

Proactive Push versus Reactive Response Artificial Intelligence for Thyroid Cancer Guideline Learning Among Resident Physicians: A Nationwide Online Randomized Controlled Trial

Hao Liu

Second Xiangya Hospital of Central South University

ShuJuan Luo

Hunan Normal University

YiLin Liu

Hunan Normal University

Huimin Jin

Second Xiangya Hospital of Central South University

Ye Fei

yefei@csu.edu.cn

Second Xiangya Hospital of Central South University

Research Article

Keywords: thyroid cancer, residency training, large language model, artificial intelligence, proactive tutoring, knowledge retention, randomized controlled trial

Posted Date: October 5th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10247340/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

Generative artificial intelligence (AI) is increasingly used in medical education, but most current tools remain reactive: learners must initiate questions before receiving assistance. Whether an AI tutor that proactively delivers contextualized, individualized learning prompts improves immediate learning and delayed retention remains uncertain. Thyroid cancer guidelines, which require frequent updating and multidisciplinary integration, provide a suitable test case.

Methods

We conducted a nationwide online, parallel-group, assessor-blinded, active-controlled randomized trial from 1 March to 1 May 2026. Resident physicians in standardized residency training were recruited online and randomized 1:1 to a proactive push AI-assisted learning group or a reactive response chatbot control group. Both groups used the same Feishu-OpenClaw platform, knowledge base, question bank, prompts, and ChatGPT-5.4 model. The proactive group received two individualized micro-quizzes daily with immediate feedback; the control group accessed the same content only when initiating interaction. The primary outcome was the 40-item total knowledge score at Week 3 (T3w). Key secondary outcomes included delayed retention at Week 7 (T7w), module scores, engagement, guideline-application self-efficacy, and user experience.

Results

A total of 120 residents were randomized; 105 (87.5%) completed the T3w primary assessment, and 90 (75.0%) completed all scheduled assessments through T7w. In intention-to-treat analyses, the adjusted T3w total score favored the proactive group but did not reach conventional statistical significance (adjusted difference, 0.65 points; 95% CI, -0.04 to 1.34; $p = 0.065$). The proactive group scored higher on the T3w adverse drug reaction module (0.89 points; 95% CI, 0.13 to 1.65; $p = 0.022$) and T7w delayed retention (1.49 points; 95% CI, 0.36 to 2.62; $p = 0.009$). It also showed more login days, more micro-quiz completions, more learner-initiated questions, and greater self-efficacy gains (2.83 points; 95% CI, 1.51 to 4.16; $p < 0.001$).

Conclusions

When content, knowledge base, and model were held constant, proactive push AI tutoring did not clearly improve the prespecified immediate total score, but it produced more consistent advantages in delayed retention, adverse drug reaction knowledge, engagement, self-efficacy, and perceived learning

experience. AI proactivity may act primarily by increasing distributed retrieval practice rather than by immediately increasing short-term total scores.

Trial registration:

Chinese Clinical Trial Registry, ChiCTR2600121101. Retrospectively registered on 25 March 2026.

Introduction

Thyroid cancer is one of the common malignancies whose clinical burden has continued to rise in China. National cancer statistics indicate that China still carries a substantial cancer burden, with thyroid cancer among the high-incidence cancers [1]. At the same time, the 5-year age-standardized relative survival rate for thyroid cancer in China reached 92.9% during 2019–2021 [2]. These epidemiological patterns suggest that thyroid cancer care has moved beyond survival alone toward consistent, evidence-based decision-making, standardized management, and long-term follow-up quality. For residency training, the educational challenge is how to transform rapidly updated thyroid cancer evidence into knowledge that residents can retain and transfer to clinical contexts.

In China, guideline-based learning is supported by national thyroid cancer diagnosis and treatment guidelines and updated Chinese Anti-Cancer Association (CACA) and Chinese Society of Clinical Oncology (CSCO)-related guidance [3, 4]. However, residents rotate across thyroid surgery, general surgery, head and neck surgery, nuclear medicine, endocrinology, medical oncology, and radiation oncology. This fragmented rotation structure makes it difficult to integrate multidisciplinary guideline knowledge during limited, discontinuous learning time.

Large language models (LLMs) are rapidly entering health professions education. Recent bibliometric and review work suggests that ChatGPT-related medical education research expanded quickly after 2023 [5–7]. A recent meta-analysis of randomized controlled trials further supports the educational relevance of generative artificial intelligence-based teaching in medical education. Yet many current applications remain grounded in a reactive paradigm: learners ask a question, and the AI system responds [8]. Such designs are useful for clarification, summarization, and case simulation, but they often provide limited support for sustained, multi-stage, context-aware learning scaffolds [9].

The central gap is therefore not simply whether AI can answer medical questions, but whether the educational behaviour of an AI system matters. Learning science consistently supports spaced practice and retrieval practice as mechanisms for durable learning, but many chatbot interventions have not systematically translated these principles into reproducible, daily, push-enabled interventions [10–12]. A proactive AI tutor is designed to address this gap: if the same content can be delivered more frequently, more personally, and with less friction, learners may receive a higher dose of retrieval practice, sustain engagement, and retain knowledge more effectively.

Emerging agentic AI applications also suggest the feasibility of personalized digital support in health-related contexts [13]. In this study, we focused on AI proactivity rather than content quality. Using the same Feishu-OpenClaw platform, knowledge base, question bank, system prompts, and ChatGPT-5.4 model, we compared a proactive push AI tutor with a functionally equivalent reactive response chatbot for thyroid cancer guideline learning. We hypothesized that, compared with reactive response AI, proactive push AI would improve immediate post-intervention knowledge and further enhance delayed retention, learning engagement, and guideline-application self-efficacy.

Methods

Study Design and Oversight

This nationwide online, parallel-group, assessor-blinded, active-controlled randomized trial compared proactive push AI-assisted learning with a reactive response AI chatbot for thyroid cancer guideline learning among resident physicians. The study was conducted from 1 March to 1 May 2026. Participants were recruited nationally through DXY, Yixuejie, Xiaohongshu, Weibo, and WeChat groups relevant to medical training. The study was approved by the Biomedical Research Ethics Committee of Hunan Normal University (approval No. Lun Li Ke 2026 (430); approved 8 January 2026). All participants provided electronic informed consent after online eligibility screening and before enrollment. All study procedures were performed in accordance with the Declaration of Helsinki and relevant institutional and national guidelines and regulations. The trial was retrospectively registered with the Chinese Clinical Trial Registry on 25 March 2026 (registration number: ChiCTR2600121101). This report was prepared in accordance with the CONSORT 2025 statement and the CONSORT-AI extension [23, 24]. The research team reviewed recruitment progress, protocol deviations, and technical or use-related adverse events at regular intervals.

Participants

Eligible participants were resident physicians aged 22 years or older who were enrolled in standardized residency training at tertiary general hospitals or oncology-specialty hospitals. Residents were in postgraduate year (PGY) 1–3 and were currently rotating, or scheduled to rotate within 6 months, in a thyroid cancer-relevant department, including thyroid surgery, general surgery, head and neck surgery, nuclear medicine, endocrinology, medical oncology, or radiation oncology. Additional inclusion criteria were daily Feishu use of at least 30 minutes, access to a smartphone for study messages, willingness to provide electronic informed consent, and commitment to complete all assessments.

Exclusion criteria were participation in systematic CACA or CSCO thyroid cancer guideline training during the previous 3 months, except routine hospital continuing medical education; completion of specialized thyroid surgery or head and neck surgery training; planned leave of more than 7 days, maternity leave, or

external training of more than 2 weeks during the study period; known inability to use LLM-based tools; and participation in a comparable AI-assisted learning trial within the previous year.

Randomization and Blinding

Eligible participants were allocated in a 1:1 ratio to the proactive push group or reactive response control group through a central electronic randomization system. Randomization was stratified by PGY level (PGY-1, PGY-2, or PGY-3) and baseline knowledge level (above or below the median) using randomly varied block sizes of 4 or 6. The allocation sequence was generated by an independent statistician using R version 4.5 and stored on an encrypted server that was inaccessible to recruitment personnel. Participants were partially blinded: the consent form stated that they would be assigned to one of two AI-assisted learning modes but did not disclose the specific manipulation of proactivity. Outcome assessors and statistical analysts remained blinded to group identity until the primary analysis was completed and used only coded group labels.

Interventions

Both groups accessed a ChatGPT-5.4-powered OpenClaw AI agent through the Feishu workplace communication platform for a 3-week intervention. The only planned design difference was whether the system initiated learning interactions. The two groups shared the same backend LLM, knowledge base, question bank, system prompt, and individualized item-selection logic. The structured knowledge base included the 2025 CACA thyroid cancer guideline, CSCO-related guidance, and adverse drug reaction (ADR) management content organized as a structured JSON database. The question bank contained 180 items. The unified system prompt specified that the AI tutor should be friendly, professional, and supportive without excessive encouragement.

In the proactive push group, the AI agent initiated one micro-quiz at 12:30 and one at 20:00 each day. Items were selected according to each learner's prior accuracy, response latency, and forgetting-curve parameters using an adaptive spaced-repetition algorithm based on the Free Spaced Repetition Scheduler (FSRS) approach [15]. After each response, participants received immediate explanatory feedback, guideline rationale, and ADR-related clinical points. They could also ask free-text questions about any guideline topic within the study scope at any time.

In the reactive response group, participants received a single message on the study start date stating that the assistant would be available for the next 3 weeks and that they could initiate questions whenever needed. Thereafter, the agent did not initiate conversations. Participants could access the system at any time, ask free-text questions, or request micro-quiz practice. When they requested a micro-quiz, the same FSRS-based selection logic and the same feedback structure were used. This design held the model, content, learning channel, system memory, and feedback framework constant while isolating proactive push as the intervention variable.

Outcomes

The primary outcome was the immediate post-intervention total knowledge score at Week 3 (T3w). The test included 40 objective single-best-answer or multiple-response items: 20 guideline items and 20 ADR items. Scores ranged from 0 to 40 and were automatically scored by the system.

Key secondary outcomes were: (1) delayed retention at Week 7 (T7w), assessed by a 40-item test that included 20 equivalent retention items and 20 clinical transfer vignettes; (2) guideline and ADR module subscores at T3w and T7w; (3) engagement indicators, including login days, number of completed micro-quizzes, push response rate, learner-initiated questions, and mean session duration; (4) guideline-application self-efficacy, measured by a 10-item scale adapted from Bandura's self-efficacy principles (pilot Cronbach's $\alpha = 0.89$); (5) technology acceptance, measured by a 12-item short form based on the Unified Theory of Acceptance and Use of Technology (UTAUT) [16]; (6) user satisfaction, assessed by net promoter score and a 5-item experience scale; and (7) weekly self-reported examples of clinical knowledge transfer.

Exploratory outcomes included subjective cognitive load, measured using the Chinese version of the National Aeronautics and Space Administration Task Load Index (NASA-TLX); perceived personalization, measured by a 5-item scale; and mediation analyses of engagement indicators. Safety events were classified as technical events, such as system downtime, message delays exceeding 24 hours, and erroneous pushes, or use-related events, such as self-reported AI-related clinical misdirection, perceived Feishu burden, or privacy concerns. Events were recorded through an electronic adverse-event form; serious events were to be reported to the ethics committee within 48 hours.

Data Collection

Baseline assessments were completed before randomization and included demographic information, training characteristics, a 40-item baseline knowledge test, the self-efficacy scale, the UTAUT short form, a short form of the previously published Motivated Strategies for Learning Questionnaire (MSLQ), and an AI ethics attitude questionnaire. The MSLQ was developed by Pintrich and colleagues and was not developed specifically for this study [25]. During the intervention, process data including logins, session duration, push responses, micro-quiz attempts and accuracy, learner-initiated questions, and technical failures were automatically recorded by the OpenClaw system.

Post-intervention assessments were conducted at T3w and included a 40-item knowledge test parallel to the baseline test, the self-efficacy scale, the UTAUT short form, NASA-TLX, net promoter score, satisfaction scale, perceived personalization scale, and open-ended feedback. Delayed retention was assessed at T7w using a 40-item test with equivalent retention items and clinical transfer vignettes. Questionnaires were administered through REDCap or Wenjuanxing browser-based forms. OpenClaw behavioural logs were exported through an application programming interface to a PostgreSQL 15 research database at prespecified intervals.

Statistical Analysis

Sample size was estimated for the primary outcome, the T3w total knowledge score. Assuming a medium effect size (Cohen's $d = 0.55$), two-sided alpha of 0.05, and power of 0.85, at least 90 participants were required. Allowing for 25%-30% attrition or withdrawal, the planned sample size was 120 participants, with 60 per group. No formal interim efficacy analysis or adaptive sample-size adjustment was planned.

The primary analysis followed the intention-to-treat principle and included all randomized participants. Missing data were handled using multiple imputation with 20 imputed datasets. The per-protocol set was defined as participants who completed at least 70% of intervention days and completed the T3w assessment. As-treated analyses were conducted as sensitivity analyses.

The primary outcome was analyzed using analysis of covariance (ANCOVA), with group as the fixed effect and baseline total knowledge score and PGY level as covariates. Results are reported as adjusted between-group differences, 95% confidence intervals (CIs), and two-sided p values; standardized effect sizes and 95% CIs are reported where calculable. Because no minimal clinically important difference was prespecified, we did not make threshold-based claims of educational importance.

Longitudinal knowledge outcomes across baseline, T3w, and T7w were analyzed using linear mixed models with fixed effects for time, group, and the time-by-group interaction, and random intercepts for participants. Count engagement outcomes were analyzed using negative binomial regression. Continuous scale outcomes were analyzed using ANCOVA of change scores or generalized linear models adjusted for baseline values. Prespecified subgroup analyses examined interactions between group and PGY level, baseline knowledge level, and prior AI use frequency. Mediation analyses used a regression framework with 5000 bootstrap resamples to estimate indirect effects. For multiple secondary outcomes, the Benjamini-Hochberg procedure was used to control the false discovery rate. Sensitivity analyses included per-protocol analyses, as-treated analyses, and robustness analyses under different missing-data assumptions. All analyses were conducted in R version 4.5, using packages including lme4, mice, mediation, and ggplot2. The statistical analysis plan was finalized and locked before database unlocking.

Results

Participant flow and baseline characteristics

Between 1 March and 1 May 2026, 120 participants were enrolled and randomized: 60 to the proactive push group and 60 to the reactive response control group. Figure 1 presents the CONSORT participant flow. In the control group, 42 participants (70.0%) completed all scheduled assessments through T7w, 8 were lost between T3w and T7w, 6 were lost before T3w, and 4 withdrew. In the proactive push group, 48 participants (80.0%) completed all scheduled assessments, 7 were lost between T3w and T7w, 3 were

lost before T3w, and 2 withdrew. Thus, 105 participants (87.5%) had available T3w primary endpoint data, and 90 (75.0%) completed all scheduled assessments through T7w. All 120 randomized participants were included in the intention-to-treat analysis, with missing outcomes handled by multiple imputation. Outcome availability and missingness patterns are shown in Supplementary Figure S1.

The two groups were broadly similar at baseline, although some imbalances remained, as expected in a modestly sized randomized trial and were addressed in adjusted analyses. The mean age was 27.32 years in the proactive group and 26.88 years in the control group. Across all participants, 68 (56.7%) were male and 52 (43.3%) were female. Mean baseline total knowledge scores were 16.60 in the proactive group and 16.98 in the control group (Table 1).

Primary Outcome

In the intention-to-treat analysis, the ANCOVA-adjusted group difference in T3w total knowledge score favored the proactive push group but did not reach the prespecified threshold for statistical significance (adjusted difference, 0.65 points; 95% CI, -0.04 to 1.34; $p = 0.065$) (Table 2; Fig. 2; Supplementary Figure S2A). Therefore, proactive push AI-assisted learning showed a directional advantage on the immediate total score, but this advantage should not be interpreted as a statistically significant improvement in the primary endpoint.

Key Secondary Outcomes

In modular analyses at T3w, the proactive push group outperformed the control group on ADR knowledge (adjusted difference, 0.89 points; 95% CI, 0.13 to 1.65; $p = 0.022$). In contrast, the adjusted difference for the T3w guideline module was -0.50 points (95% CI, -1.28 to 0.27; $p = 0.203$), providing no clear evidence of benefit for that module.

At T7w, the proactive push group had a higher delayed retention total score than the control group (adjusted difference, 1.49 points; 95% CI, 0.36 to 2.62; $p = 0.009$). Decomposition of this outcome showed a higher original-item retention score in the proactive group (adjusted difference, 1.01 points; 95% CI, 0.17 to 1.85; $p = 0.019$), but the difference in transfer vignette scores was not statistically significant (adjusted difference, 0.35 points; 95% CI, -0.53 to 1.22; $p = 0.438$) (Table 2; Supplementary Figure S2B-C).

The longitudinal mixed model showed similar baseline trajectories, followed by separation at T3w and T7w (Fig. 3). The distributions of score changes from baseline to T3w, baseline to T7w, and T3w to T7w further supported a relative retention advantage for the proactive push group (Fig. 3; Supplementary Figures S2B-C and S7A-B). These findings suggest that proactive push primarily improved overall delayed retention and original-item retention, whereas evidence for transfer to new clinical vignettes was limited.

Engagement and Behavioral Mechanisms

The proactive push group showed substantially higher engagement across core indicators. Over the 3-week intervention, mean login days were 20.22 in the proactive group and 11.73 in the control group (adjusted difference, 8.56 days; 95% CI, 7.26 to 9.87; $p < 0.001$). Mean total micro-quiz completions were 53.57 and 19.88, respectively (adjusted difference, 34.11; 95% CI, 30.75 to 37.47; $p < 0.001$). Mean learner-initiated questions were 23.27 and 13.45, respectively (adjusted difference, 10.30; 95% CI, 8.75 to 11.85; $p < 0.001$). Time-trend plots indicated that the proactive group maintained high-frequency learning activity throughout the intervention, with consistently higher micro-quiz completion and login frequency. Daily trajectories also showed more regular patterns of login, micro-quiz completion, learner-initiated questioning, and push response in the proactive group (Table 3; Fig. 4; Supplementary Figure S3A-D). These findings suggest that proactive push increased not only overall use but also the regularity of distributed learning.

Self-Efficacy and User Experience

Self-efficacy gains from baseline to T3w were greater in the proactive push group than in the control group (mean change, 6.89 versus 4.02; adjusted difference, 2.83; 95% CI, 1.51 to 4.16; $p < 0.001$) (Table 2; Supplementary Figure S5A). The proactive group also reported higher post-intervention technology acceptance, satisfaction, and perceived personalization (Table 3; Supplementary Figure S5B-D). NASA-TLX scores were lower in the proactive group, but the difference was not statistically significant, suggesting that proactive push did not increase subjective cognitive load (Table 3).

Safety Events

No technical or use-related adverse events were reported in either group during the study.

Subgroup, Mediation, and Sensitivity Analyses

Subgroup analyses showed a consistent direction of effect across PGY levels, baseline knowledge strata, and prior AI use frequency, with no significant interaction effects (Supplementary Figure S4). Mediation analysis suggested that total micro-quiz completions partially mediated the relationship between group assignment and T3w knowledge score (Supplementary Figure S6). Sensitivity analyses, including per-protocol analyses, as-treated analyses, and multiple imputation under alternative missing-data assumptions, yielded results in the same direction. These analyses support the robustness of the engagement and delayed-retention findings, while the primary endpoint remains a non-significant directional result.

Discussion

In this nationwide online randomized trial, proactive push AI-assisted learning did not clearly improve the prespecified immediate total knowledge score at T3w. However, it showed more consistent advantages in delayed knowledge retention, ADR knowledge, guideline-application self-efficacy, engagement, and subjective learning experience. These findings suggest that proactive AI may work less by producing an

immediate score gain and more by sustaining learning behaviour and converting a higher dose of retrieval practice into better medium-term retention.

The adjusted difference for the immediate primary outcome was 0.65 points, representing a small effect and falling short of statistical significance. Therefore, the present study should not be framed as proof that proactive AI improves immediate total knowledge. The educationally more informative signal appears in the secondary outcomes. The proactive group performed better on ADR knowledge and delayed retention at T7w, suggesting that proactive micro-quizzing may be especially useful for granular, practice-relevant content that is easy to forget and difficult to consolidate through passive reading or occasional self-initiated querying alone.

This interpretation aligns with the spacing and testing effects: distributed retrieval practice tends to support longer-term retention more effectively than massed or passive exposure [15, 19]. In the present trial, the AI tutor operationalized this principle by using adaptive item selection and embedding two daily retrieval opportunities into participants' routine communication platform. Rather than simply serving as an external source of answers, the proactive agent repeatedly prompted learners to retrieve and elaborate guideline knowledge.

The delayed-retention advantage also addresses a concern frequently raised in AI-assisted learning: that external AI memory might encourage cognitive offloading and weaken internal knowledge consolidation [9]. Our findings suggest a more nuanced interpretation. When AI use is designed around regular retrieval practice rather than passive information access, it may mitigate cognitive offloading and instead function as a scaffold for elaborative rehearsal. Engagement indicators provide convergent mechanism evidence: the proactive group completed roughly three times as many micro-quizzes as the control group and logged in on nearly twice as many days, transforming fragmented daily time into structured learning opportunities.

Prior LLM studies in medical education have largely examined passive, on-demand chatbot interfaces [5–9, 20]. Such studies have established feasibility, acceptability, and knowledge-assessment performance, but they have rarely isolated AI proactivity as an educational design variable. Chung and colleagues, for example, showed that GPT-4 Turbo, used in a passive response mode with structured clinical context and advanced prompting, achieved meaningful performance in perioperative risk prediction [21]. The present study extends this literature by experimentally manipulating interaction initiation while holding content and model constant. The results suggest that engagement, a prerequisite for any educational intervention, is highly sensitive to proactivity design. Mediation findings further support engagement dose as one plausible mechanism linking proactive design to learning outcomes [17–18].

These findings have practical implications for postgraduate medical education. First, implementation through Feishu, an existing workplace communication channel, may have improved ecological validity and reduced adoption friction. The completion rate through T7w was 80.0% in the proactive group compared with 70.0% in the control group, suggesting that embedding AI tutoring into an existing

workflow may be preferable to requiring learners to adopt a separate educational application. Second, the self-hosted OpenClaw configuration demonstrates the feasibility of deploying AI tutoring in a nationally recruited online education study while supporting institutional data governance. Third, the module-specific pattern suggests that proactive AI tutoring may be most valuable for reinforcement-intensive content, such as ADR recognition and guideline updates, whereas broader conceptual frameworks may require complementary case-based discussion, workplace teaching, or supervised clinical practice. Future AI education designs may therefore combine proactive push for reinforcement with reactive response for exploratory learning.

Limitations

Several limitations should be considered. First, the prespecified primary endpoint did not meet the conventional statistical significance threshold ($p = 0.065$). All claims about educational benefit must therefore be interpreted in this context, and significant secondary outcomes should not be used to retrospectively redefine the primary conclusion. Second, although T7w total and original-item retention favored proactive AI tutoring, transfer vignette scores did not differ significantly, suggesting that the current intervention strengthened consolidation more than cross-context transfer. Third, the intervention lasted only 3 weeks. This duration matched the practical rhythm of residency rotations but cannot determine whether proactivity advantages persist over longer teaching cycles. Fourth, although contamination controls were implemented, cross-group information exchange and subclinical contamination could not be completely excluded in a nationally recruited, chat-platform-based trial. Fifth, participants were mainly residents from tertiary hospitals who used Feishu frequently and had relatively high digital readiness; generalizability to primary care settings, learners with lower digital engagement, or other communication platforms should be tested. Sixth, no usual-education group without AI support was included, so the trial estimates the comparative effect of proactive versus reactive AI within an active-control framework rather than the absolute effect of AI tutoring versus traditional education.

Conclusion

With the knowledge content, knowledge base, and LLM held constant, proactive push AI-assisted learning did not clearly improve the immediate total knowledge score, but it showed more consistent advantages in delayed retention, ADR knowledge, self-efficacy, engagement, and perceived learning experience. These findings suggest that, in residency education, AI proactivity primarily changes the frequency and rhythm of learning, which in turn may affect knowledge retention. Proactivity should therefore be considered a core design parameter of LLM-based medical education tools rather than a superficial interface feature.

Declarations

Consent for publication

Not applicable. The manuscript contains no identifiable individual participant data.

Competing interests

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Ethics approval and consent to participate

The study was approved by the Biomedical Research Ethics Committee of Hunan Normal University (approval No. 20260430; approved 8 January 2026). All study procedures involving human participants and human data were performed in accordance with the Declaration of Helsinki and relevant institutional and national guidelines and regulations. All participants provided electronic informed consent after eligibility screening and before enrollment.

Availability of data and materials

Deidentified individual participant data that support the findings of this study, together with the data dictionary, study protocol, statistical analysis plan, and analysis code, are available from the corresponding author on reasonable request after publication, subject to approval by the responsible ethics committee and execution of an appropriate data-use agreement. Raw Feishu/OpenClaw free-text interaction logs are not publicly available because they may contain potentially identifiable information. Aggregate data supporting the conclusions are included in this article and its supplementary files.

Funding

None.

Acknowledgements

The authors acknowledge and express their deepest gratitude to the participants of this study. We acknowledged the contribution of those who helped in improving the English language writing of this manuscript.

Author Contributions

Hao Liu and Ye Fei conceived and designed the study. Hao Liu developed the study protocol, coordinated participant recruitment, managed data collection, performed data analysis, and drafted the manuscript. ShuJuan Luo contributed to participant recruitment, data collection, study discussion, and language editing. YiLin Liu contributed to questionnaire administration, data verification, and organization of study materials. Huimin Jin contributed to clinical content review, interpretation of thyroid cancer guideline-related findings, and manuscript revision. Ye Fei supervised the project, provided methodological and

clinical oversight, critically reviewed the manuscript for important intellectual content, and approved the final version. All authors read and approved the final manuscript.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used ChatGPT-4o for language polishing. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

1. Han BF, Zheng RS, Zeng HM, et al. Age-specific characteristics of cancer incidence and mortality in China, 2024. *China Cancer*. 2026;35(03):163–71.
2. Lin Y, Wu Y. Trends in incidence and overdiagnosis of thyroid cancer in China, Japan, and South Korea. *Cancer Sci*. 2023;114(10):4052–62. 10.1111/cas.15909.
3. Lin YS, Wang RF, Huang R, Wen Q, Cao W, Chen LB, et al. Chinese management guidelines for radioactive iodine-refractory differentiated thyroid cancer (2025 edition). *Eur J Nucl Med Mol Imaging*. 2025;52(10):3859–76. 10.1007/s00259-025-07222-1.
4. National Health Commission of the People's Republic of China. National guidelines for diagnosis and treatment of thyroid cancer 2022 in China (English version). *Chin J Cancer Res*. 2022;34(3):131–50. 10.21147/j.issn.1000-9604.2022.03.01.
5. Zhang Y, Xie X, Xu Q. ChatGPT in medical education: Bibliometric and visual analysis. *JMIR Med Educ*. 2025;11:e72356. 10.2196/72356.
6. Sato H, Ogasawara K, Sakurai H. Performance evaluation of 18 generative AI models in the 2024 Japanese pharmacist licensing examination: Comparative study. *JMIR Med Educ*. 2025;11:e76925. 10.2196/76925.
7. Lin Y, Luo Z, Ye Z, Zhong N, Zhao L, Zhang L, et al. Applications, challenges, and prospects of generative artificial intelligence empowering medical education: Scoping review. *JMIR Med Educ*. 2025;11:e71125. 10.2196/71125.
8. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *N Engl J Med*. 2023;388(13):1233–9.
9. Gallifant J, Fiske A, Levites Strekalova YA et al. Peer review of GPT-4 technical report and systems card. *PLOS Digit Health*. 2024;3(1):e0000417. 10.1371/journal.pdig.0000417
10. Han ER, Yeo S, Kim MJ, Lee YH, Park KH, Roh H. Medical education trends for future physicians in the era of advanced technology and artificial intelligence: An integrative review. *BMC Med Educ*. 2019;19(1):460. 10.1186/s12909-019-1891-5.
11. Xie Q, Zhu Y, Zhao X. Factors associated with healthy lifestyles in Chinese older adults with chronic conditions: A comparison of online and offline social support, mediated by health management self-

- efficacy and moderated by online patient-centered communication. *Digit Health*. 2026;12:20552076261443840. 10.1177/20552076261443840.
12. Wang C, Wang L, Zhang X, Qi H. The influencing factors of medical postgraduates' usage intention toward artificial intelligence-generated content tools in academic research: Qualitative analysis. *J Med Internet Res*. 2026;28:e77928. 10.2196/77928.
 13. Goh HL, Sancenon V, Chu BMX, et al. Personalised health plan development using agentic AI in Singapore's national preventive care programme: A pilot study. *NPJ Digit Med*. 2026;9(1):332. 10.1038/s41746-026-02514-8.
 14. Cruz Rivera S, Liu X, Chan AW, Denniston AK, Calvert MJ, SPIRIT-AI and CONSORT-AI Working Group. Guidelines for clinical trial protocols for interventions involving artificial intelligence: The SPIRIT-AI extension. *Nat Med*. 2020;26(9):1351–63. 10.1038/s41591-020-1037-7.
 15. Tabibian B, Upadhyay U, De A, Zarezade A, Scholkopf B, Gomez-Rodriguez M. Enhancing human learning via spaced repetition optimization. *Proc Natl Acad Sci U S A*. 2019;116(10):3988–93. 10.1073/pnas.1815156116.
 16. Nadal C, Sas C, Doherty G. Technology acceptance in mobile health: Scoping review of definitions, models, and measurement. *J Med Internet Res*. 2020;22(7):e17256. 10.2196/17256.
 17. Li J, Yin K, Wang Y, Jiang X, Chen D. Effectiveness of generative artificial intelligence-based teaching versus traditional teaching methods in medical education: A meta-analysis of randomized controlled trials. *BMC Med Educ*. 2025;25(1):1175. 10.1186/s12909-025-07750-2.
 18. Geng EH, Powell BJ, Goss CW, Lewis CC, Sales AE, Kim B. When the parts are greater than the whole: How understanding mechanisms can advance implementation research. *Implement Sci*. 2025;20(1):22. 10.1186/s13012-025-01427-6.
 19. Shaham N, Chandra J, Kreiman G, Sompolinsky H. Stochastic consolidation of lifelong memory. *Sci Rep*. 2022;12(1):13107. 10.1038/s41598-022-16407-9.
 20. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, Chartash D. How does ChatGPT perform on the United States Medical Licensing Examination? The implications of large language models for medical education and knowledge assessment. *JMIR Med Educ*. 2023;9:e45312. 10.2196/45312.
 21. Chung P, Fong CT, Walters AM, Aghaeepour N, Yetisgen M, O'Reilly-Shah VN. Large language model capabilities in perioperative risk prediction and prognostication. *JAMA Surg*. 2024;159(8):928–37. 10.1001/jamasurg.2024.1621.
 22. Baumel A, Muench F, Edan S, Kane JM. Objective user engagement with mental health apps: Systematic search and panel-based usage analysis. *J Med Internet Res*. 2019;21(9):e14567. 10.2196/14567.
 23. Hopewell S, Chan AW, Collins GS, Hróbjartsson A, Moher D, Schulz KF, et al. CONSORT 2025 statement: updated guideline for reporting randomised trials. *BMJ*. 2025;389:e081123. 10.1136/bmj-2024-081123.

24. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK, SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nat Med. 2020;26(9):1364–74. 10.1038/s41591-020-1034-x.
25. Pintrich PR, Smith DAF, Garcia T, McKeachie WJ. A manual for the use of the Motivated Strategies for Learning Questionnaire (MSLQ). Ann Arbor: National Center for Research to Improve Postsecondary Teaching and Learning: University of Michigan; 1991.

Tables

Table 1. Baseline Characteristics of the Study Participants

Characteristic	Reactive response control (n = 60)	Proactive push AI (n = 60)
Age, years	26.88 (2.79)	27.32 (2.47)
Thyroid-related rotation months	5.85 (4.07)	6.37 (3.85)
Prior guideline learning times	2.40 (1.62)	2.68 (1.67)
Daily Feishu use, min	132.97 (62.54)	135.43 (56.29)
Baseline total score	16.98 (4.07)	16.60 (3.09)
Baseline guideline score	8.25 (2.12)	8.20 (1.94)
Baseline ADR score	8.73 (2.57)	8.40 (1.94)
Baseline self-efficacy total	33.88 (4.19)	35.00 (3.53)
Baseline UTAUT total	43.40 (5.02)	44.27 (4.62)
Male sex	35 (58.3%)	33 (55.0%)
PGY-1	22 (36.7%)	14 (23.3%)
PGY-2	19 (31.7%)	29 (48.3%)
PGY-3	19 (31.7%)	17 (28.3%)
High baseline knowledge stratum	19 (31.7%)	11 (18.3%)
Daily AI use	13 (21.7%)	15 (25.0%)

Table 2. Primary and Secondary Outcomes: Knowledge Scores at T3w and T7w

Outcome	Reactive response control (n = 60)	Proactive push AI (n = 60)	Adjusted effect	95% CI	p value	Effect size (d)
T3w total score	20.44 (3.97)	20.15 (3.79)	0.65	-0.04 to 1.34	0.065	-0.08
T3w guideline score	10.56 (2.52)	9.98 (2.46)	-0.50	-1.28 to 0.27	0.203	-0.23
T3w ADR score	9.88 (2.32)	10.16 (2.54)	0.89	0.13 to 1.65	0.022	0.12
T7w total score	18.67 (4.07)	19.58 (4.16)	1.49	0.36 to 2.62	0.010	0.22
T7w original score	9.21 (2.37)	10.27 (2.38)	1.01	0.17 to 1.85	0.019	0.44
T7w transfer score	9.45 (2.47)	9.31 (2.54)	0.35	-0.53 to 1.22	0.438	-0.06
Change from baseline to T3w	2.70 (1.54)	3.35 (1.83)	0.65	-0.04 to 1.34	0.065	0.38
Change from baseline to T7w	0.98 (2.37)	2.42 (2.83)	1.49	0.36 to 2.62	0.010	0.55
Change from T3w to T7w	-1.64 (1.54)	-0.85 (1.62)	0.85	0.16 to 1.55	0.016	0.50
Self-efficacy change	4.02 (3.43)	6.89 (3.18)	2.83	1.51 to 4.16	<0.001	0.87

Table 3. Engagement, Self-Efficacy, and User Experience

Measure	Reactive response control (n = 60)	Proactive push AI (n = 60)	Adjusted effect	95% CI	p value	FDR-adjusted p value
Login days over 3 weeks	11.73 (4.12)	20.22 (3.09)	8.56	7.26 to 9.87	<0.001	<0.001
Total micro-quiz completions	19.88 (7.34)	53.57 (10.75)	34.11	30.75 to 37.47	<0.001	<0.001
Total proactive questions	13.45 (4.88)	23.27 (4.96)	10.30	8.75 to 11.85	<0.001	<0.001
Post UTAUT mean	3.01 (0.34)	3.47 (0.37)	0.46	0.31 to 0.61	<0.001	<0.001
Satisfaction mean	3.05 (0.49)	3.34 (0.53)	0.25	0.05 to 0.46	0.015	0.017
Personalization mean	2.65 (0.44)	3.32 (0.43)	0.69	0.52 to 0.86	<0.001	<0.001
NASA-TLX mean	50.69 (8.13)	48.20 (6.47)	-2.68	-5.77 to 0.41	0.089	0.089
Push response rate (intervention only)	—	0.52 (0.06)	—	—	—	—

Figures

Proactive Chat-Native AI Tutoring versus Reactive Chatbot for Thyroid Cancer Guideline Learning

A Multi-Center Feishu-Embedded OpenClaw Randomized Controlled Trial

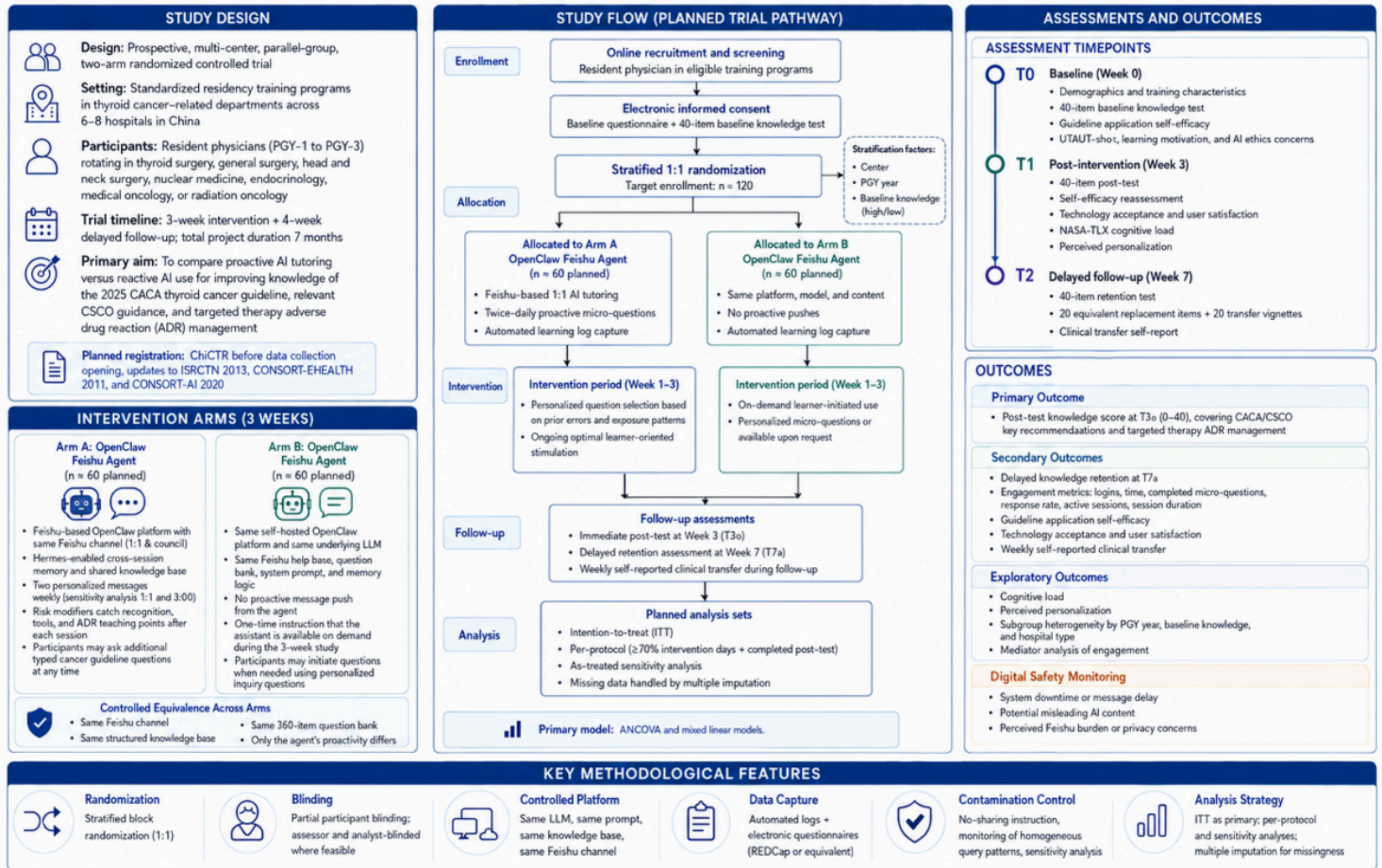


Figure 1

Schematic Overview of the Experimental Design

This figure summarizes the study design, randomization, intervention arms, and assessment schedule. Eligible resident physicians were randomized 1:1 to the proactive push AI-assisted learning group or the reactive chatbot control group. Both groups used the same Feishu-embedded OpenClaw platform, knowledge base, and large language model framework. Assessments were conducted at baseline, Week 3 (T3w), and Week 7 (T7w).

Figure 2. Main and key secondary outcome effects

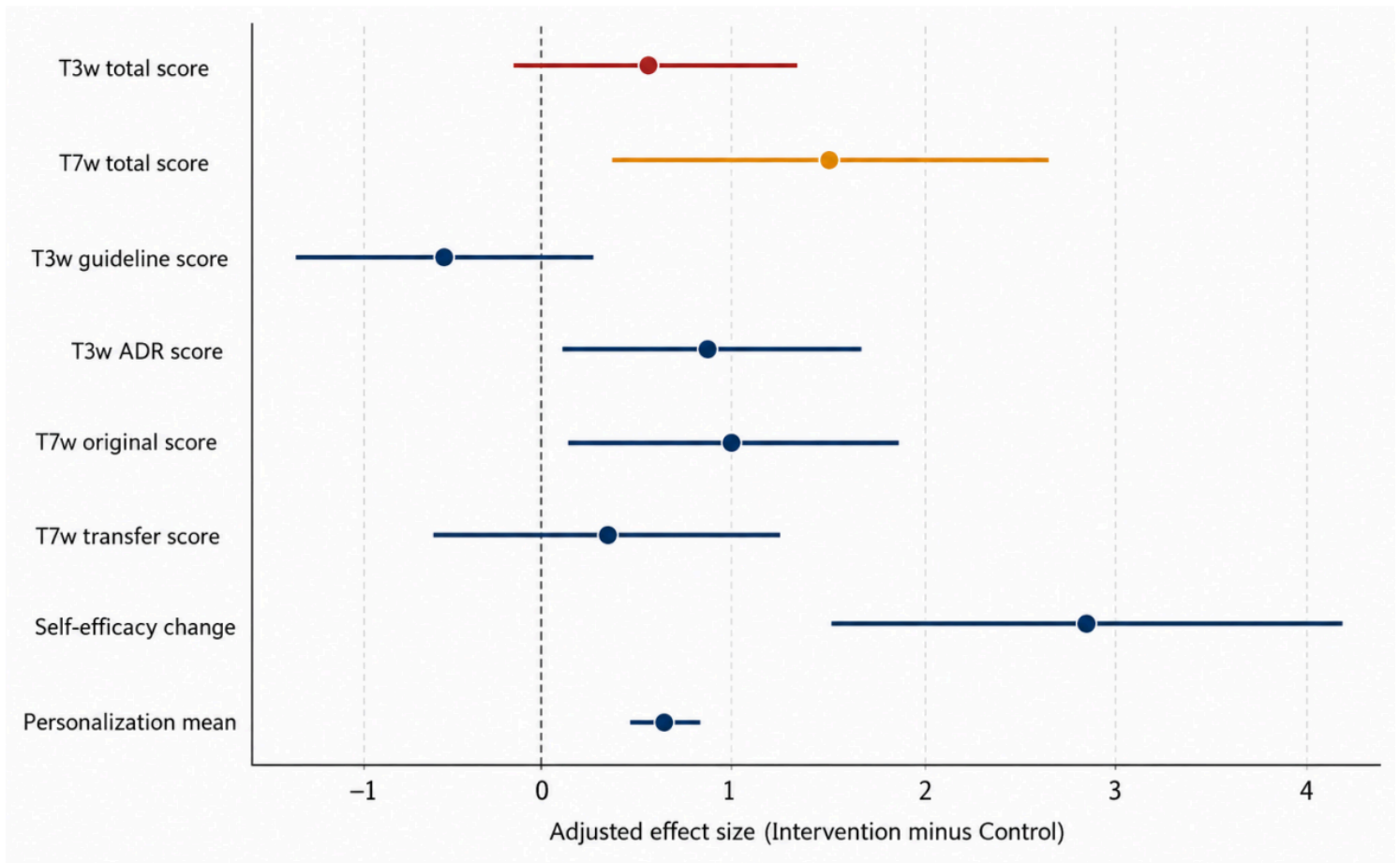


Figure 2

Main and key secondary outcome effects

(1) T3w total knowledge score; (2) T3w guideline module score; (3) T3w ADR module score; (4) T7w total retention score; (5) T7w original-item retention score; (6) T7w transfer-item score; (7) change from baseline to T3w; (8) change from baseline to T7w; (9) change from T3w to T7w; and (10) self-efficacy change score. Positive values indicate superiority of the proactive push AI-assisted learning intervention.

Figure 3. Knowledge trajectory over time

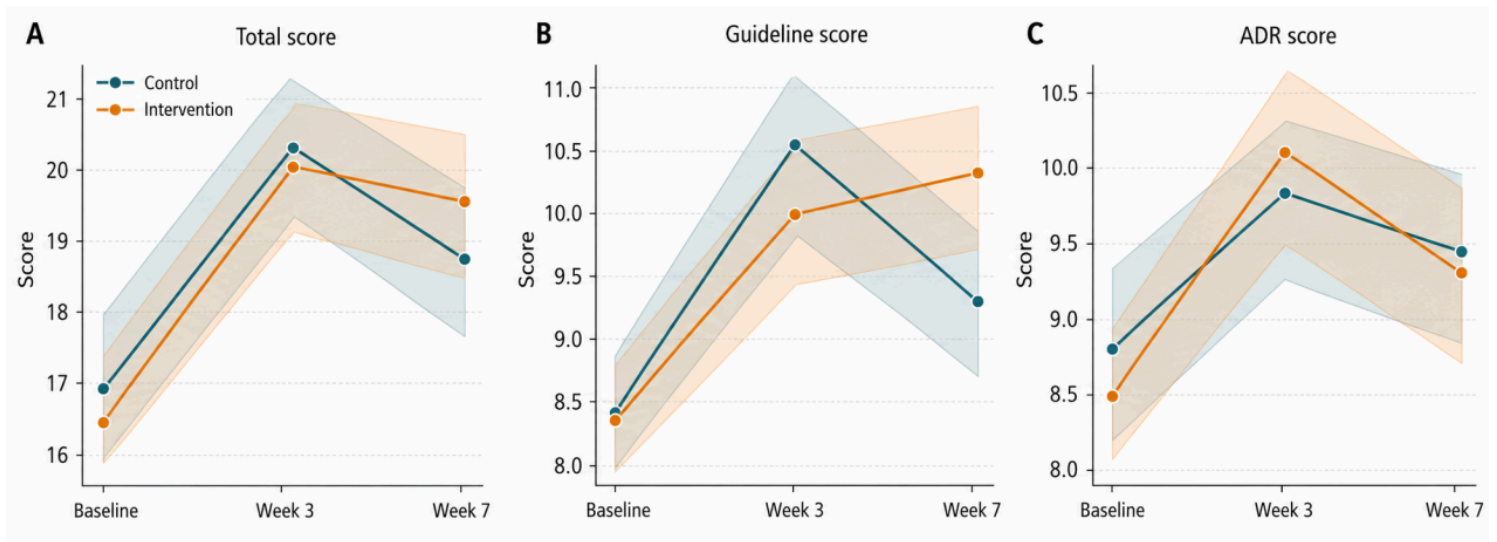


Figure 3

Knowledge trajectory over time

Longitudinal changes in knowledge scores from baseline to Week 7.

(A) Total knowledge score.

(B) Guideline module score.

(C) ADR module score.

Lines represent group mean scores, and shaded areas indicate 95% confidence intervals

Figure 4. Engagement mechanism panels

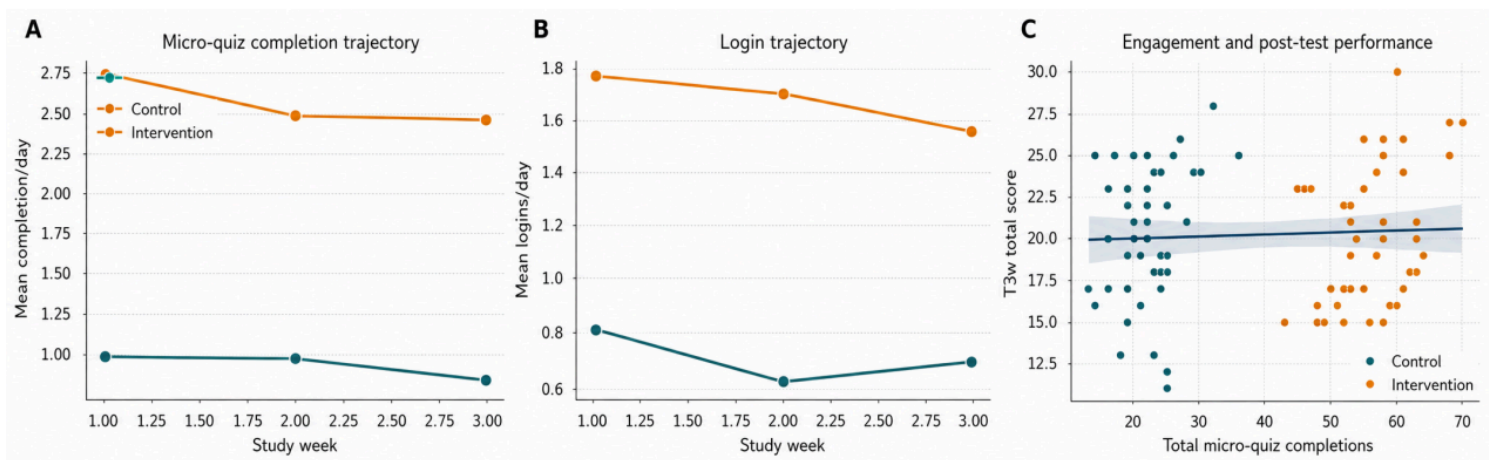


Figure 4

Engagement mechanism panels

Engagement trajectories and their association with learning outcomes.

(A) Mean daily micro-quiz completion across study weeks.

(B) Mean daily login frequency across study weeks.

(C) Association between total micro-quiz completions and T3w post-test performance.

Supplementary Files

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

- [FigureS1MissingnessHeatmap.pdf](#)
- [FigureS2ScoreChangeDistributions.pdf](#)
- [FigureS3DailyEngagementTrajectory.pdf](#)
- [FigureS4SubgroupForest.pdf](#)
- [FigureS5PRODistributions.pdf](#)
- [FigureS6MediationDiagram.pdf](#)
- [FigureS7SpaghettiPlot.pdf](#)
- [CONSORT2025ChecklistCompleted.pdf](#)
- [SupplementaryFigureLegends.docx](#)