HNUE JOURNAL OF SCIENCE
Journal of Educational Science 2026, Volume 71, Issue 4, pp. 88-98
ISSN: 2354-1075; Website: https://hnuejs.edu.vn/es
DOI: 10.18173/2354-1075.2026-0060

DEVELOPING A VIRTUAL ASSISTANT TO SUPPORT INFORMATICS LESSON PLANNING

Nguyen Hong Phong, Do Minh Quan, Le Xuan Hien*, Pham Tho Hoan
School of Mathematics and Computer Science, Hanoi National University of Education, Hanoi, Vietnam
*Corresponding author: Le Xuan Hien, email: lexuanhien@hnue.edu.vn
Received: March 27, 2026. Revised: May 20, 2026. Accepted: June 09, 2026.

Abstract. Preparing high school Informatics lesson plans is time-consuming, as teachers must meet the knowledge, competency, and quality requirements of the 2018 General Education Curriculum while organizing textbook content in accordance with the structure of Official Dispatch 5512. Large Language Models (LLMs) can assist teachers; however, hallucination – where models generate content that deviates from the curriculum – is the primary obstacle. This study develops a GraphRAG-based virtual assistant, combining Gemini with a Neo4j knowledge graph that encodes the high school Informatics curriculum and the Digital Competence Framework prescribed by Circular 02/2025/TT-BGDĐT and operationalized for school students in Official Dispatch 3456 (MOET, 2025b). The novelty of this research lies in a two-layer control mechanism: an immutable zone extracting data verbatim from the knowledge graph, and a creative zone for the model to flexibly design pedagogical activities. The pedagogical design of the system is justified through the Backward Design and constructive alignment frameworks, clearly defining the structural support role of the virtual assistant while preserving the teacher's pedagogical autonomy. A survey of 35 participants yielded an overall mean score of 4.52/5 (SD = 0.60; $\alpha = 0.898$) across 22 criteria in four groups. A within-subject comparative experiment ($n = 20$) showed that the system was rated higher than a standalone LLM across all four core criteria ($p < 0.05$, Wilcoxon signed-rank test).
Keywords: Virtual assistant, lesson planning, GraphRAG, knowledge graph, large language model.

1. Introduction

The 2018 General Education Curriculum (GEC), promulgated under Circular No. 32/2018/TT-BGDĐT, shifts the focus from knowledge transmission to the development of learners' essential qualities and core competencies (Ministry of Education and Training [MOET], 2018a, 2018b). For the Informatics subject at the high school level, the curriculum defines informatics competence as comprising five components (NLa–NLe) (MOET, 2018a). Official Dispatch No. 5512/BGDĐT-GDTrH (MOET, 2020) further prescribes a standardized lesson plan template structured around four sequential learning activities (warm-up, knowledge formation, practice, and application), each requiring a highly detailed implementation framework. In addition, Circular No. 02/2025/TT-BGDĐT (MOET, 2025a) prescribes a Digital Competence Framework for learners, and Official Dispatch No. 3456 (MOET, 2025b) provides implementation guidance for school students, requiring teachers to integrate digital literacy into instructional activities. Consequently, teachers must simultaneously satisfy the quality, competency, and knowledge objectives stipulated in the 2018 GEC, organize textbook content according to the Official Dispatch 5512 template, and embed digital competence, which results in substantial preparation time per lesson.

Large Language Models (LLMs) have shown considerable potential in assisting teachers to design instructional content. Prior work consistently highlights both their capacity for content personalization and instructional design as well as the necessity of quality-assurance mechanisms against factually incorrect outputs (Kasneci et al., 2023; Baidoo-Anu & Owusu Ansah, 2023).

Experimental results by Zheng et al. (2025) further indicate that integrating structured knowledge substantially improves the quality of generated lesson plans compared to standalone LLMs. Nevertheless, a core challenge remains: the hallucination problem, in which models generate plausible but factually incorrect content (Huang et al., 2025). Retrieval-Augmented Generation (RAG) (Gao et al., 2024) and its graph-based extension, GraphRAG (Edge et al., 2024), offer a viable approach in education. Specifically, by integrating knowledge graphs, these technologies leverage the ability of graphs to effectively represent the interconnected and hierarchical structure of curricula (Abu-Salih & Alotaibi, 2024).

Most studies applying LLMs in education (Kasneci et al., 2023; Baidoo-Anu & Owusu Ansah, 2023; Zheng et al., 2025) have focused on international contexts. Zheng et al. (2025), for instance, developed Lesson Plan LM, which integrates an LLM with a knowledge base derived from over 100,000 real-world lesson plans in China; however, the system represents curriculum information only as discrete data fields rather than as a structured knowledge graph encoding the national curriculum. To our knowledge, no prior work has addressed a lesson plan preparation assistant for the Vietnamese high school informatics curriculum that simultaneously satisfies the constraints of the 2018 GEC and Official Dispatch 5512. Accordingly, this study makes three primary contributions. First, we construct a domain-specific Neo4j knowledge graph encoding the full Vietnamese high school informatics curriculum from the "Kết nối tri thức với cuộc sống" textbook series, together with the Digital Competence Framework specified in Circular 02/2025/TT-BGDĐT and Official Dispatch 3456 – representing the first publicly documented structured encoding of this curriculum. Second, we develop a GraphRAG-based virtual assistant integrated with Gemini that drafts lesson plans compliant with Official Dispatch 5512, and propose a two-layer control mechanism separating an immutable zone (curriculum data retrieved from the graph) from a creative zone (pedagogical decisions left to the teacher), thereby mitigating hallucination at the data layer while preserving teachers' pedagogical autonomy. Third, a within-subject empirical comparison (n = 20) shows that the knowledge-graph-augmented system produces lesson plans rated significantly higher than those from a standalone LLM in terms of structural compliance and curriculum accuracy.

2. Literature review and research approach

2.1. Pedagogical theoretical framework

Two theoretical frameworks help explain why the four-activity structure of Official Dispatch 5512 (MOET, 2020) lends itself to partial automation. Under the backward design framework (Wiggins & McTighe, 2005), instruction is planned by first defining learning outcomes, followed by assessment evidence, and finally instructional activities – a sequence that mirrors the objective-first logic of Official Dispatch 5512. The constructive alignment framework (Biggs, 1996) adds a further constraint: objectives, tasks, and assessment must be internally consistent within each activity. In Official Dispatch 5512, this consistency is operationalized through four mandatory components (Objectives, Content, Products, and Organization of Implementation) along with the four implementation steps described in Section 1. This rule-governed regularity makes the templated portion amenable to computational support, while the teacher retains autonomy over pedagogical decisions – such as selecting methods, adjusting content, and designing activities (Celik, 2023; Kasneci et al., 2023).

When the TPACK framework (Mishra & Koehler, 2006) is overlaid on these two lenses, three design principles for the support system emerge. The first principle concerns the anchoring of objectives: following the backward design framework, the curriculum's achievement standards must be fixed before any content is generated, rather than inferred by the large language model. The second principle concerns structural completeness: the following constructive alignment framework, every generated activity must contain all four mandatory components (Objectives, Content, Products, and the Organization of Implementation). The third principle concerns the human-machine boundary: following TPACK, the system should handle the technological and

curricular-structure dimensions while the teacher retains creative control over methods, techniques, and contextual adaptation (Celik, 2023).

These three principles strictly govern the design decisions presented in Section 3: the knowledge graph anchors objectives as fixed entities (Section 3.1), while prompt constraints and mandatory schemas ensure structural completeness in alignment with Official Dispatch 5512. Finally, the two-layer control mechanism enforces the human-machine boundary by separating immutable curriculum data from the teacher's pedagogical autonomy (Section 3.2.2).

2.2. Technological approach

While traditional vector Retrieval-Augmented Generation (RAG) relies on semantic similarity, which often overlooks the hierarchical constraints of structured data (Gao et al., 2024), educational curricula inherently possess a complex relational structure linking lessons to specific grades, topics, and objectives. To faithfully encode these multidimensional relationships, this study adopts the GraphRAG framework (Edge et al., 2024). By replacing vector databases with a knowledge graph, GraphRAG enables precise, relationship-based querying. In the context of lesson planning, this approach retrieves a lesson's complete curricular context via a single query, thereby ensuring accurate prompt augmentation and overcoming the limitations of approximate semantic searches.

3. Research findings

3.1. Knowledge Graph Construction

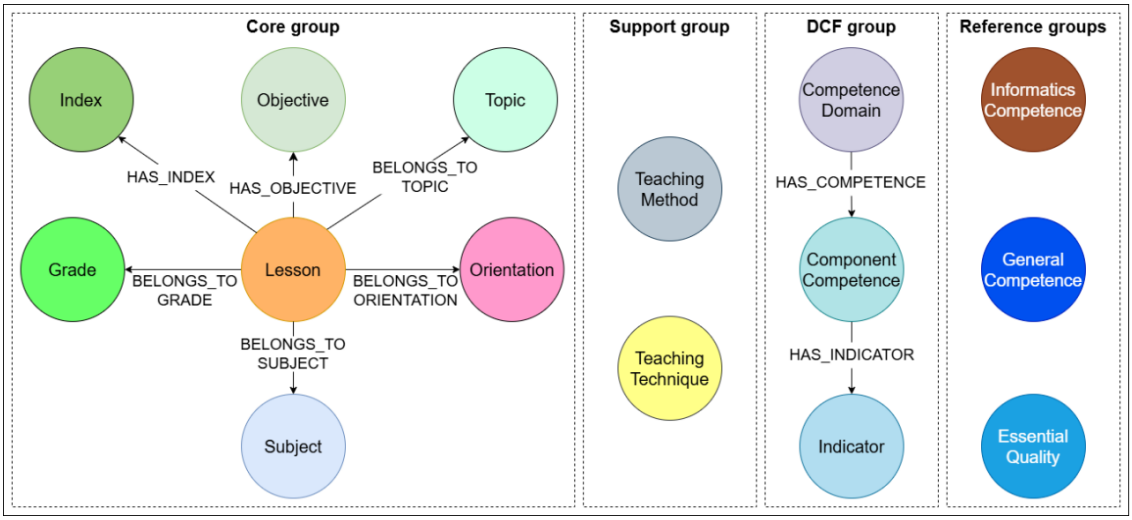

Figure 1. Structure of the High School Informatics curriculum knowledge graph

The knowledge graph is built on Neo4j, encoding the high school informatics curriculum from the "Kết nối tri thức với cuộc sống" textbook series, guided by the General Education Curriculum for Informatics (MOET, 2018a) and the Overall Program (MOET, 2018b). It is organized into 15 entity types across four clusters – core curriculum, teaching support, reference, and the Digital Competence Framework (DCF) – as detailed in Table 1 and illustrated in Figure 1. The DCF comprises six domains and 24 component competencies across four proficiency levels and eight bands under Circular No. 02/2025/TT-BGDĐT (MOET, 2025a); Official Dispatch No. 3456 (MOET, 2025b) maps these to five grade groups, from Basic 1 (grades 1–3) to Advanced 1 (grades 10–12). As this study focuses on the high school level, the graph encodes only the 55 Advanced 1 indicators extracted from the Official Dispatch 3456 appendix.

The knowledge graph was constructed manually in three phases (Figure 2): (1) analyzing the General Education Curriculum for Informatics, the Overall Program (MOET, 2018a, 2018b), and textbooks to identify entities and relationships; (2) structuring the data into CSV files; and (3)

importing the data into Neo4j via Python scripts using batch Cypher queries. Manual construction ensures high curriculum-data accuracy – which is a prerequisite for the two-layer control mechanism.

Table 1. Summary of main entity types in the knowledge graph

Entity Type	Description	Group
Lesson	Grade 10, 11, 12 lessons (theory, practice)	Core
Grade	Grade 10, 11, 12	Core
Topic	Content topic	Core
Objective	Knowledge objective of each lesson	Core
Index	Ordered lesson content index	Core
Subject	Subject	Core
Orientation	Orientation (General, Applied IT, Computer Science)	Core
TeachingMethod	Teaching method with implementation procedure	Support
TeachingTechnique	Teaching technique with implementation procedure	Support
InformaticsCompetence	Specific manifestations of the 5 Informatics competence components (NLa – NLe)	Reference
GeneralCompetence	3 general competencies with manifestations (self-control and self-study, communication and collaboration, problem-solving and creativity)	Reference
EssentialQuality	5 essential qualities with manifestations (patriotism, kindness, diligence, honesty, responsibility)	Reference
CompetenceDomain	6 digital competence domains	DCF
ComponentCompetence	24 component competencies	DCF
Indicator	55 specific indicators at Advanced 1 level (high school, grades 10 – 12) with codes	DCF

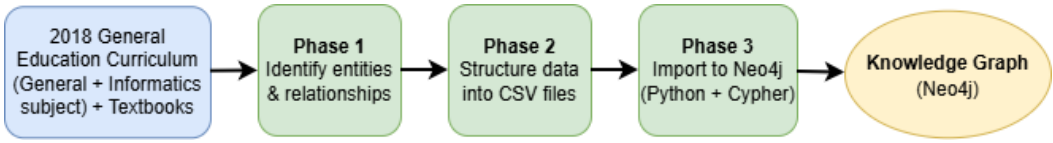

Figure 2. The process of encoding informatics knowledge data

This structure enables retrieval of a lesson's full curriculum data – knowledge objectives, content indices, identification information (grade, topic, orientation), and textbook content – via a single Cypher query, alongside the separate retrieval of reference materials (manifestations of informatics competence, general competencies, and essential qualities) for prompt injection. The next section presents the architecture for integrating the knowledge graph into the virtual assistant.

3.2. Virtual Assistant Architecture

3.2.1. Overall Architecture

The virtual assistant is deployed as five Docker containers (Figure 3): (1) PostgreSQL 16 Alpine for managing users, permissions, lesson plans, shared content, classes, and API token limits; (2) Neo4j Community Edition for hosting the curriculum knowledge graph; (3) a FastAPI/Python backend (Ramírez, 2018) to handle business logic and LLM communication; (4) a React/Nginx frontend to

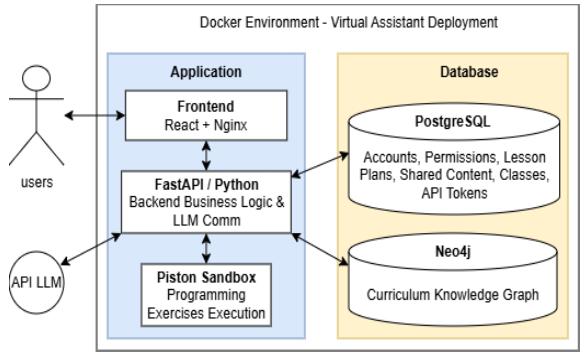

Figure 3. Overall architecture of the virtual assistant supporting lesson plan preparation

serve the UI and proxy API requests; and (5) Piston, which serves as an isolated sandbox for executing programming exercises.

3.2.2. Two-layer Control Mechanism and Prompt Engineering

The assistant's core mechanism separates the prompt into two distinct zones. The immutable zone contains curriculum data extracted from the knowledge graph which is divided in two categories: lesson identification data (name, grade, topic, orientation, and lesson type), which must remain unaltered; and curricular-content data (content index, knowledge objectives, implementation procedures for teaching methods and techniques, manifestations of Informatics competence, general competencies, essential qualities), which must be adhered strictly in accordance with the lesson plan regulations. The LLM is strictly prohibited from fabricating or modifying any information within this zone.

The creative zone is where the LLM operates as an instructional designer, with output grounded in the immutable zone. The system prompt forbids the LLM from merely name-mentioning teaching methods or techniques; instead, it must design a detailed pedagogical script for each activity following the four-step implementation process defined in Section 1. Through these steps, the LLM translates the chosen implementation procedures of the chosen teaching method (retrieved from the knowledge graph) into concrete actions for both the teacher and students, ensuring that the lesson plan reflects the core characteristics of the method rather than just listing its name.

The prompt comprises two main components. First, the system instruction defines the persona of a 'High School Informatics Pedagogical Expert' and imposes four strict rules on output accuracy and the precise description of pedagogical actions. Second, the dynamic prompt synthesizes contextual data via five XML tags (Neo4j data, activity configuration, manifestations of competencies and qualities, textbook content, and teacher information) and incorporates constraints based on the instructional setting (regular classroom versus a computer lab). The system utilizes the Gemini 3 Pro Preview model to generate pedagogical content that strictly adheres to the lesson plan framework mandated by Official Dispatch 5512 (MOET, 2020).

3.2.3. Response Parser and Self-Correction Mechanism

The JSON output from the LLM is not always syntactically valid; for instance, the model may emit invalid escape characters or the response may be truncated when the content is too long. To handle these issues, the assistant implements a multi-strategy response parser that progressively falls back from standard JSON parsing to character-level repair, truncation recovery, and marker-based extraction, thereby ensuring a usable result in nearly all cases. After parsing, the assistant verifies the presence of the six mandatory sections. If any section is missing, a self-correction (retry) mechanism re-prompts the LLM with a summary of the generated content and a request to supplement the missing parts. This multi-strategy approach, combined with the self-correction mechanism, is designed to improve the successful parsing rate.

3.3. Web Application and Interactive Process

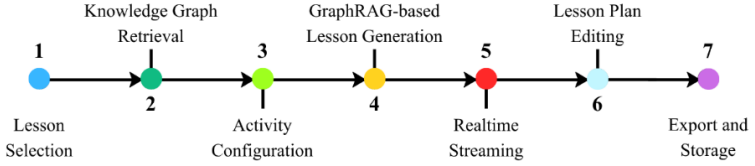

Figure 4. Lesson Planning Workflow

As illustrated in Figure 4, the lesson plan drafting workflow on the web-based platform (React/TypeScript) follows a tightly coupled interaction model. The teacher's lesson selection triggers the retrieval of the corresponding metadata and content indices from the knowledge graph. The teacher then configures the pedagogical scenario by allocating teaching methods, techniques,

and learning spaces to each instructional activity. Based on this configuration, the GraphRAG-based generation process is streamed in real time into a WYSIWYG editor for instant review and refinement. The system also provides auxiliary utilities: including interactive mind map generation, PDF export, persistent storage on PostgreSQL, and a classroom management module.

4. Assessment of GraphRAG-based virtual assistant

4.1. Evaluation Methodology

The quality evaluation was conducted through a survey involving 35 participants (n = 35): two University of education lecturers (5.7%), 25 high school teachers (71.4%), and eight K72 intern students (22.9%). By teaching seniority, 62.9% had over 10 years of experience, with the remainder at 5–10 years (8.6%), 3–5 years (5.7%), and under 3 years (22.9%) — a distribution that closely aligns with the system's target user base. Given the limited sample size, results serve as an initial feasibility study without sufficient statistical power to generalize to all high school Informatics teachers nationwide. The procedure involved three steps: (1) training on system operation; (2) each participant independently generating two lesson plans (one theory, and one practical); and (3) evaluating the plans on a five-point Likert scale (1 = strongly disagree, 5 = strongly agree). The evaluation form comprised 22 criteria divided in four groups, developed in accordance with the 2018 GEC (MOET, 2018a, 2018b), Official Dispatch 5512 (MOET, 2020), the Digital Competence Framework under Circular No. 02/2025/TT-BGDĐT (MOET, 2025a) and Official Dispatch 3456 (MOET, 2025b), and the Technology Acceptance Model (TAM) (Davis, 1989). Table 2 details the evaluation form.

Table 2. Quality evaluation form for lesson plans generated with the virtual assistant

Group	Code	Criteria
1. Accuracy of Objective Content	TC1	Knowledge objectives are accurate and ensure the required outcomes of the lesson.
	TC2	Informatics competence components NLa, NLb, NLc are identified appropriately for the lesson content.
	TC3	Informatics competence components NLd, NLe are identified appropriately for the organization of teaching activities.
	TC4	General competencies are selected appropriately, with specific manifestations linked to teaching activities.
	TC5	Essential qualities are selected appropriately, with specific manifestations linked to teaching activities.
	TC6	Digital competence manifestations (if any) are specifically tailored to the lesson content.
	TC7	The objectives of each activity are accurate and linked to the knowledge content formed in that activity.
	TC8	The overall objectives of the activities cover the general objectives of the lesson.
2. Compliance with Official Dispatch 5512 Structure	TC9	The lesson plan contains all four main sections: I. Objectives, II. Teaching Equipment and Materials, III. Teaching Process, IV. Appendices.
	TC10	The teaching process includes all four activities: warm-up, knowledge formation, practice, application.
	TC11	Each activity contains all four components: objectives, content, products, and implementation organization.
	TC12	The implementation organization section includes all four steps: task assignment; task execution; reporting and discussion; conclusion and assessment.
3. Quality of Content and Activity Design	TC13	Teaching equipment and materials are listed appropriately and feasibly.
	TC14	Knowledge formation content closely follows the textbook content.
	TC15	Practice questions cover the formed knowledge.
	TC16	Application questions are linked to real-world contexts.
	TC17	The organization of teaching activities is appropriate and capable of achieving the set objectives.

	TC18	The implementation organization (four steps) correctly reflects the characteristics of the teaching method and technique selected by the teacher.
	TC19	Study sheets / multiple-choice questions / programming exercises (if any) can be used in actual teaching.
4. Technology Acceptance (TAM)	TC20	The generated lesson plan can be used after minor edits.
	TC21	The system significantly reduces lesson plan preparation time compared to manual drafting.
	TC22	The system interface is intuitive and easy to use.

In addition to the 22-criterion survey, a within-subject comparative experiment was conducted on a separate sample of 20 high school teachers ($n = 20$), to compare lesson plans generated by the proposed system with those produced by an LLM without knowledge graph integration (referred to as a standalone LLM). Each teacher prepared the same lesson using both tools in a randomized order, and subsequently self-assessed the outputs against four core criteria (Table 3) on a five-point Likert scale. The Wilcoxon signed-rank test ($\alpha = 0.05$) was employed because the data were paired and ordinal. The experimental procedure and results are detailed in Section 4.3.

Table 3. Core criteria for system comparative evaluation

Code	Core criteria	Evaluation standard specifications
SS1	Content accuracy	The generated lesson plan contains no false or fabricated information (i.e., hallucinations). Furthermore, the knowledge objectives and digital competence indicators strictly align with the achievement standards of the 2018 General Education Curriculum, Circular No. 02/2025/TT-BGDĐT, and Official Dispatch No. 3456.
SS2	Pedagogical structure	The lesson generated plans exhibit a clear hierarchy, strictly adhering to the four-activity process (including Warm-up, Knowledge Formation, Practice, and Application). Each activity incorporates all four mandatory components (Objectives, Content, Products, and Organization of Implementation), with the latter component being fully deployed through four-steps implementation process as stipulated by Official Dispatch No. 5512.
SS3	Logical consistency	Ensures a logical and seamless connection among the components: lesson objectives, teaching content, expected teaching equipment and materials, and assessment methods.
SS4	Language and usability	The generated lesson plan utilizes standard, coherent pedagogical language, free from redundant, or robotic AI expressions. Consequently, the output achieves a high degree of completeness, thereby requiring minimal effort and time from teachers to refine it for practical teaching application.

4.2. Results and Discussion

Table 4 presents the mean and standard deviation (SD) for each criterion, while table 5 summarizes the results across the four criteria groups. The overall mean reached $M = 4.52$ ($SD = 0.60$), with all 22 criteria scoring between 4.29 and 4.69 — which indicates that the system meets requirements across all dimensions.

Table 4. Evaluation results by individual criteria (n = 35)

Code	**TC1**	**TC2**	**TC3**	**TC4**	**TC5**	**TC6**	**TC7**	**TC8**	**TC9**	**TC10**	**TC11**
Mean	4.57	4.54	4.43	4.60	4.51	4.29	4.54	4.54	4.63	4.60	4.66
SD	0.56	0.66	0.61	0.55	0.61	0.79	0.61	0.56	0.49	0.60	0.54
Code	**TC12**	**TC13**	**TC14**	**TC15**	**TC16**	**TC17**	**TC18**	**TC19**	**TC20**	**TC21**	**TC22**
Mean	4.69	4.37	4.46	4.49	4.46	4.54	4.34	4.40	4.43	4.66	4.60
SD	0.47	0.77	0.61	0.56	0.61	0.61	0.76	0.60	0.61	0.48	0.55

Group 1: Accuracy of Objective Content ($M = 4.50$; $SD = 0.62$). The score for seven criteria (excluding TC6) ranged from 4.43–4.60, with criterion TC4 (general competencies) scoring highest, confirming that verbatim extraction of data from the knowledge graph works effectively. The only exception criterion is TC6 (manifestations of digital competence), which receives the lowest in the entire survey ($M = 4.29$; $SD = 0.79$). Although the system supplies the indicator

descriptions to the LLM along with two worked examples, the digital competence framework comprises six domains with 55 indicators; two examples are insufficient for the model to generalize effectively, producing outputs that occasionally lack specificity.

Group 2: Compliance with Official Dispatch 5512 Structure (M = 4.65; SD = 0.53). This category achieves the highest mean score among the four groups. Specifically, criterion TC12 (organization of implementation encompassing with all four steps) reached M = 4.69 (SD = 0.47), representing the top score and lowest variance overall. This finding validates the effectiveness of two technical solutions: enforcing the six mandatory sections via structured JSON output, and deploying the automatic retry mechanism to supplement any missing components. Furthermore, the gap between the mean score of Group 2 and Group 3 (ΔM = 0.21) demonstrates that while the system effectively controls structural completeness through the immutable zone, the creative content with the creative zone exhibits greater variability, which is fully consistent with the proposed two-layer control mechanism.

Group 3: Quality of Content and Activity Design (M = 4.44; SD = 0.65). This category exhibited the lowest mean score and the highest variance among the four groups. Within this group, criterion TC18 (reflecting the characteristics of the selected teaching method) scored the lowest (M = 4.34; SD = 0.76). This lower score is attributed to the fact that knowledge graph currently stores only generic method names and general procedures, rather than granular, method-specific steps that would enable the LLM to differentiate among them more precisely. Similarly, criterion TC13 (appropriate allocation of teaching equipment, M = 4.37; SD = 0.77) also showed high substantial variability, which reflects the structural differences in infrastructure across different schools. However, criterion TC17 (appropriateness of activity organization, M = 4.54) indicates that overall organization of instructional activity remains highly feasible, despite the current limitations of granular method differentiation.

Group 4: Technology Acceptance (TAM) (M = 4.56; SD = 0.55). With this category, criterion TC21 (reducing lesson preparation time) reached M = 4.66 (SD = 0.48), which aligns closely with the perceived usefulness (PU) construct of the Technology Acceptance of Model (TAM) (Davis, 1989). Conversely, criterion TC20 (usability after minor editing) scored the lowest in this group (M = 4.43; SD = 0.61). This finding indicates that while teachers view the generated output as a high-quality draft that requires further refinement rather than a finished product, it directly corroborates the design orientation of the system: the teacher retains ultimate authority over all pedagogical content.

Table 5. Summary of evaluation results by four criteria groups (n = 35)

Group	Criteria	Mean	SD
1	Accuracy of Objective Content (TC1–TC8)	4.50	0.62
2	Compliance with Official Dispatch 5512 Structure (TC9–TC12)	4.65	0.53
3	Quality of Content and Activity Design (TC13–TC19)	4.44	0.65
4	Technology Acceptance – TAM (TC20–TC22)	4.56	0.55
Overall Average		**4.52**	**0.60**

Note: Mean (5-point scale); SD = Standard Deviation.

The two university lecturers scored notably lower overall (4.00) compared to the high school teachers (4.56) and student interns (4.52). The largest discrepancies were observed in Group 1 (3.88 vs. 4.55) and Group 3 (3.93 vs. 4.47), which reflects the stricter evaluation standards imposed by experts with deep pedagogical backgrounds. However, given the limited sample of only two experts, these findings indicate a trend rather than a statistically robust conclusion. Conversely, high school teachers and student interns scored similarly (4.56 vs. 4.52 respectively), demonstrating a consistent acceptance of the system among those directly engaged in classroom instruction, regardless of their teaching seniority (Table 6).

Table 6. Mean scores (5-point scale) classified by professional role

Evaluation group	n	n1	n2	n3	n4	Overall
University of education lecturers (professionals)	2	3.88	4.50	3.93	3.83	4.00
High school teachers	25	4.55	4.66	4.47	4.63	4.56
K72 intern students	8	4.52	4.62	4.45	4.54	4.52

The overall scale yielded α = 0.898, suggesting adequate shared variance among the 22 criteria, though this high value is partly driven by the substantial item count rather than strict unidimensionality (Taber, 2018). Specifically, Groups 1 to3 each exceeded α = 0.79; Group 4, which comprises only three criteria, returned $\alpha = 0.636$. As Taber (2018) notes, the commonly 0.70 threshold is merely heuristic, and Cronbach's alpha tend to underestimate internal consistency when a scale contains few items — therefore, Group 4's lower coefficient is fully interpretable within the context of its limited item count (Table 7). The evaluation results highlight criteria TC6 and TC18 as the two metrics most in need of improvement; corresponding remedial strategies are discussed in Section 5.

Table 7. Cronbach's Alpha by criteria group and overall scale (n = 35)

Group	Criteria	k	α
1	Accuracy of Objective Content (TC1 - TC8)	8	0.838
2	Compliance with Official Dispatch 5512 Structure (TC9 -TC12)	4	0.875
3	Quality of Content and Activity Design (TC13 -TC19)	7	0.797
4	Technology Acceptance - TAM (TC20 -TC22)	3	0.636
Overall scale		22	0.898

4.3. Comparison with an LLM Without Knowledge Graph Integration

The comparative experiment aimed to quantify the improvement contributed by the knowledge graph, the two-layer control mechanism, and the JSON format enforcement compared to a standalone LLM. Following the procedure outlined in Section 4.1, 20 high school teachers prepared the same lesson using a standalone LLM and the proposed system in a randomized order, and subsequently self-assessed both lesson plans on the four core criteria (SS1–SS4) using a five-point Likert scale. Table 8 reports the Wilcoxon signed-rank test results for each criterion.

Across all criteria, the proposed system significantly outperformed the standalone LLM (overall $M = 4.19$ versus 2.99), with 80% of teachers rating it higher on average. As shown in Table 8, the most substantial improvements were observed in Pedagogical structure (SS2) and Language and usability (SS4), both reaching very large effect sizes ($r \geq 0.79$). This finding demonstrates that the two-layer control mechanism and JSON format enforcement effectively compel adherence to Official Dispatch 5512 and improve pedagogical language quality. Content accuracy (SS1) and Logical consistency (SS3) also showed significant improvements with large effect sizes. While these latter criteria still fundamentally rely on the intrinsic reasoning capabilities of the LLM shared by both tools, the knowledge graph crucially anchors the generated content by accurately supplying knowledge objectives and digital competence indicators, thereby maintaining better consistency and reducing hallucinations compared to the standalone LLM.

Table 8. Comparison results between the proposed system and the standalone LLM (n = 20)

Criteria	Standalone LLM M (SD)	The System M (SD)	Z	p	R
SS1 – Content accuracy	3.25 (0.91)	4.20 (0.83)	–2.611	0.007	0.58
SS2 – Pedagogical structure	2.85 (1.04)	4.25 (0.72)	–3.516	< 0.001	0.79
SS3 – Logical consistency	2.85 (1.14)	3.95 (0.94)	–2.482	0.011	0.55
SS4 – Language and usability	3.00 (0.92)	4.35 (0.59)	–3.621	< 0.001	0.81
Overall mean	**2.99 (1.00)**	**4.19 (0.78)**			

Note: Z = Wilcoxon signed-rank test statistic; p = exact two-tailed significance; r = effect size ($r = |Z| / \sqrt{N}$).

5. Conclusions

This study successfully developed a GraphRAG-based virtual assistant featuring a two-layer control mechanism to assist high school informatics teachers in generating lesson plans that comply with Official Dispatch 5512 (MOET, 2020) and current digital competence standards, specifically Circular No. 02/2025/TT-BGDĐT (MOET, 2025a) and Official Dispatch No. 3456 (MOET, 2025b). The evaluation results from the 35 participants affirmed the system's high feasibility across content accuracy, structural compliance, activity design, and user acceptance, thereby proving its value as a support tool that preserves teachers' professional autonomy. Furthermore, a comparative experiment (n = 20) demonstrated that the system significantly outperformed a standalone LLM across all core criteria ($p < 0.05$), especially in pedagogical structure and language quality.

Several limitations remain to be addressed. First, the evaluation relies primarily on self-reporting with a modest sample size (n = 35); consequently, a large-scale, blind comparative experiment between the AI-generated lesson plans and those prepared by expert teachers is needed to robustly affirm the system's pedagogical effectiveness. Second, the quality of the creative zone remains constrained by the LLM's intrinsic reasoning capability: specifically, concretizing digital competence indicators (TC6) and reflecting teaching method characteristics (TC18) achieved the lowest scores, indicating that the domain examples in the current prompt are insufficient for effective generalization. Third, the generated supplementary resources currently support only multiple-choice quizzes and essay-style study sheets, and the system exclusively operates on individual lessons, lacking support for project-based teaching and long-term unit planning.

Future work will focus on three areas: (1) adding dynamic domain examples to concretize digital competence indicators and describe the distinct characteristics of each teaching method, thereby directly addressing the previously identified TC6 and TC18 limitations above; (2) developing a semi-automated knowledge graph construction process to accelerate the system's expansion to other academic subjects – primarily natural sciences – in order to facilitate interdisciplinary STEM lesson plans; and (3) supporting project-based teaching and diversifying the supplementary resource formats.

Notes for contributors: *Nguyen Hong Phong and Do Minh Quan are undergraduate students, and MSc Le Xuan Hien and Assoc.Prof. Dr Pham Tho Hoan are lecturers at the School of Mathematics and Computer Science, Hanoi National University of Education, Vietnam. Author 1: conceptualization, methodology, software, data collection, visualization, writing original draft; author 2: translation, writing -review and editing; author 3: conceptualization, development orientation, funding acquisition, evaluation and editing support; author 4: supervision, writing -review and editing.*

Conflicts of interest: *The authors declare no conflicts of interest.*

Acknowledgments: *The authors would like to thank the educational experts, high school teachers, and student interns for their valuable participation in the evaluation process.*

REFERENCES

Abu-Salih, B., & Alotaibi, S. (2024). A systematic literature review of knowledge graph construction and application in education. *Heliyon, 10*(3), e25383. https://doi.org/10.1016/j.heliyon.2024.e25383

Baidoo-Anu, D., & Owusu Ansah, L. (2023). Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. *Journal of AI, 7*(1), 52–62. https://doi.org/10.61969/jai.1337500

Biggs, J. (1996). Enhancing teaching through constructive alignment. *Higher Education, 32*(3), 347–364. https://doi.org/10.1007/BF00138871

Celik, I. (2023). Towards Intelligent-TPACK: An empirical study on teachers' professional knowledge to ethically integrate artificial intelligence (AI)-based tools into education. *Computers in Human Behavior, 138,* 107468. https://doi.org/10.1016/j.chb.2022.107468

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly, 13*(3), 319–340. https://doi.org/10.2307/249008

Edge, D., Trinh, H., Cheng, N., Bradley, J., Chao, A., Mody, A., Truitt, S., Metropolitansky, D., Ness, R. O., & Larson, J. (2024). *From local to global: A Graph RAG approach to query-focused summarization.* arXiv. https://doi.org/10.48550/arXiv.2404.16130

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., & Wang, H. (2024). *Retrieval-Augmented Generation for large language models: A survey.* arXiv. https://doi.org/10.48550/arXiv.2312.10997

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems, 43*(2), Article 42. https://doi.org/10.1145/3703155

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., … Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103,* 102274. https://doi.org/10.1016/j.lindif.2023.102274

Ministry of Education and Training. (2018a). *Chương trình giáo dục phổ thông môn Tin học* [General Education Curriculum for Informatics] (Circular No. 32/2018/TT-BGDĐT, December 26, 2018).

Ministry of Education and Training. (2018b). *Chương trình giáo dục phổ thông – Chương trình tổng thể* [General Education Curriculum – Overall Program] (Circular No. 32/2018/TT-BGDĐT, December 26, 2018).

Ministry of Education and Training. (2020). *Công văn số 5512/BGDĐT-GDTrH về việc xây dựng và tổ chức thực hiện kế hoạch giáo dục của nhà trường* [Official Dispatch No. 5512 on school education plan development] (December 18, 2020).

Ministry of Education and Training. *(2025a). Khung năng lực số cho người học* [Digital Competence Framework for learners] (Circular No. 02/2025/TT-BGDĐT, January 24, 2025).

Ministry of Education and Training. (2025b). *Hướng dẫn triển khai thực hiện Khung năng lực số cho học sinh phổ thông và học viên giáo dục thường xuyên* [Guidance on implementing the Digital Competence Framework for school students and continuing education learners] (Official Dispatch No. 3456/BGDĐT-GDPT, June 27, 2025).

Mishra, P., & Koehler, M. J. (2006). Technological pedagogical content knowledge: A framework for teacher knowledge. *Teachers College Record, 108*(6), 1017–1054. https://doi.org/10.1111/j.1467-9620.2006.00684.x

Ramírez, S. (2018). *FastAPI: Modern, fast web framework for building APIs with Python.* https://fastapi.tiangolo.com/

Taber, K. S. (2018). The use of Cronbach's alpha when developing and reporting research instruments in science education. *Research in Science Education, 48*(6), 1273–1296. https://doi.org/10.1007/s11165-016-9602-2

Wiggins, G., & McTighe, J. (2005). *Understanding by design (Expanded 2nd ed.). Association for Supervision and Curriculum Development*.

Zheng, Y., Huang, S., Zeng, X., Huang, Y., Liu, Z., & Luo, W. (2025). Knowledge-enhanced large language models for automatic lesson plan generation. *Humanities and Social Sciences Communications, 12,* 1784. https://doi.org/10.1057/s41599-025-06004-2