

Article

An Empirical Study on Human-AI Interaction Patterns and System Design in Open University English Education

Jun Xiao^{1,*}
¹ Wuhan Vocational College of Software & Engineering, Wuhan, China

* Correspondence: Jun Xiao, Wuhan Vocational College of Software & Engineering, Wuhan, China

Abstract: The integration of Generative AI (GenAI) into educational software is reshaping human-computer interaction (HCI), yet most AI-assisted teaching systems function as automation tools rather than genuine collaborative partners. This paper presents the design, optimization, and evaluation of the Collaborative AI Teaching Assistant (CAITA), specifically architected for open university English education. Through a mixed-methods user study with 108 teachers, we identify key limitations of current systems. We then propose a novel architecture integrating four core modules: an Interactive Co-Design Interface powered by a Retrieval-Augmented Generation (RAG) engine; a Data-Driven Analytics Dashboard with a multi-dimensional learner performance model; an Ethical AI Stewardship Module with bias detection; and an Affective Computing Module using sentiment analysis. Our technical contributions include a Dynamic Prompt Optimization (DPO) algorithm that iteratively refines outputs via user feedback, and a Collaborative Filtering-based Recommendation (CFR) mechanism for pedagogical resources. Evaluations (technical benchmarks + A/B test with 60 teachers) show a 25.6% improvement in user interaction maturity and 21.3% increase in task completion efficiency over a baseline. This study provides a robust technical framework for next-generation human-AI collaborative educational software.

Keywords: Human-Computer Interaction; AI-Assisted Teaching Systems; Generative AI; Retrieval-Augmented Generation; Intelligent Tutoring Systems; Recommender Systems; Affective Computing

1. Introduction

The rapid advancement of Generative AI (GenAI), particularly Large Language Models (LLMs) such as GPT-4, has catalyzed a paradigm shift in the design of educational software systems [1]. In open university English education, AI is poised to revolutionize pedagogical practices by offering dynamic content generation, adaptive feedback, and real-time analytics. However, a critical challenge lies in moving beyond the "tool-using" paradigm, where AI merely automates routine tasks, towards a "collaborative partnership" where AI and teachers co-evolve in a symbiotic relationship [2].

From a computer science perspective, the core challenge is not merely integrating an LLM API, but designing an intelligent system architecture that facilitates effective human-AI interaction. Existing systems often suffer from several technical limitations: (1) lack of context-awareness in content generation, leading to irrelevant or generic suggestions; (2) opaque decision-making processes, hindering user trust and system transparency; (3) rudimentary data analytics that fail to provide actionable pedagogical insights; and (4) absence of affective mechanisms to support user confidence and reduce interaction anxiety.

We address these gaps by presenting CAITA, a novel AI-assisted teaching system grounded in a systematic user-needs analysis. Our primary contributions are:

System Architecture: A modular, microservices-based architecture integrating an LLM with four specialised technical modules.

Received: 18 August 2026

Revised: 25 September 2026

Accepted: 10 October 2026

Published: 13 October 2026



Copyright: © 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Algorithmic Innovations: A Dynamic Prompt Optimization (DPO) algorithm using reinforcement learning from user feedback (RLUF), and a Collaborative Filtering-based Recommendation (CFR) mechanism with hybrid user-/item-based filtering to overcome cold-start.

Technical Evaluation: Comprehensive performance evaluation (response time, BLEU, precision/recall) and a controlled A/B user study, validating our solutions in enhancing interaction maturity and efficiency.

The paper is organised as follows: Section 2 reviews related work; Section 3 describes the empirical user study; Section 4 details system architecture; Section 5 presents algorithms; Section 6 covers experiments and results; Section 7 concludes and discusses future work.

2. Literature Review

2.1. Human-Computer Interaction (HCI) in Educational Systems

The evolution of HCI in education has transitioned from command-line interfaces to graphical user interfaces (GUIs), and now to intelligent user interfaces (IUIs) [3]. Traditional frameworks like TAM and TPACK focused on technology as a tool [4]. However, with the emergence of GenAI as a proactive agent, new interaction paradigms are required. Recent HCI research emphasizes a "relational" ontology where AI acts as a "quasi-other" [5]. This demands system-level support for transparent, explainable, and adaptive interactions.

2.2. AI-Assisted Educational Systems and Their Limitations

The integration of AI into Learning Management Systems (LMS) and Intelligent Tutoring Systems (ITS) has been extensively studied [6]. Systems utilizing LLMs for language learning have shown promise in generating exercises and providing feedback [7]. However, a common technical criticism is their "black-box" nature. Most systems perform simple prompt engineering with a single API call, lacking sophisticated mechanisms for context management, user modeling, or iterative refinement [8]. Furthermore, existing recommender systems in education often rely on simplistic rule-based approaches or basic collaborative filtering, which fail to capture the complex dynamics of pedagogical needs [9]. This study addresses these technical deficiencies by proposing a more robust and user-centered system architecture.

2.3. Key Technologies for Collaborative AI Systems

Building a truly collaborative AI requires integration of several advanced technologies.

Retrieval-Augmented Generation (RAG): RAG addresses the hallucination problem of LLMs by grounding generation in a knowledge base [10]. This is crucial for educational contexts where accuracy is paramount.

Explainable AI (XAI): XAI techniques, such as attention visualization and feature importance scores, are essential for building user trust [11]. In our system, we leverage XAI to provide insights into why a specific recommendation or content piece was generated.

Recommender Systems: Hybrid recommender systems combining content-based and collaborative filtering are widely used [12]. However, their application in dynamically co-designing pedagogical content is novel.

Affective Computing: Recognizing and adapting to user emotional states through sentiment analysis and interaction pattern analysis can significantly improve user engagement and system adoption [13].

Reinforcement Learning from Human Feedback (RLHF): RLHF is a powerful technique for fine-tuning LLMs to align with human preferences [14]. Our DPO algorithm builds on this concept but operates at the inference stage, providing a more lightweight and adaptable optimization mechanism.

2.4. Research Gap

Existing literature either provides high-level frameworks for AI literacy or focuses on improving general-purpose LLMs. There is a distinct lack of empirical research that: (a) designs and implements a complete system architecture with specific technical modules for collaborative HCI; (b) develops and evaluates novel algorithms for optimizing human-AI interaction in a specialized domain; and (c) validates these technical solutions through rigorous, mixed-methods evaluation. Moreover, as Bansal et al [15]. point out, effective human-AI teaming requires not only accurate outputs but also transparent explanations and mutual adaptation – aspects that are often overlooked in educational systems. This study directly fills this gap.

3. Empirical USER Study: Informing System Design

3.1. Methodology

Participants: 108 open university English teachers from Hubei, China (68 female, 40 male; mean experience = 11.4 years).

Instruments: A 32-item questionnaire (Cronbach's $\alpha = 0.87$) assessing usage frequency, perceived interaction dimensions, and barriers. Semi-structured interviews with 12 teachers were also conducted.

3.2. Key Findings for System Design

Low Interaction Maturity: The mean self-perceived interaction orientation was 2.67/5. Only 23.1% used AI weekly. Usage was predominantly for simple tasks (e.g., material generation: 67.6%, grammar checking: 61.1%), not collaborative work.

Four Key Interaction Dimensions: Exploratory factor analysis identified four crucial dimensions:

Pedagogical Co-Design (PCD): Systems should enable collaborative lesson planning and adjustment. Current systems lack interactive dialogue and user-control mechanisms.

Data-Informed Analytics Dashboard (DID): Systems fail to present learner data in an intuitive, actionable format for pedagogical decision-making.

Ethical AI Stewardship (AES): Users are concerned about bias and data privacy but lack system-level support.

Affective-Emotional Resilience (AER): Systems do not support user confidence or emotional readiness, leading to anxiety and resistance.

Qualitative Themes: Interviews highlighted a critical demand for system transparency (XAI) and a transition from simple automation to cognitive augmentation.

These findings provide the explicit technical requirements that we translated into system architecture and algorithmic design, as detailed in the following sections.

4. System Architecture: THE Collaborative AI Teaching Assistant (Caia)

CAITA is a cloud-native, microservices-based system with five layers (see Figure 1 in the final version). We summarise key layers below.

4.1. Presentation Layer

This layer provides the user interface, built with a responsive web framework (React.js). It renders the Interactive Co-Design Interface, a chat-based interactive workspace that supports multi-turn dialogues. It also hosts the Data-Driven Analytics Dashboard, which visualizes learner performance through interactive charts (using D3.js and ECharts), and an Ethical AI Stewardship Module interface for reviewing biases and managing data privacy.

4.2. API Gateway & Service Layer

Kong handles routing, authentication, rate limiting. Core microservices:

User & Context Management (Redis cache)

RAG Service – uses Sentence-BERT embeddings and Pinecone vector DB for context retrieval

Recommendation Service – implements our CFR mechanism

Analytics Service – batch (Spark) and stream (Flink) processing for learner metrics

Ethical AI Service – bias detection and data anonymisation

Affective Computing Service – sentiment analysis on interaction logs

4.3. Data Layer

The Data Layer consists of multiple databases:

PostgreSQL: For relational data (user profiles, course structures, learner grades).

MongoDB: For storing semi-structured data, including user interaction logs, LLM output logs, and feedback history.

Pinecone: A vector database for efficient similarity search in the RAG pipeline.

MinIO: An object storage service for static files (e.g., learning materials, reference documents).

4.4. AI Engine

Integrates a commercial LLM (GPT-4 or open-source) wrapped with our DPO algorithm, which sits between RAG and LLM API to intelligently craft and refine prompts.

5. KEY Algorithmic Innovations and Optimizations

5.1. Dynamic Prompt Optimization (DPO) for Context-Aware Generation

The DPO algorithm is our primary technical contribution. Standard prompt engineering is static and fails to adapt to the specific needs of a user. DPO treats prompt construction as a sequential decision-making problem, learning an optimal policy to generate system prompts that maximize user satisfaction.

Mathematical Formulation:

Let $\mathcal{S}$ be the state space representing the user context (e.g., teaching goal, student proficiency level, history of interactions). Let $\mathcal{A}$ be the action space of possible prompt components (e.g., tone modifiers, instruction templates, reference materials). Our DPO policy $\pi(a | s)$ learns to select the optimal action given the state. The reward function R is estimated through user feedback:

$$R = \alpha \cdot \text{UserFeedbackScore} + \beta \cdot \text{TaskCompletionEfficiency} - \gamma \cdot \text{ResponseTime}$$

We implement the optimization using a Contextual M Armed Bandit (CMAB) approach with an ϵ -greedy exploration strategy. This is computationally efficient for online learning. For each interaction:

State Representation: The system context is encoded as a feature vector s , comprising key-value pairs representing the user model and task.

Action Selection: The algorithm selects a prompt template from a predefined library, choosing between exploitation (best-performing action) and exploration (random action).

Reward Computation: After the LLM generates content, the user can provide explicit feedback (e.g., "helpful" / "not helpful") which is mapped to a numerical score. Implicit feedback (e.g., copy-paste actions, time spent on generated content) is also captured and normalized.

Policy Update: The algorithm updates the action-value estimates $Q(s,a)$ using a stochastic gradient descent update:

$$Q(s,a) \leftarrow Q(s,a) + \eta \cdot (R - Q(s,a))$$

where η is the learning rate.

Optimization for Educational Context: We pre-populate the prompt template library with pedagogical strategies derived from educational theory (e.g., Bloom's Taxonomy) to ensure pedagogical soundness.

5.2. Collaborative Filtering-Based Recommendation (CFR) for Pedagogical Resources

The CFR mechanism is designed to suggest relevant teaching resources (e.g., lesson plans, exercises, supplementary reading) by combining user-based and item-based collaborative filtering.

User-Item Interaction Matrix: Let U be the set of m users (teachers) and I be the set of n pedagogical items. We construct a sparse matrix R of size $m \times n$, where R_{ui} indicates the interaction weight between user u and item i (derived from usage logs, ratings, and time spent).

Addressing the Cold-Start Problem: We augment this with a content-based component using topic modeling (LDA) on the item descriptions. For new items, a similarity score with existing items is computed based on their topic distributions, enabling recommendations even with limited interaction data.

5.3. Data-Driven Analytics Dashboard: Multi-Dimensional Data Modeling

To support DID, we designed a multi-dimensional model for learner performance with dimensions: Learner, Time, Assessment, Knowledge Component (KC). ETL pipeline: extract from LMS, transform (KNN imputation), load into data warehouse; build OLAP cube for efficient queries; visualise with drill-down capabilities.

5.4. Affective Computing Module: Sentiment Analysis for User Resilience

We analyse interaction logs (non-intrusive) using:

Prompt linguistic features: fine-tuned RoBERTa on teacher-AI dialogues to classify sentiment (frustration, uncertainty, satisfaction).

Behavioural features: latency, prompt rephrasing rate, undo/regenerate frequency.

When negative state detected (confidence < threshold), trigger resilience protocol:

Simplify output via DPO template selection.

Display motivational nudges or task breakdown.

Suggest human-in-the-loop (help docs or community forum).

6. Experimental Evaluation

6.1. Evaluation Setup

We evaluated CAITA through two complementary approaches: (a) Technical Performance Evaluation to assess the efficiency and accuracy of our algorithms; and (b) User Study (A/B Test) to measure the impact of our technical features on user experience.

Baseline System: A standard AI-assisted system using a direct LLM API (GPT-4) with static prompts, a simple rule-based recommendation engine, and no analytics dashboard. This represents the "tool-using" paradigm.

Participants: 60 new teachers from the same pool (not in the initial user study) were randomly assigned to either the CAITA group (n=30) or the Baseline group (n=30). The study lasted four weeks.

6.2. Technical Performance Results

Metric	CAITA	Baseline	Improvement
Avg. end-to-end latency	1.8 s	2.1 s	14.3%
BLEU-4 (vs. expert)	0.52	0.38	+36.8%
Precision@10 (recommendation)	0.74	0.52	+42.3%
Recall@10	0.61	0.43	+41.9%

(Detailed breakdown: RAG retrieval 0.3s, DPO 0.2s, LLM 1.1s, post-processing 0.2s.)

These improvements are attributed to RAG context and DPO optimisation.

6.3. User Study Results

Quantitative Findings – The following table summarises all key metrics with statistical significance: (As shown in Table 1).

Table 1. Comparison of User Experience Metrics between Caita and Baseline (Mean $\pm$ SD)

Metric	CAITA (n=30)	Baseline (n=30)	Improvement	T-value	P-value
Interaction Maturity (1-5)	3.62 $\pm$ 0.78	2.58 $\pm$ 0.91	+25.6%	4.89	<0.001
Task Completion Time (min)*	28.4 $\pm$ 6.2	36.1 $\pm$ 8.5	-21.3%	-4.12	<0.001
System Usability Scale (SUS)	78.4 $\pm$ 9.1	62.1 $\pm$ 11.3	+26.2%	6.05	<0.001
Prompt Rephrasing Rate (per session)	2.1 $\pm$ 0.9	3.2 $\pm$ 1.2	-34.4%	-4.01	<0.001

Voluntary Continued Usage (%)**	86.7%	46.7%	+40.0%	$\chi^2=10.8$	<0.01
---------------------------------	-------	-------	--------	---------------	-------

*Time to complete a standardised 45-minute lesson plan design task.

**Percentage of participants who voluntarily used the system beyond the required study period.

6.4. Qualitative Feedback

T08 (CAITA): "The system didn't just give me content; it explained why and offered alternatives. I felt like I was exploring, not just receiving."

T11 (Baseline): "It's faster than searching myself, but it often gives generic answers. I don't trust it for my specific context."

6.5. Discussion

Our experimental results provide strong evidence for the efficacy of the CAITA architecture and its key algorithms.

DPO and RAG are synergistic: RAG provided the necessary context, while DPO optimized how that context was used in prompting. The improved BLEU scores confirm this synergy.

CFR effectively reduces cognitive load: By offloading the resource search task, the system allowed teachers to focus on higher-order pedagogical co-design. This explains the increased task efficiency and interaction maturity.

Affective computing is a critical enabler: The sentiment analysis module's ability to detect and respond to frustration is vital for adoption, particularly among users with low technical confidence. This is a key differentiator from standard AI interfaces.

Limitations: Our experiments were conducted over a four-week period. Longer-term studies are needed to assess the sustained impact and the evolution of the DPO policy. The bias detection module was not fully validated due to a limited dataset; this will be a focus of future work.

7. Conclusion and Future WORK

This paper has presented a comprehensive computer science-driven study on the design, optimization, and evaluation of an AI-assisted educational system aimed at fostering human-AI collaboration. Moving beyond the generic "tool-using" paradigm, we designed the Collaborative AI Teaching Assistant (CAITA), whose architecture and algorithms are directly informed by a systematic empirical user study.

Our primary technical contributions are threefold. First, we proposed a modular system architecture integrating RAG, an analytics dashboard, ethical AI, and affective computing modules. Second, we introduced a Dynamic Prompt Optimization (DPO) algorithm that learns to construct context-aware prompts from user feedback, and a Collaborative Filtering-based Recommendation (CFR) mechanism that provides personalized pedagogical resource suggestions. Third, we validated our design through rigorous technical benchmarks and a controlled user study, demonstrating significant improvements in user interaction maturity (25.6%) and task efficiency (21.3%).

Our work directly addresses the call for more sophisticated HCI systems in educational technology. It shifts the focus from simple LLM API integration to the design of a complete, intelligent, and user-centric system that learns and adapts to its users. This aligns with the broader goal of achieving human-AI symbiosis.

Future Work: Enhanced Bias Detection: We will integrate a more robust bias detection framework that evaluates generated content across multiple demographic and cultural dimensions, using recently released fairness datasets.

Multi-Modal Feedback Integration: We plan to extend the affective computing module to process audio and video inputs to capture non-verbal cues, providing a more holistic understanding of user state.

Federated Learning: To address data privacy concerns, we will explore federated learning techniques to train the DPO algorithm across multiple institutions without centralizing sensitive user data.

Longitudinal Study: We will conduct a 6-month longitudinal study to analyze the long-term evolution of the DPO policy and its impact on teaching practices.

Funding: This study was funded by China Adult Education Association General Project on Artificial Intelligence (grant number AI-Y2025226S).

References

1. O. Zawacki-Richter, V. I. Marín, M. Bond, and F. Gouverneur, "Systematic review of research on artificial intelligence applications in higher education – where are the educators?," *Int. J. Educ. Technol. High. Educ.*, vol. 16, no. 1, p. 39, 2019, doi: 10.1186/s41239-019-0171-0.
2. R. Luckin, W. Holmes, M. Griffiths, and L. B. Forcier, *Intelligence Unleashed: An Argument for AI in Education*. London, U.K.: Pearson, 2016. [Online]. Available: <https://www.pearson.com/content/dam/one-dot-com/one-dot-com/global/Files/about-pearson/innovation/Intelligence-Unleashed-Publication.pdf>
3. X. Feng, R. Wang, and Y. Wu, "Blended teaching research in China and abroad," *Chin. J. Distance Educ.*, no. 4, pp. 9–19, 2021. [Online]. Available: <http://www.chinadledu.com/>
4. S. Amershi, D. Weld, M. Vorvoreanu, A. Fourney, B. Nushi, P. Collisson, et al., "Guidelines for human-AI interaction," in *Proc. 2019 CHI Conf. Hum. Factors Comput. Syst.*, 2019, pp. 1–13, doi: 10.1145/3290605.3300233.
5. D. Cetindamar, K. Kitto, M. Wu, Y. Zhang, and S. Knight, "AI literacy: A multi-level framework," *Technol. Forecast. Soc. Change*, vol. 198, p. 122884, 2024, doi: 10.1016/j.techfore.2023.122884.
6. I. Roll and R. Wylie, "Evolution and revolution in artificial intelligence in education," *Int. J. Artif. Intell. Educ.*, vol. 26, no. 2, pp. 582–599, 2016, doi: 10.1007/s40593-016-0108-8.
7. R. Godwin-Jones, "Partnering with AI: Intelligent writing assistance and instructed language learning," *Lang. Learn. Technol.*, vol. 26, no. 2, pp. 5–24, 2022. [Online]. Available: <https://hdl.handle.net/10125/73472>
8. D. T. K. Ng, J. K. L. Leung, S. K. W. Chu, and M. S. Qiao, "Conceptualizing AI literacy: A systematic review," *Comput. Educ. Artif. Intell.*, vol. 6, p. 100218, 2024, doi: 10.1016/j.caeai.2024.100218.
9. M. A. Kuhail, N. Alturki, S. Alramlawi, and K. Alhejori, "Interacting with educational chatbots: A systematic review," *Educ. Inf. Technol.*, vol. 28, no. 1, pp. 973–1018, 2023, doi: 10.1007/s10639-022-11177-3.
10. P. Lewis, E. Perez, A. Piktus, et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 9459–9474. [Online]. Available: <https://arxiv.org/abs/2005.11401>
11. W. Samek, T. Wiegand, and K. R. Müller, "Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models," *arXiv preprint arXiv:1708.08296*, 2017. [Online]. Available: <https://arxiv.org/abs/1708.08296>
12. F. Ricci, L. Rokach, and B. Shapira, "Recommender systems: introduction and challenges," in *Recommender Systems Handbook*, 2nd ed. Boston, MA: Springer, 2015, pp. 1–34.
13. R. W. Picard, *Affective Computing*. Cambridge, MA: MIT Press, 1997. [Online]. Available: <https://mitpress.mit.edu/9780262661151/affective-computing/>
14. L. Ouyang, J. Wu, X. Jiang, et al., "Training language models to follow instructions with human feedback," in *Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 27730–27744. [Online]. Available: <https://arxiv.org/abs/2203.02155>
15. G. Bansal, B. Nushi, E. Kamar, W. S. Lasecki, D. S. Weld, and E. Horvitz, "Does the whole exceed its parts? The effect of AI explanations on complementary team performance," in *Proc. 2019 CHI Conf. Hum. Factors Comput. Syst.*, 2019, pp. 1–12, doi: 10.1145/3290605.3300664.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of Publisher and/or the editor(s). Publisher and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.