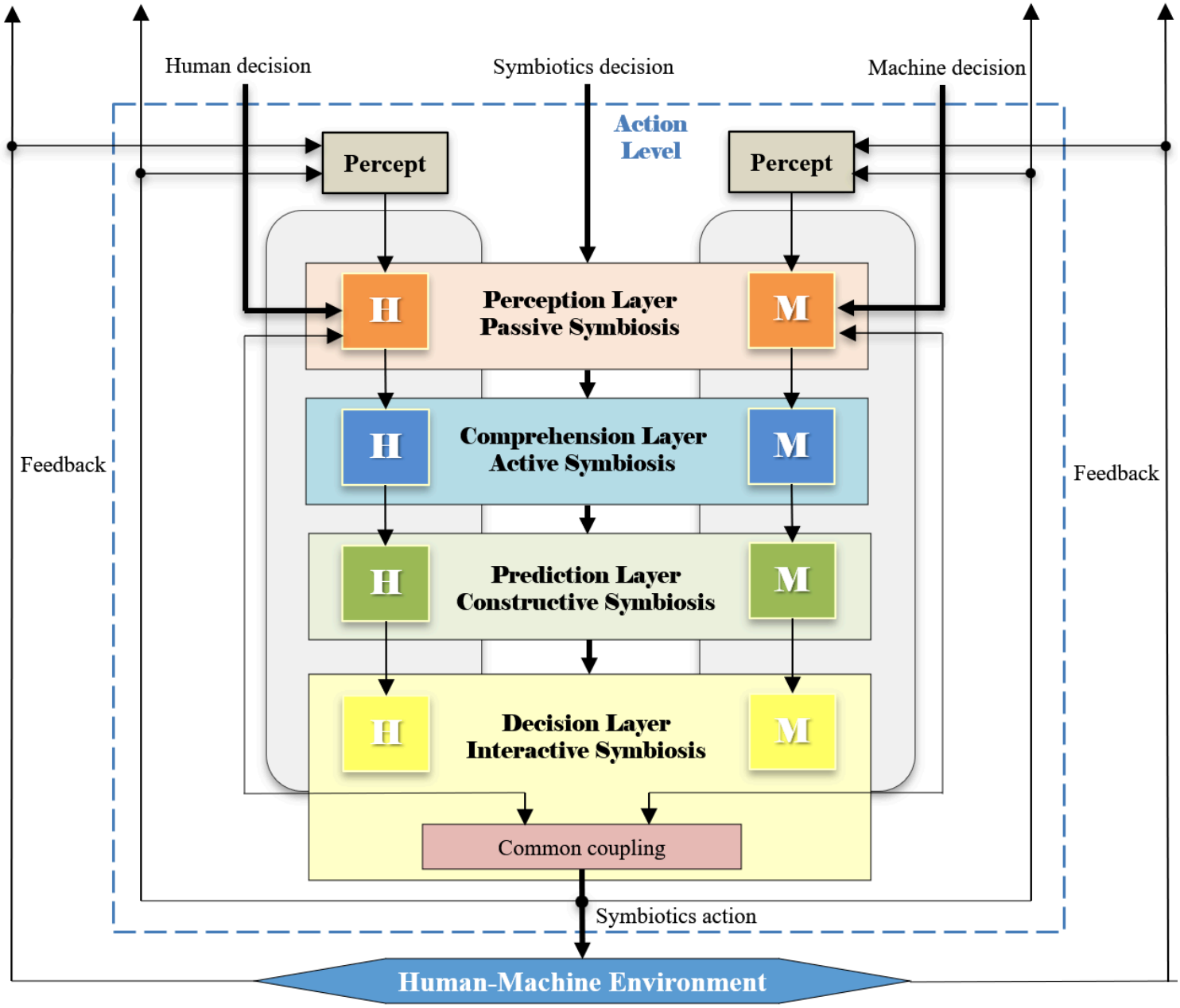


Figure 1

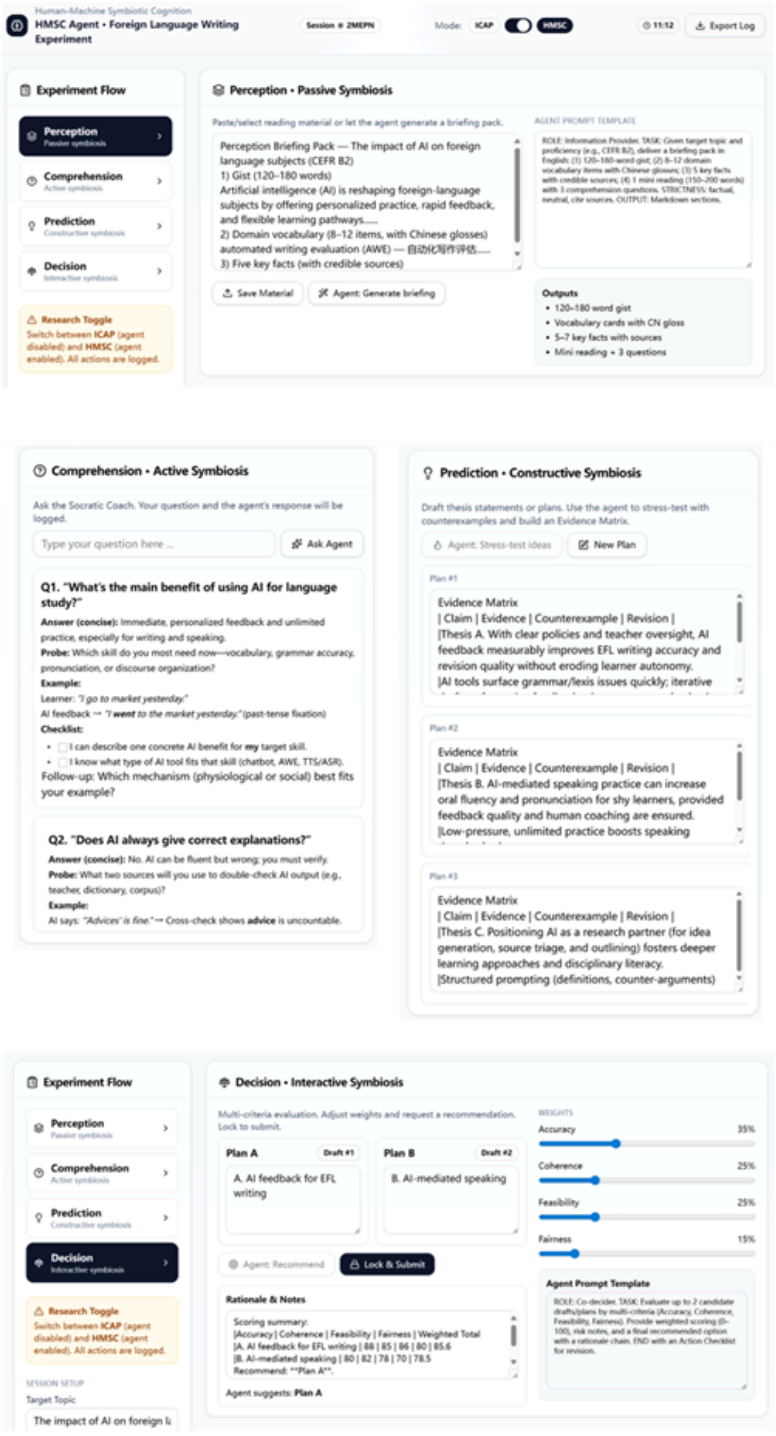


Figure 2

Examples of functions available across the four collaboration modes.

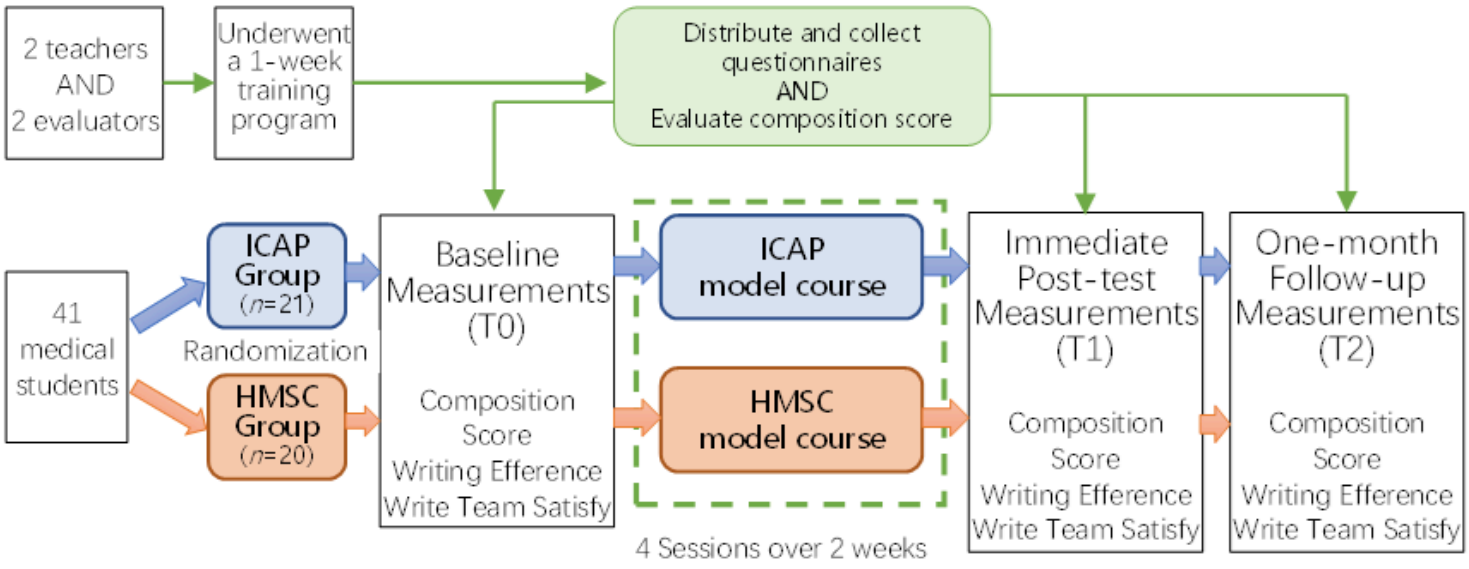


Figure 3

Flowchart of experimental procedure

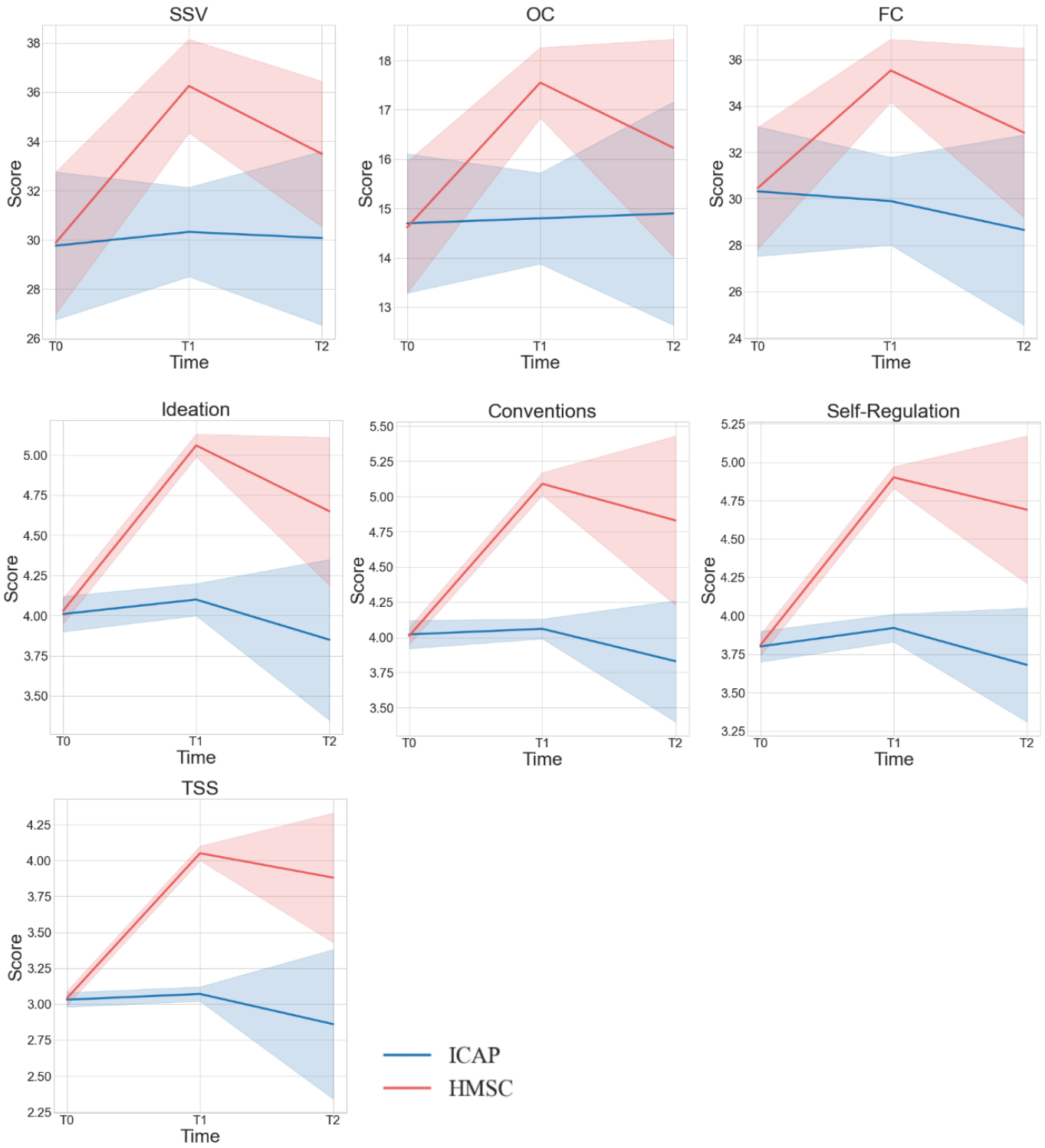


Figure 4

Longitudinal changes in study variables by group.

Note: 1. The writing assessment comprises three components: SSV, OC, and FC, referring to Sentence Structures and Vocabulary, Organization and Coherence, and Format and Content, respectively. 2. The

Self-Efficacy for Writing Scale (SEWS) consists of three subscales: Ideation, Conventions, and Self-Regulation. 3. TSS stands for the Teamwork Satisfaction Scale.