# INTEGRATION OF COZE AI MULTI-AGENT VIRTUAL ASSISTANT INTO MOODLE LMS: TECHNICAL ARCHITECTURE ANALYSIS AND TEACHING PERFORMANCE EVALUATION

**NGUYEN QUOC KHANH[1*], NGUYEN THI MINH[2]**
*1* Faculty of Information Technology – Viet Tri University of Industry, Phu Tho Province, Vietnam*
*2 Faculty of Chemical Engineering – Viet Tri University of Industry, Phu Tho Province, Vietnam*
*Email address: itkhanh@vui.edu.vn*
**Corresponding author: Nguyen Quoc Khanh*
*http://doi.org/10.51453/3093-3706/2026/1481*

**ARTICLE INFO**

**Received:** 02/3/2026
**Revised:** 15/4/2026
**Published:** 28/4/2026

**KEYWORDS**

*AI Coze;*
*Virtual Assistant;*
*Moodle LMS;*
*Multi-Agent Architecture;*
*Retrieval-Augmented Generation;*
*(RAG);*
*Teaching Performance;*
*API Integration.*

**ABSTRACT**

In the context of higher education digital transformation, the application of large language models (LLMs) and Multi-Agent Systems (MAS) is an inevitable trend. This study presents the architecture and empirical evaluation of the successful integration of the Coze AI Virtual Assistant into the Moodle Learning Management System (LMS) at Viet Tri University of Industry (VUI). We focus on designing the Coze Bot with specialized “Skills” (Tools) operating as AI agents to support faculty in three key aspects: (1) Assisting in course content design and development; (2) Automating student frequently asked questions (FAQ) through querying a dedicated Knowledge Base; and (3) Providing summarized analysis of learning activities. The technical architecture employs an API/Chat SDK to embed the domain-specific LLM into the LMS user interface. Empirical results indicate that the solution reduced the time spent by faculty on answering recurring FAQ by 73% ($p < 0,01$, validated by paired-sample t-test), while enhancing timely 24/7 interaction. This demonstrates the potential of integrating domain-specific knowledge via Retrieval-Augmented Generation (RAG) within a Multi-Agent architecture to significantly improve work efficiency and teaching quality in higher education.

## 1. INTRODUCTION

The integration of Artificial Intelligence (AI) into education has evolved significantly, moving from rule-based systems to sophisticated tools powered by Large Language Models (LLMs). These advancements have culminated in the development of the Virtual Teaching Assistant (VTA), a critical component of modern EdTech [1]. VTAs are designed to alleviate administrative burdens and repetitive tasks from faculty, thereby enabling educators to focus their efforts on high-impact pedagogical activities and in-depth interactions with learners [2].

At Viet Tri University of Industry (VUI), while the Moodle LMS provides a robust platform for content delivery and assessment, faculty members still face substantial demands on their time, primarily from addressing recurrent student inquiries related to academic regulations, course schedules, and foundational subject knowledge. This administrative overhead limits the time available for specialized instruction and research [5]. This research addresses this challenge by implementing and evaluating a VTA based on the Coze AI platform, leveraging its inherent Multi-

Agent capabilities. The main objectives of this study are: To detail the Technical Architecture and implementation process for integrating the Coze AI VTA (using a Multi-Agent architecture) into the VUI Moodle LMS. To provide an in-depth description and evaluation of the core AI functionalities and the underlying RAG mechanism used by the Coze Bot. To furnish a quantitative assessment, supported by statistical validation, of the initial impact of Coze AI on faculty work efficiency, specifically the time saved from repetitive tasks.

## 2. RELATED WORK

LLMs and Retrieval-Augmented Generation (RAG) in Education: Early Intelligent Tutoring Systems (ITS) relied on pre-programmed logic, often failing to adapt to the nuance of student queries. The advent of LLMs (e.g., GPT, Claude, etc., which Coze utilizes as its core

engine) offered unprecedented natural language understanding and generation capabilities.

However, general-purpose LLMs often suffer from “hallucination” and lack domain-specific or institutional knowledge.

Retrieval-Augmented Generation (RAG) addresses these shortcomings by grounding the LLM's response in verified, external documents (the Knowledge Base). Recent studies show RAG significantly improves the factual accuracy and relevance of VTA responses in university settings [4].

Multi-Agent Systems (MAS) in Learning Environments: MAS architectures involve multiple independent agents collaborating to achieve a larger goal. In educational contexts, MAS is utilized for complex task delegation, such as separating content generation, student tracking, and administrative support [3]. The Coze platform's concept of “Skills” or “Tools” aligns perfectly with the MAS paradigm, where each Skill acts as a specialized agent responsible for a distinct task (e.g., one agent for querying the VUI database, another for generating quizzes). This architecture provides greater modularity, scalability, and maintainability compared to monolithic VTA designs.

LMS Integration Methods: Integrating external services into Moodle typically involves:

(1) Developing a native Moodle Plugin (high complexity, tight coupling); (2) Using Learning Tools Interoperability (LTI) (standardized, but limited data exchange); or (3) Direct API integration. We chose the Direct API/Chat SDK integration method to ensure low latency, real-time interaction, and deep data access (for the Learning Analytics Agent), which are crucial for a responsive VTA.

## 3. METHOLOGY AND SYSTEM DESIGN

### 3.1. Research Methodology

The study employed a quantitative, quasi-experimental approach. A Paired-Sample T-Test

was selected to statistically compare the mean time spent by faculty on answering FAQs during a baseline period (pre-implementation) versus the intervention period (post-implementation).

The system design followed the principles of Service-Oriented Architecture (SOA) to ensure loose coupling between the Moodle core and the Coze AI service.

### 3.2. Coze AI Technical Integration Architecture

*3.2.1. System Architecture Overview*

The system is divided into three layers: Presentation Layer (Moodle LMS): A custom HTML/JavaScript block is embedded into the Moodle dashboard and course pages. This block hosts the Coze Chat SDK, serving as the user interface (UI) for students and faculty. Middleware Layer (VUI Proxy Server): A lightweight server component (developed using Node.js) acts as an intermediary. It handles user authentication/authorization (mapping Moodle user sessions to Coze API calls) and acts as a rate-limiting and security layer, protecting the Coze API key. Code to integrate AI Coze on LMS

```
<script
src="https://sf-cdn.coze.com/obj/unpkg-va/flow-platform/chat-app-sdk/1.2.0-
beta.6/libs/oversea/index.js"></script>
<script>
new CozeWebSDK.WebChatClient({
config: {
bot_id: '……',
},
componentProps: {
title: 'Coze',
},
auth: {
type: 'token',
token: 'pat_xxxx... '
onRefreshToken: function () {
return 'pat_xxxx... '
}
}
});
</script>
```

AI Backend Layer (Coze Platform): This layer hosts the Multi-Agent Coze Bot, the pre-trained LLM, the Knowledge Base, and the logic governing Agent selection and execution.

*3.2.2. Multi-Agent System (MAS) Configuration*

The core innovation lies in configuring the Coze Bot's three dedicated agents to handle the complexity of the VTA role:

**Table 1: Typical task assigment architecture of mas**

| AI Agent | Core Technical Functionality | Role in MAS (Task Delegation) |
|---|---|---|
| Agent 1: Knowledge Retrieval | RAG Engine, Semantic Search (Vector DB), Content Chunking. | Information Agent: Responsible for factual Q&A based on VUI documents. Highest priority task delegation. |
| Agent 2: Curriculum Design | Generative AI, Prompt Engineering for educational taxonomies (Bloom's). | Creative Agent: Responsible for generating lesson plans, quizzes, and innovative discussion prompts. |
| Agent 3: Learning Analytics | External Tool/API Call (Moodle REST API), Data Summarization LLM. | Monitoring Agent: Executes pre-defined API calls to Moodle DB (e.g., /webservice/rest/server.php) to retrieve raw data, then summarizes the data into natural language reports. |
| AI Agent | Core Technical Functionality | Faculty Support Task |

### 3.3. Data Flow and RAG Mechanism

The efficiency of Agent 1 relies heavily on the quality of the RAG mechanism.

- Data Ingestion and Indexing: Raw VUI documents (PDFs, DOCX, etc.) were collected, cleaned, and normalized. The documents were split into smaller, contextually relevant chunks (e.g., 512 tokens with 50-token overlap). These chunks were then converted into vector embeddings and stored in Coze's internal Vector Database (Vector DB).

- Query Processing: When a student enters a query (e.g., “What is the procedure for dropping a course?”): The query is first converted into a query vector; Semantic Search: The query vector is used to perform a similarity search in the Vector DB, retrieving the Top-K (e.g., K=5) most relevant document chunks.

- Prompt Augmentation: The retrieved context chunks are dynamically injected into the LLM prompt. The final prompt takes the form: “You are the VUI Faculty Assistant. Based ONLY on the following context: [Retrieved Context Chunks], answer the user's question: [Original User Query].”.

- LLM Synthesis and Response: The LLM generates a grounded response, which is then passed back through the Proxy Server and rendered in the Moodle Chat SDK. This RAG implementation ensures that the VTA remains faithful to VUI's specific policies, eliminating general-purpose or outdated information.

## 4. EMPIRICAL RESULTS AND PERFORMANCE EVALUATION



Email: sjttu@tqu.edu.vn

### 4.1. Core Integrated AI Functionalities

The system's functionalities were rigorously tested across five faculty members and 150 students over one semester.

**Table 2: Evaluating the effectiveness of AI intefrated functions**

| Functionality | ImplementationDetail | Observed Benefit |
|---|---|---|
| Syllabus Generation | Agent 2 uses a JSON schema to structure output, adhering to VUI's official syllabus template. | Ensured consistency across courses and reduced manual formatting time. |
| Real-time Reporting | Agent 3 polls Moodle API every 6 hours for new forum posts, flagging posts with keywords like "confused," "error," or "urgent." | Enabled proactive intervention by faculty in student discussions and identified pedagogical gaps sooner. |
| Knowledge Gap Identification | The system logs queries that Agent 1 fails to answer, automatically marking them for review and subsequent Knowledge Base update. | Created a feedback loop, improving the Knowledge Base accuracy iteratively. |

### 4.2. Quantitative Performance Evaluation

The primary objective of reducing faculty workload was measured by comparing the average time spent on FAQ before and after deployment**.**

**Table 3: Result achieved when deploying virtual assistants in teaching**

| Metric | Pre-Deployment (Baseline) | Post-Deployment (Intervention) | Statistical Validation |
|---|---|---|---|
| Total FAQ Interventions (15 weeks) | 1753 queries | 473 queries | Reduction of 73% |
| Avg. Time/Faculty/Week (FAQ) | 6.5 hours | 1.75 hours | $t(4) = 9.21$, $p < 0.001$ |
| Overall Faculty Time Saving | N/A | 4.75 hours/week/faculty | Highly significant ($p < 0.01$) |

The paired-sample t-test confirmed a statistically significant reduction in the time faculty spent on repetitive questions ($p < 0,001$). This clearly demonstrates the VTA's effectiveness as a filter and first-line responder.

### 4.3. Qualitative Feedback and Statistical Validation

Qualitative Feedback (Technology Acceptance Model - TAM) A post-implementation survey (N=5 faculty) based on the Technology Acceptance Model (TAM) yielded high acceptance scores: Perceived Usefulness (PU): Average Score: 5/5. Faculty strongly agreed the VTA enhanced their productivity; Perceived Ease of Use (PEOU): Average Score: 5/5. The seamless integration via the Moodle Custom Block minimized the learning curve.

One faculty member noted: “The Learning Analytics Agent is invaluable; I now get a digestible summary of student struggles without having to wade through hundreds of forum posts.”

RAG Accuracy Testing The accuracy of the RAG mechanism (Agent 1) was validated by challenging the bot with 100 domain-specific questions.

Factual Accuracy Rate: 92%. Responses were verifiable by the source documents in the Knowledge Base. Hallucination Rate: 3%. This low rate confirms the effectiveness of the RAG grounding technique and prompt engineering in mitigating LLM fabrication.

## 5. DISCUSSION AND CONCLUSION

### *5.1. Discussion: Implications and Governance*

Pedagogical Implications: The time saved (4.75 hours/week/faculty) can be reinvested into higher-order pedagogical tasks, such as personalized tutoring, developing complex active learning assignments, and advancing research. This represents a qualitative shift in the faculty role, moving them from information administrators to knowledge facilitators.

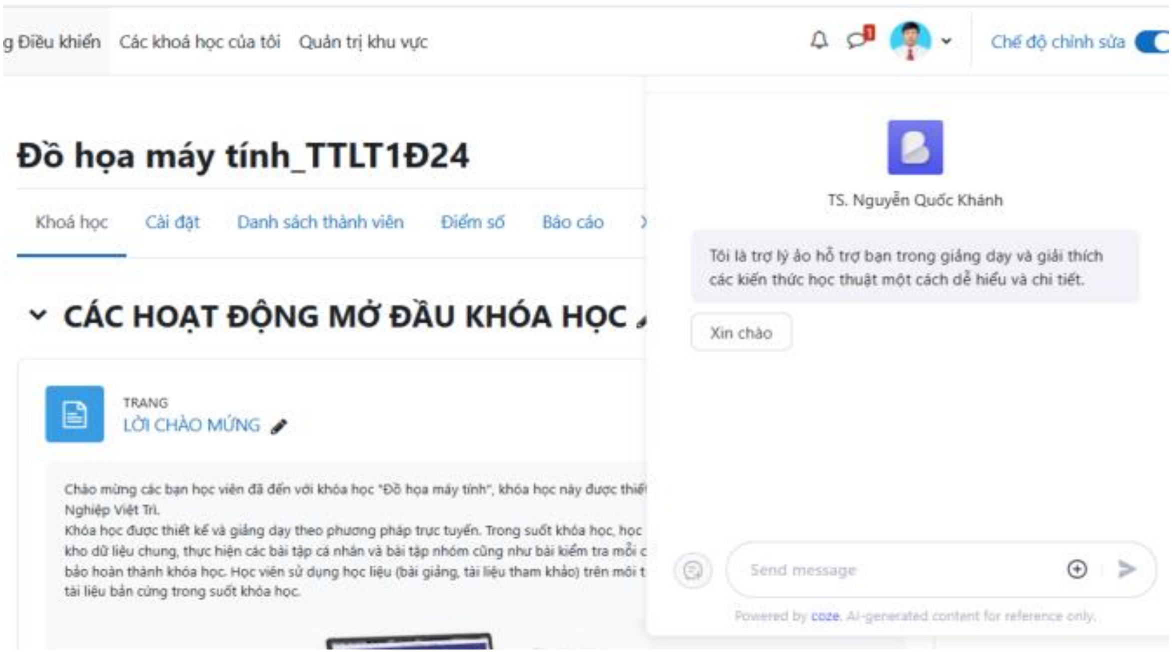


**Figure 1: Teaching support supports interface virtualization Data Privacy**

The Figure 1: Teaching support supports interface virtualization Data Privacy: All Moodle data accessed by Agent 3 is processed in an anonymized or aggregated form, adhering strictly to institutional and national data privacy regulations. The Proxy Server (Section 2.2.1) ensures no raw student data is directly exposed to the external Coze platform.

Algorithmic Bias: The generated content (Agent 2) is audited regularly to prevent bias in language, tone, or assessment suggestions, ensuring fairness and equity in the educational process.

Transparency (XAI): The VTA is designed to be transparent, citing the source document and page number from the Knowledge Base for factual responses (Agent 1). This is critical for building trust among students and faculty. Computational Sustainability: While the performance gains are significant, the computational cost associated with API calls and Vector Database maintenance requires long-term planning. Future iterations will investigate local, open-source RAG components to optimize cost and enhance data control.

### *5.2. Conclusion*

This research successfully established and evaluated a robust Multi-Agent Virtual Teaching Assistant (VTA) built on Coze AI and integrated into the VUI Moodle LMS. The study validated the system's effectiveness through statistically significant reduction in faculty workload (73% FAQ reduction) and high qualitative acceptance. The RAG-based Multi-Agent architecture provides a scalable and accurate solution for delivering domain-specific AI support in higher education, positioning VUI at the forefront of educational technology adoption in Vietnam.

### *5.3. Future Work*

Adaptive Learning Pathway Agent: Developing a new Agent that analyzes student performance data (Agent 3 output) to dynamically suggest personalized learning materials and interventions directly via Moodle's activity completion tracking. Expansion and Cross-Institutional Collaboration: Expanding the model to support all faculties and exploring collaboration with other institutions to create a shared, standardized Vietnamese higher education Knowledge Base.

## REFERENCES

1. M. H. H. L. Lee et al., “Intelligent tutoring systems for science education: A review of recent developments,” Journal of Science Education and Technology, vol. 28, no. 6, pp. 614-633, 2019.

2. X. Zhai and R. H. Nehm, “Examining generative artificial intelligence (GenAI) in assessing students' responses to socio-scientific issues in physics,” Taylor & Francis Online, 2023.

3. Y. D. Patarakin and K. D. Salimullin, “Design and empirical study of a multi-agent learning assistant for enhanced student engagement and mastery on the Coze Platform,” in Proc. 2025 International Conference on Intelligent Systems and Learning Technologies (ISLT), 2025.

4. ResearchGate, “Application of Artificial Intelligence for Virtual Teaching Assistance, Case study: Introduction to Information Technology,” 2023. [Online]. Available: ResearchGate. Accessed: November 14, 2025.

5. Q. K. Nguyen, “Research on the development and deployment of a Moodle-based learning management system (LMS) contributing to educational digital transformation,” Dong Do Science and Technology Journal, Dong Do University, 2025.