SAGE: A Guardrailed LLM Tutoring System with Scaffolded Self-Regulated Learning for Adult Education

Kwan Hui Lim
Singapore University of Social Sciences
School of Business
Singapore, Singapore
khlim@suss.edu.sg

Jing Ren
Singapore University of Social Sciences
School of Business
Singapore, Singapore
jingren@suss.edu.sg

Yimiao Zhang
Singapore University of Social Sciences
School of Business
Singapore, Singapore
yzhang@suss.edu.sg

Stefanie Chye
Singapore University of Social Sciences
Teaching & Learning Centre
Singapore, Singapore
stefaniechyeyl@suss.edu.sg

Amy Wong
Singapore University of Social Sciences
School of Business
Singapore, Singapore
amywongom@suss.edu.sg

Abstract

We present SAGE (Scaffolded, Audited, Guided Education), a deployable guardrailed LLM tutoring system for adult learners. Left unrestricted, chat assistants often hand over the answer, which can improve in-the-moment performance but may erode the self-regulation that supports durable learning. SAGE wraps a commercial LLM in interface-level guardrails that require learners to engage before the model answers: a lightweight intent classifier routes each turn to one of three self-regulated-learning scaffolds (write-your-approach-first, paraphrase-before-proceeding, attempt-before-reveal), so a request to "just give me the answer" is held behind a guided step rather than answered outright. As a secondary safeguard, an online counterfactual fairness audit checks whether the guardrailed tutor responds differently across age band, language dominance, prior education, and digital fluency, while an agentic rectifier regenerates any flagged turn; a pseudonymous interaction log supports per-cohort analytics. The demonstration runs the same business-education prompt through an unrestricted LLM and through SAGE side by side, allowing an attendee to watch the scaffold gate hold the answer, then exercise the counterfactual swap, inspect the per-attribute audit panel, and trigger the rectifier. The system is open-source and available at https://github.com/kwanhui/public-llmLearnFair-CIKM2026 to allow other groups to deploy and audit their own scaffolds.

CCS Concepts

• **Computing methodologies** → Intelligent agents; **Natural language processing**; • **Social and professional topics** → **Computing education**; Adult education; • **Human-centered computing** → *Interactive systems and tools.*



This work is licensed under a Creative Commons Attribution 4.0 International License.
CIKM '26, Rome, Italy
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2539-5/2026/11
https://doi.org/10.1145/3799682.3840260

Keywords

LLM tutoring, algorithmic fairness, self-regulated learning, guardrails, adult education

ACM Reference Format:
Kwan Hui Lim, Jing Ren, Yimiao Zhang, Stefanie Chye, and Amy Wong. 2026. SAGE: A Guardrailed LLM Tutoring System with Scaffolded Self-Regulated Learning for Adult Education. In *Proceedings of the 35th ACM International Conference on Information and Knowledge Management (CIKM '26), November 07–11, 2026, Rome, Italy*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3799682.3840260

1 Introduction

Singapore's Continuing Education and Training (CET) sector now serves working adults whose study increasingly takes place through commercial chat assistants outside classroom guardrails. Recent randomised evidence shows that unrestricted use of frontier LLMs raises in-the-moment task scores yet lowers post-access grades, while a hint-based guardrail wrapped around the same model preserves the gain [5]. Lee et al. [14] report similar trust-driven offloading among working knowledge workers, while Adarkwah [1] describes the resulting GenAI divide among adult learners.

Our aim is to preserve the self-regulation that unrestricted answer-giving tends to erode. Rather than allowing the model to answer on demand, SAGE inserts guided steps that require learners to plan, attempt, or paraphrase before the tutor responds. Equity is a secondary concern: almost all GenAI-in-education trials report only group means [3], leaving it unclear whether such a guardrail helps every learner equally, and SAGE therefore audits its responses across protected demographic attributes.

This paper presents SAGE, a deployable classroom artefact built around this guardrail, and contributes: (i) three interface-level self-regulated-learning (SRL) scaffolds (`WriteApproachFirst`, `ParaphraseBeforeProceed`, `AttemptBeforeReveal`) that make engagement a precondition for the model's answer, grounded in established theory [6, 28]; (ii) a Zod-schema LLM-as-classifier that labels each learner turn into a four-way intent set and routes it to the appropriate scaffold; (iii) as a secondary safeguard,

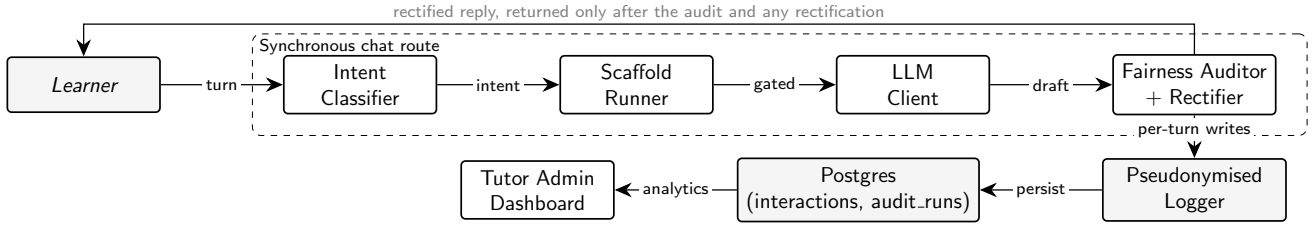


Figure 1: Per-turn data flow. The reply passes through the auditor and rectifier before it returns to the learner, so a flagged turn is regenerated with a fairness-correction prefix ahead of delivery. The dashed envelope marks the synchronous chat route; the pseudonymised logger records only per-turn metadata and hashed audit rows, but not raw learner text.

an online counterfactual fairness audit on tutor-generated text across age band, language dominance, prior education, and digital fluency, with a rectifier loop that regenerates flagged turns before they reach the learner [13, 15]; and (iv) a pseudonymous HMAC-SHA256 interaction log with documented salt and retention policies. The reference implementation (https://github.com/kwanhui/public-llmLearnFair-CIKM2026) is deployable as a Next.js application with a hosted demo linked from the repository, and a tutor admin flow for configuring scaffolds, audit thresholds, and student invite URLs per cohort. The system targets CET and university educators who want a response-level guardrail layer for their classrooms, and information-systems and learning-analytics researchers who need a reference implementation to study scaffolding policies, audit thresholds, and intent labels. This paper presents the architecture and three walkthrough scenarios that exercise the scaffolds, audit, and rectifier end to end.

2 System Architecture

Figure 1 shows the four components of SAGE that process each learner turn. A learner request enters the chat route, passes through the intent classifier, scaffold runner, and LLM client, and is then audited, rectified if flagged, and logged before the reply reaches the learner. For our demo, the runtime is a Next.js 16 application deployed on Vercel with a Postgres-backed log.

2.1 Intent Classification

Each learner turn is labelled `direct_answer`, `conceptual`, `clarification`, or `meta` via an LLM-as-classifier call with a Zod-validated structured-output schema, removing the need for fine-tuning and keeping the label set stable across model upgrades. The label determines which scaffold the runner activates and is logged with each turn for downstream analytics. The four-label schema and production multi-turn pattern follow Liu et al. [18]; the classifier runs on the same low-cost LLM backend as the chat reply and audit probes (`gpt-4o-mini`, temperature 0.2), keeping per-turn overhead low.

2.2 Scaffold Enforcement

Three state machines, each gated by the classified intent, act as preconditions on the LLM call and keep the decision deterministic, similar to our agentic tour replanner [16]. `WriteApproachFirst` fires on `direct_answer` and holds the response until the learner contributes at least 160 characters of their own approach across several distinct words, rejecting padded filler or repeated single words used to clear the count, operationalising Zimmerman's forethought phase [28]. `ParaphraseBeforeProceed` fires on `direct_answer` or `conceptual` turns after the first tutor reply and requires a Jaccard token overlap of at least 0.3 between the learner's next message and the previous tutor reply, bringing performance monitoring under explicit interface control [28]. `AttemptBeforeReveal` fires on `conceptual` and surfaces a guess-first prompt before the explanation, in line with the desirable-difficulties literature [6] and the enforced-move pattern from MetaTutor [2].

2.3 Counterfactual Fairness Audit

After each tutor turn, the auditor runs one counterfactual probe per configured attribute value across four protected attributes: age band, language dominance, prior education, and digital fluency. The probe grid is a deployment-level audit configuration file: the business-education file used in this demonstration lists four attributes with three values each, so the audit issues twelve probe calls per turn, while another deployment can ship a different grid. A probe prefixes the system prompt with a marker sentence of the form "The learner describes themselves as: {value}.", drawn from that file's per-attribute value set, calls the LLM with the same learner prompt, and scores the response on two text metrics: `explanation_depth` (a length-led elaboration score that sums the response's word count with its list-marker, definition- or example-cue, and follow-up-question counts, then maps the total to the unit interval through a smooth saturating transform $1 - e^{-x}$, so longer answers spread across the range rather than all collapsing to the maximum) and `jargon_density` (the share of tokens that match the cohort's domain glossary, a tutor-curated list shipped with each scenario). A probe is flagged when the within-attribute range on either metric crosses a per-cohort threshold (the business-education cohort that drives the demo uses 0.20 and 0.10 respectively, pending empirical calibration), invoking an agentic rectifier loop whose retry-with-correction pattern is adapted from RALLM-POI [15] and regenerates the response with a fairness-correction prefix before it is returned, capped at a per-cohort retry budget (one regeneration by default) so that a persistently flagged turn does not stall the learner. The probe-and-rectify design follows the counterfactual-fairness frame of Kusner et al. [13] and ports the deployment-time audit pattern of Maughan et al. [19] from classifier outputs to generated tutor text. Both metrics are engineering proxies rather than

validated constructs: they score properties of the response text, not what a learner understood, and we have not calibrated them against learner-perceived fairness, learning outcomes, or group-level education measures such as ABROCA [11]. `jargon_density` also inherits the coverage of its glossary: a term the tutor admin omits is invisible to the audit. The audit aims to demonstrate that a runtime equity check can run per turn and drive a rectifier, rather than that these two numbers are the right measures of pedagogical equity.

2.4 Privacy-Preserving Logging

The interaction logger writes one row per turn to Postgres, keyed by an HMAC-SHA256 hash of the learner identifier using a server-side salt that is never persisted alongside the row. Learner message text is never written to the log; each per-turn row holds only the intent label, scaffold state, and token counts, while the audit log stores a non-reversible hash of each probed prompt rather than its text, allowing an analyst to compute aggregate metrics without accessing raw text. The 90-day retention policy and salt-rotation policy are configured per deployment, with a hard refusal to hash when the salt is unset, and a tutor admin can erase an individual learner on request, following the per-turn pseudonymisation pattern explored in our prior privacy work [20]. In a small cohort, the combination of per-turn fields (intent label, scaffold state, timestamp) can still single out a learner. Hence pseudonymisation is not anonymisation but our combined use of short retention, salt rotation, and per-learner erasure limits this potential exposure.

2.5 Deployment Characteristics

The hosted demo runs the same code path as the cohort-aware student route, with side-by-side comparison forced on and the database write path short-circuited; a one-click deploy button in the README provisions a private instance. The intent classifier adds one LLM call ahead of the tutor reply; the scaffold runner uses local state logic and adds none. After the reply is generated, the audit issues all its probes as one concurrent batch, so a guarded turn waits on a single probe round whose length is determined by the slowest probe, rather than twelve sequential calls. A flagged turn adds at most the capped regeneration to that round before the response is shown. Per-turn audit cost scales with the size of the probe grid, and an operator can lower it by shipping a smaller grid or pointing the audit backend to a cheaper model than the chat model, which is configured separately.

3 Demonstration

The hosted demo runs the full pipeline against a public sandbox cohort, and a short walkthrough video is available at https://youtu.be/4FyYClwt7Ic. The visitor's first view of /demo provides a scenario picker, the SRL scaffold sidebar, side-by-side comparison columns, counterfactual swap controls, and a pseudonymisation note.

3.1 Scenario 1: "Just Write It For Me"

The viewer types *"Just write the executive summary of my churn analysis for me"* into both columns. The unrestricted column writes the entire summary and hands it over, effectively completing the learner's task for them. The guardrailed column classifies the turn as `direct_answer`, fires `WriteApproachFirst`, and renders an inline scaffold prompt asking the learner to draft their own approach before any answer is shown. Once the learner contributes their own draft, the tutor helps refine it rather than replacing it, and the audit badge resolves to a per-attribute fairness verdict beside the message. Figure 2 (left) shows this side-by-side interaction.

3.2 Scenario 2: Liquidity Ratio Drop

The second scenario asks *"Our company's current ratio dropped from 1.8 to 1.1 over a year. What does that mean for liquidity?"*. This turn is classified as `conceptual` and triggers `AttemptBeforeReveal`. The tutor first surfaces a guess-first gate ("Take a quick guess first, even if you are unsure: what do you think the answer might be, and why?") and only releases the full explanation once the learner attempts an answer. If the audit detects a depth gap on any protected attribute above the configured threshold, the rectifier loop regenerates the explanation with a fairness-correction prefix before it is shown, and a `rectified (n)` chip on the audit badge marks the corrected turn and its retry count. The viewer can force a flag by widening the gap, for example by pinning one swap to secondary education and another to postgraduate, and watch the rectifier pull the explanation back within the threshold.

3.3 Live Fairness Audit Panel

A swap control sits above the chat input, with one dropdown per protected attribute populated from the cohort configuration. The viewer changes one dropdown and clicks Re-run last turn; the counterfactual variant appears as a card below the original assistant message, making any divergence visible side by side rather than only as a metric. The inline audit badge expands on click to a per-probe panel with explanation-depth and jargon-density bars and a flag indicator for each crossed-threshold probe.

A user with tutor access can switch to /admin/cohorts/<id>/analytics to view the cohort dashboard shown in Figure 2 (right), which provides per-scaffold pass rates, divergence-by-attribute over time, and a CSV export of the pseudonymised log.

4 Related Work

LLMs in education: learning harms and the equity gap. Randomised trials of generative AI in learning report mixed results. Bastani et al. [5] found that students given unrestricted GPT-4 outscored a no-AI control during practice but underperformed it on a closed-book post-test, while a hint-based guardrail variant preserved the gains. Barcaui [4] reports a similar retention loss, while Fan et al. [9] find fewer metacognitive moves and no transfer gain among ChatGPT users; Yan et al. [27] describe this as a performance-versus-learning gap. Xu et al. [26] show that metacognitive prompts partly close this gap, while Lee et al. [14] report similar trust-driven offloading among knowledge workers. Adarkwah [1] identifies a GenAI divide for adult learners that these studies do not address: all report group means, but none audit subgroup outcomes by age, language, prior education, or digital fluency.

Self-regulated learning and interface-level scaffolds. Our scaffolds draw on established theory. Zimmerman [28] models self-regulated learning as a forethought–performance–reflection cycle,

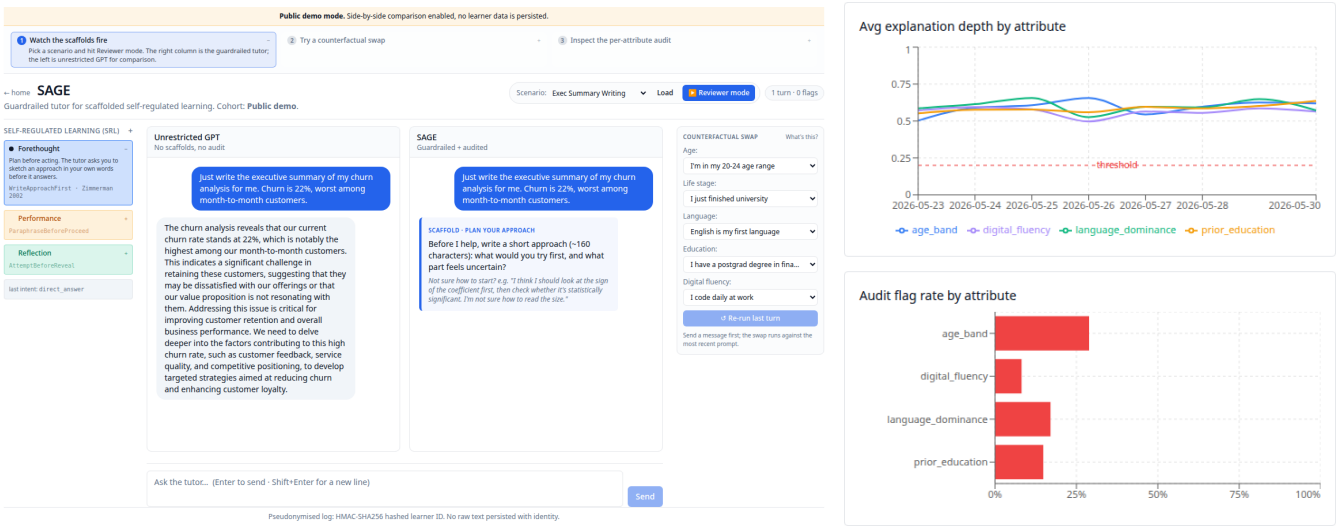


Figure 2: Left: the scaffold gate firing on a direct-answer request; unrestricted GPT writes the whole executive summary while SAGE holds it behind a plan-first scaffold prompt. Right: the per-cohort fairness analytics (synthetic seeded data): average explanation depth by protected attribute against the configured threshold, audit flag rate by attribute, scaffold pass rate, and a CSV export; learner identifiers are HMAC-SHA256 hashes and no raw text is shown.

while Bjork and Bjork [6] argue that desirable difficulties can depress immediate performance while improving retention. Cronbach and Snow [8] formalised aptitude–treatment interactions, highlighting that a given form of support may benefit some learners more than others, which Sana and Fenesi [22] revisit for adaptive instruction. MetaTutor [2] showed that a tutoring system can enforce such moves directly through the interface. We extend this approach to free-form LLM chat: WriteApproachFirst makes forethought a precondition for the first answer, ParaphraseBeforeProceed makes monitoring a precondition for the next turn, and AttemptBeforeReveal makes productive struggle a precondition for direct solutions.

Counterfactual fairness for generated tutor text. Kusner et al. [13] formalise fairness as invariance under counterfactual swaps of a protected attribute, with Hardt et al. [12] providing the group-level baseline. Gallegos et al. [10] show that LLM prompt-perturbation audits are almost always conducted offline over fixed templates, while Maughan et al. [19] port counterfactual auditing to deployment but only for classifier outputs, without intervening on the model. In education, fairness work remains at the classifier or course-outcome level: Gardner et al. [11] introduce ABROCA for MOOC predictors, and Baker and Hawn [3] survey the demographic-gap evidence behind our four attributes; lab work has studied fairness–accuracy trade-offs [17] and per-turn guarantees on user history [20]. We extend this approach to LLM-generated explanations: for each turn, we swap demographic markers in the system prompt and trigger a rectifier when divergence in explanation depth or jargon density exceeds the configured threshold.

Positioning against existing systems. The closest systems are LLM tutors such as ChatTutor [7] and Ruffle&Riley [23], which chain LLMs for content delivery without a runtime equity check, and guardrail toolkits such as NeMo Guardrails [21] that provide general-purpose rails without a pedagogical model. A concurrent multi-agent framework that shares our system's name [24] plans and orchestrates pedagogical dialogue across agents, but does not audit its own output for equity, and FairTutor [25] routes between models on cost and pedagogical quality rather than intervening on the generated text of a single turn. SAGE differs from these toolkits in architecture and scope, rather than a measured comparison. Where Bastani et al. [5] show that a single hint-style guardrail preserves the learning gain, SAGE generalises this approach into a deployable system: it enforces preconditions across all three phases of Zimmerman's cycle, routes them based on classified intent, and adds a per-turn online equity audit that no hint-based guardrail or existing tutor performs.

5 Conclusion

In this paper, we presented and released SAGE, a deployable LLM tutoring system that uses interface-level scaffolds to support self-regulated learning rather than handing over answers. The system integrates three SRL scaffolds, intent-based routing, an online fairness audit with rectification, and pseudonymous interaction logging. We demonstrated the system through scenarios that show how the scaffolds, fairness audit, and rectifier operate during learner–tutor interactions. As this demonstration focuses on the system implementation, we have not evaluated its effects on real learners, benchmarked it against existing guardrail toolkits, or validated the audit metrics as measures of learning outcomes. As part of future work, we will evaluate intent classification and scaffold completion, measure post-test retention against an unguarded control, and examine how the audit metrics relate to learner-rated explanation quality before recalibrating the audit thresholds.

GenAI Usage Disclosure

Generative AI tools were used to improve the prose, writing, and structure of this paper, as well as to refactor and document code for the system and repository. The authors reviewed all content and take full responsibility for it. Separately, commercial large language models are used as the backbone of the system described in this work, while SAGE provides the guardrails, fairness audit, and interface around this model usage for the research project.

References

[1] Michael Agyemang Adarkwah. 2024. GenAI-Infused Adult Learning in the Digital Era: A Conceptual Framework for Higher Education. *Adult Learning* (2024). doi:10.1177/10451595241271161

[2] Roger Azevedo, François Bouchet, Melissa Duffy, Jason Harley, Michelle Taub, Gregory Trevors, Elizabeth Cloude, Daryn Dever, Megan Wiedbusch, Franz Wortha, and Rebeca Cerezo. 2022. Lessons Learned and Future Directions of MetaTutor: Leveraging Multichannel Data to Scaffold Self-Regulated Learning With an Intelligent Tutoring System. *Frontiers in Psychology* 13 (2022), 813632. doi:10.3389/fpsyg.2022.813632

[3] Ryan S. Baker and Aaron Hawn. 2022. Algorithmic Bias in Education. *International Journal of Artificial Intelligence in Education (IJAIED)* (2022). doi:10.1007/s40593-021-00285-9

[4] André Barcaui. 2025. ChatGPT as a Cognitive Crutch: Evidence from a Randomized Controlled Trial on Knowledge Retention. *Social Sciences & Humanities Open* (2025). doi:10.1016/j.ssaho.2025.102287

[5] Hamsa Bastani, Osbert Bastani, Alp Sungu, Haosen Ge, Özge Kabakçı, and Rei Mariman. 2025. Generative AI Without Guardrails Can Harm Learning: Evidence from High School Mathematics. *Proceedings of the National Academy of Sciences (PNAS)* (2025). doi:10.1073/pnas.2422633122

[6] Elizabeth Ligon Bjork and Robert A. Bjork. 2011. Making Things Hard on Yourself, But in a Good Way: Creating Desirable Difficulties to Enhance Learning. In *Psychology and the Real World: Essays Illustrating Fundamental Contributions to Society*, Morton Ann Gernsbacher, Richard W. Pew, Leaetta M. Hough, and James R. Pomerantz (Eds.). Worth Publishers.

[7] Yulin Chen, Ning Ding, Hai-Tao Zheng, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2024. Empowering Private Tutoring by Chaining Large Language Models. In *Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM)*. 354–364. doi:10.1145/3627673.3679665

[8] Lee J. Cronbach and Richard E. Snow. 1977. *Aptitudes and Instructional Methods: A Handbook for Research on Interactions*. Irvington Publishers.

[9] Yizhou Fan, Luzhen Tang, Huixiao Le, Kejie Shen, Shufang Tan, Yueying Zhao, Yuan Shen, Xinyu Li, and Dragan Gašević. 2025. Beware of Metacognitive Laziness: Effects of Generative Artificial Intelligence on Learning Motivation, Processes, and Performance. *British Journal of Educational Technology* 56 (2025), 489–530. doi:10.1111/bjet.13544

[10] Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and Fairness in Large Language Models: A Survey. *Computational Linguistics* 50, 3 (2024), 1097–1179. doi:10.1162/coli_a_00524

[11] Josh Gardner, Christopher Brooks, and Ryan Baker. 2019. Evaluating the Fairness of Predictive Student Models Through Slicing Analysis. In *Proceedings of the International Conference on Learning Analytics and Knowledge (LAK)*. doi:10.1145/3303772.3303791

[12] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of Opportunity in Supervised Learning. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*.

[13] Matt J. Kusner, Joshua R. Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual Fairness. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*.

[14] Hao-Ping Lee, Advait Sarkar, Lev Tankelevitch, Ian Drosos, Sean Rintel, Richard Banks, and Nicholas Wilson. 2025. The Impact of Generative AI on Critical Thinking: Self-Reported Reductions in Cognitive Effort and Confidence Effects from a Survey of Knowledge Workers. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI)*. doi:10.1145/3706598.3713778

[15] Kunrong Li and Kwan Hui Lim. 2025. RALLM-POI: Retrieval-Augmented LLM for Zero-Shot Next POI Recommendation with Geographical Reranking. In *Proceedings of the Pacific Rim International Conference on Artificial Intelligence (PRICAI)*. 210–218.

[16] Kwan Hui Lim, Aldy Gunawan, and Carmen Kar Hang Lee. 2026. Real-Time Adaptive Tour Recommendation with Agentic AI. In *Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM)*. doi:10.1145/3799682.3840259

[17] Junhua Liu, Roy Ka-Wei Lee, and Kwan Hui Lim. 2025. Understanding Fairness-Accuracy Trade-offs in Machine Learning Models: Does Promoting Fairness Undermine Performance?. In *Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. 137–145.

[18] Junhua Liu, Yong Keat Tan, Bin Fu, and Kwan Hui Lim. 2026. Balancing Accuracy and Efficiency in Multi-Turn Intent Classification for LLM-Powered Dialog Systems in Production. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 40073–40082.

[19] Krystal Maughan, Ivoline C. Ngong, and Joseph P. Near. 2022. Prediction Sensitivity: Continual Audit of Counterfactual Fairness in Deployed Classifiers. arXiv preprint. arXiv:2202.04504

[20] Wenchuan Mu and Kwan Hui Lim. 2025. Bayesian Privacy Guarantee for User History in Sequential Recommendation Using Randomised Response. In *Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM)*. 5041–5046.

[21] Traian Rebedea, Razvan Dinu, Makesh Narsimhan Sreedhar, Christopher Parisien, and Jonathan Cohen. 2023. NeMo Guardrails: A Toolkit for Controllable and Safe LLM Applications with Programmable Rails. In *Proceedings of EMNLP — System Demonstrations*.

[22] Faria Sana and Barbara Fenesi. 2025. Working Memory and Instructional Fit: Reintroducing Aptitude–Treatment Interaction in Education Research. *Behavioral Sciences* 15, 6 (2025), 765. doi:10.3390/bs15060765

[23] Robin Schmucker, Meng Xia, Amos Azaria, and Tom Mitchell. 2024. Ruffle&Riley: Insights from Designing and Evaluating a Large Language Model-Based Conversational Tutoring System. In *Proceedings of the International Conference on Artificial Intelligence in Education (AIED) (Lecture Notes in Computer Science, vol. 14829)*. Springer. doi:10.1007/978-3-031-64302-6_6

[24] Van-Khanh Tran, Van-Khai Dang, and Duc-Huy Nguyen. 2026. SAGE: A Compositional Multi-Agent LLM Framework with Pedagogical Reasoning for Structured Collaborative Problem Solving. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 40872–40879. doi:10.1609/aaai.v40i48.42120

[25] Qingyang Xu. 2026. FairTutor: Equity-Aware Pedagogical LLM Routing for Budget-Constrained AI Tutoring. arXiv preprint. arXiv:2606.20713

[26] Xiaoqing Xu, Lifang Qiao, Nuo Cheng, Hongxia Liu, and Wei Zhao. 2025. Enhancing Self-Regulated Learning and Learning Experience in Generative AI Environments: The Critical Role of Metacognitive Support. *British Journal of Educational Technology* 56, 5 (2025), 1842–1863. doi:10.1111/bjet.13599

[27] Lixiang Yan, Samuel Greiff, Jason M. Lodge, and Dragan Gašević. 2025. Distinguishing Performance Gains from Learning When Using Generative AI. *Nature Reviews Psychology* (2025). doi:10.1038/s44159-025-00467-5

[28] Barry J. Zimmerman. 2002. Becoming a Self-Regulated Learner: An Overview. *Theory into Practice* 41, 2 (2002), 64–70. doi:10.1207/s15430421tip4102_2