

A Design-Science Proof-of-Concept for Constraining LLM Pedagogical Decisions Using Knowledge Graphs

Susana Ayu Handayani

Yogyakarta State University

Priyo Nugroho Adi

`priyo_nugroho.adi@tu-dresden.de`

TU Dresden

Thomas Köhler

TU Dresden

Mochamad Bruri Triyono

Yogyakarta State University

Priyanto Priyanto

Yogyakarta State University

Siswantoyo Siswantoyo

Yogyakarta State University

Rita Fransina Maruanaya

TU Dresden

Article

Keywords: Large Language Models (LLMs), Knowledge Graph Grounding, Neuro-Symbolic Artificial Intelligence, Explainable Educational AI, Pedagogical Hallucination Mitigation, Trustworthy, Constraint-Based Generative AI

Posted Date: October 12th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11046385/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Artificial intelligence–driven educational systems increasingly rely on Large Language Models (LLMs) to deliver adaptive recommendations, personalized learning pathways, and automated pedagogical support. However, despite their linguistic fluency and contextual reasoning capabilities, LLMs remain vulnerable to pedagogical hallucinations, including logically invalid prerequisite recommendations, over-prescriptive interventions, and non-verifiable instructional guidance. These limitations pose substantial risks for educational trustworthiness, learner autonomy, and explainable decision-making within gamified learning environments. This study investigates whether symbolic grounding through Knowledge Graphs (KGs) can constrain the generative action space of LLMs to produce logically verifiable and pedagogically valid outputs. Adopting a Design Science Research (DSR) paradigm, the study develops a controlled proof-of-concept artifact that integrates explicit pedagogical semantic triples representing learner states, occupational prerequisites, and educational interventions. A quasi-experimental comparative evaluation was conducted across five state-of-the-art LLMs (GPT-5.3, Sonnet 4.6, Gemini 3.1 Pro, Mistral Large 2, and DeepSeek) under ungrounded and KG-grounded conditions. The findings demonstrate that KG grounding systematically eliminates over-prescriptive recommendations, enforces minimal intervention compliance, and improves logical action validity and explicit reasoning traceability across all evaluated models. Rather than functioning as post-generation hallucination filtering, the proposed approach proactively restricts invalid pedagogical actions through symbolic constraint structures embedded within the reasoning process itself. The study contributes empirical evidence that neuro-symbolic integration can improve explainability, determinism, and trustworthiness in educational AI systems while preserving adaptive personalization capabilities. Although deliberately bounded as a proof-of-concept rather than a production-scale educational platform, the findings establish foundational support for knowledge-grounded generative AI architectures in high-stakes domains requiring transparent and logically constrained decision-making.

Introduction

Large Language Models have emerged as transformative forces in computational linguistics, demonstrating unprecedented capabilities in contextual reasoning, semantic understanding, and adaptive knowledge synthesis. These models now operate across diverse domains from clinical decision support to educational tutoring, where their fluency and apparent understanding inspire widespread deployment. Yet this apparent intelligence masks a critical vulnerability: the systematic generation of fluent but factually unsupported content, a phenomenon termed hallucination. In educational contexts, this vulnerability becomes acute. Educators and learners entrust AI-driven tutoring systems with high-stakes pedagogical decisions, course recommendations, prerequisite identification, and learning path optimization, where factual unreliability threatens instructional validity and student trust.

The challenge intensifies within gamified learning ecosystems, where adaptive personalization has become a core design principle. Modern educational gamification leverages AI to dynamically adjust

difficulty, provide real-time feedback, and tailor interventions to individual learner profiles. Such systems promise equitable access to personalized education while avoiding the "one-size-fits-all" constraints of traditional curricula. However, contemporary LLM-based educational systems operate without explicit guardrails. They generate recommendations based on statistical probability learned from training corpora, not from verified instructional knowledge. An ungrounded LLM tasked with recommending learning interventions may suggest non-existent courses, misidentify prerequisite chains, or prescribe over-interventionist advice when no action is warranted, outcomes that undermine both pedagogical validity and learner autonomy.

This problem is fundamentally structural, not merely a training data issue. Recent comprehensive surveys highlight the multidimensional nature of LLM hallucination, identifying it as a persistent challenge across the entire model development lifecycle, from data collection and architecture design through inference stages. Alansari and Luqman (2026) demonstrate that hallucinations arise not from factual memory failures but from the inherent tendency of neural systems to generate fluent continuations unconstrained by external truth conditions. Similarly, investigations into fairness in natural language processing reveal that LLMs encode and amplify societal biases unless explicitly constrained through structured knowledge frameworks. Zadid et al. (2026) argue that achieving fairness in NLP requires integration of socio-technical approaches that combine technical interventions with community participation and transparent governance.

Parallel to these challenges, knowledge graphs have emerged as formal mechanisms for representing structured domain knowledge. Unlike unstructured training data, knowledge graphs encode explicit semantic relationships through ontological triples: subject-predicate-object statements that capture the logical dependencies underlying complex systems. In medical decision support, Ray (2026) documents how knowledge graphs enable lightweight LLM applications on resource-constrained IoT devices by providing grounded reasoning pathways that reduce the model's computational burden. In industrial applications, Yu et al. (2026) demonstrate that hybrid architectures combining LLMs with retrieval-augmented generation (RAG) and knowledge-based reasoning significantly improve task-specific performance and model interpretability.

The pedagogical domain presents a unique opportunity to test this hypothesis. Educational systems operate within relatively well-defined logical constraints: prerequisite hierarchies follow deterministic rules, learning outcomes map systematically to capability levels (e.g., CEFR language proficiency frameworks), and intervention effectiveness is measurable against defined objectives. Unlike open-ended knowledge domains where "correct" answers are contested, educational decision-making can be grounded in explicit logical rules derivable from pedagogical science and curriculum design.

Yet despite growing evidence that structured knowledge constrains generative systems, a critical gap persists in the literature. Existing reviews of LLM applications in education, such as the comprehensive study by Adi et al. (2026a), emphasize adaptive personalization and engagement enhancement but largely sidestep the epistemic question: How can educators ensure that AI-driven pedagogical

recommendations are not merely fluent but logically valid and grounded in verified instructional knowledge? Contemporary studies of multimodal LLMs in medical decision support (Warkiani et al., 2026) and AI-enhanced pharmaceutical systems (Chen et al., 2026) demonstrate that domain-specific knowledge integration significantly improves output reliability. Yet comparable investigations in educational AI remain limited.

Furthermore, the intersection of knowledge graphs, LLM grounding, and gamified learning has not been systematically examined as an integrated research problem. While Park et al. (2026) document BIM-LLM integration in construction workflows and identify validation gaps, and while Al-Amin et al. (2026) trace the historical evolution of chatbots through increasingly sophisticated conversational paradigms, no prior work explicitly tests whether symbolic knowledge representation can enforce hard logical constraints on LLM pedagogical outputs within a controlled experimental framework.

This study addresses this gap through a deliberately scoped design-science proof-of-concept. Rather than proposing a production-ready tutoring system or measuring learning outcomes, we isolate a specific technical hypothesis: Can explicit knowledge graph structures restrict the generative action space of LLMs such that pedagogical outputs become logically verifiable and deterministically valid? By comparing ungrounded baseline LLM behavior against a knowledge-graph-constrained prototype, we provide empirical evidence of constraint-induced reasoning correction. This approach aligns with established design-science research principles articulated by Hevner et al. (2004), which emphasize building and evaluating artifacts to extend human and organizational capabilities through rigorous, bounded experimentation.

The contribution of this work is explicitly modest and carefully bounded: We demonstrate that knowledge graph grounding prevents over-prescriptive pedagogical hallucinations and enforces minimal intervention compliance. We do not claim that this approach scales to complex, real-world educational ecosystems. We do not measure learner outcomes or claim pedagogical superiority. Instead, we establish a foundational proof-of-concept that symbolic knowledge representation can function as a corrective mechanism for unconstrained neural generation, a finding with implications extending beyond education to any high-stakes domain where AI recommendations must be both fluent and logically sound.

The structure of this manuscript follows this progression: the literature review synthesizes current knowledge across the constituent domains of knowledge graphs, LLM applications, educational gamification, and explainable AI, identifying specific methodological and conceptual gaps. The methodology section describes the design science research paradigm and the construction of both baseline and prototype artifacts. The results present empirical findings from quasi-experimental scenario testing. The discussion contextualizes these findings within broader questions of AI trustworthiness and pedagogical validity. Finally, the conclusion outlines implications for future research in symbolic-neural hybrid systems and adaptive knowledge representation.

1.1. Scope and Contribution Clarification

This study presents a controlled design-science proof-of-concept that investigates how symbolic scaffolding via a pedagogical Knowledge Graph (KG) can impose strict logical constraints on Large Language Model (LLM) outputs in gamified educational scenarios. Rather than proposing a production-ready tutoring system or evaluating learner outcomes, the study deliberately isolates the decision-making behavior of LLMs under constrained and unconstrained conditions. The contribution lies in empirically demonstrating how explicit knowledge structures can restrict the LLM action space, thereby preventing over-prescriptive and logically invalid pedagogical recommendations under well-defined instructional rules.

Literature Review

This section synthesizes current knowledge across knowledge graphs, LLM grounding mechanisms, educational AI, and explainable AI systems, organizing research into integrated thematic areas rather than sequential paper summaries. This synthetic approach identifies specific research gaps and positions the present work within a coherent scholarly landscape.

2.1. Knowledge Graphs, Graph Neural Networks, and Symbolic Reasoning

Knowledge graphs represent a fundamental shift in how artificial intelligence systems encode and reason for structured knowledge. Unlike the implicit, distributed representations learned by neural networks, knowledge graphs materialize domain expertise as explicit semantic relationships: ontological triples that capture causality, hierarchy, and constraint. Saeidi et al. (2026) provide a comprehensive survey of graph convolutional networks (GCNs), demonstrating the integration of GCNs across eight advanced learning paradigms, including recurrent, attention-based, and reinforcement learning architectures. Their analysis reveals a decisive architectural transition: from computationally constrained spectral GCNs toward scalable spatial GCNs that enable inductive reasoning on unseen data. Critically, they identify interpretability enhancement and resolution of the over-smoothing problem in deep GCNs as key future directions, suggesting that interpretability, not merely predictive accuracy, has become a priority in graph-based learning.

Beyond computer vision and network analysis, graph intelligence has proven instrumental in cybersecurity and critical infrastructure monitoring. Ahmad et al. (2026) examine how graph-based representations capture the relational dependencies underlying IoT and Cyber-Physical System security, moving beyond reactive network intrusion detection toward physics-aware hybrid learning and trustworthy causal reasoning. Their framework is particularly instructive: by representing system states and their transitions graphically, security systems achieve both improved detection performance and enhanced human-interpretable explanations. This integration of symbolic representation with neural

learning demonstrates a generalizable principle that explicit structure constrains generative systems toward more reliable outputs.

2.2. Large Language Models: Capabilities, Limitations, and the Hallucination Challenge

Large Language Models have achieved remarkable capabilities in natural language understanding, semantic reasoning, and knowledge synthesis. Yet these capabilities coexist with systematic vulnerabilities. Alansari and Luqman (2026) provide the most comprehensive contemporary analysis of LLM hallucination, identifying it not as a surface phenomenon but as a deep architectural issue spanning the entire model development lifecycle. Their taxonomy distinguishes hallucination types, analyzes root causes from data collection through inference, and systematically categorizes detection and mitigation strategies. Critically, they demonstrate that hallucination arises from the fundamental objective of neural language models, maximizing the likelihood of coherent text, which can diverge fundamentally from maximizing factual accuracy. This insight suggests that post-generation filtering alone cannot solve the problem; instead, architectural constraints are necessary.

Ray (2026) examines LLM integration into resource-constrained IoT ecosystems, synthesizing evidence that lightweight LLMs (sub-billion to 2-billion parameter models) can achieve comparable performance to larger models when combined with knowledge-based retrieval and explicit structural constraints. This finding extends the hallucination literature: even highly capable models become more reliable when their reasoning is scaffolded by external knowledge structures. In industrial domains, Yu et al. (2026) document how LLMs function across the product lifecycle through a systematic combination of prompt engineering, retrieval-augmented generation, and parameter-efficient fine-tuning. Their assessment emphasizes that effectiveness depends critically on domain-specific knowledge integration and standardized evaluation frameworks—suggesting that generic LLMs require domain-grounding to function effectively in specialized contexts.

2.3. Educational AI, Gamification, and Adaptive Personalization

Educational technology has undergone a paradigm shift toward AI-driven personalization and adaptive learning. Adi et al. (2026a) provides a comprehensive integrative review examining how artificial intelligence, particularly generative AI, can correct persistent limitations in educational gamification. They identify that while gamification successfully drives engagement, it often fosters extrinsic reward dependency and inequitable learning experiences. Critically, they propose an "AI Corrective Role Framework" wherein AI functions not as a content generator but as a learner-centered mechanism that personalizes difficulty, replaces superficial rewards with intelligent feedback, and enhances inclusivity.

This framework suggests that pedagogically appropriate AI must be both adaptive and constrained, which is responsive to individual needs yet grounded in principled educational logic.

The present study builds directly on this theoretical foundation. Adi et al. (2026a) argue that AI-enhanced gamification requires integration of learner profile data, adaptive algorithms, and equitable design principles. They emphasize that AI's role in education should enhance educator judgment, not replace it, and should preserve transparency such that learners and educators understand the basis for AI recommendations. These principles align precisely with the knowledge-graph-rounding approach: by making pedagogical constraints explicit, symbolic scaffolding preserves the interpretability and pedagogical validity that Adi et al. identify as essential to trustworthy educational AI.

Methodology

This research adopts the design science research (DSR) paradigm as articulated by Hevner et al. (2004) and operationalized through the process framework of Peffers et al. (2007). Design science is fundamentally concerned with the creation of artifacts, including innovative technologies, systems, processes, or frameworks that extend human and organizational capabilities. Hevner et al. emphasizes that design-science research maintains rigor not through population-level statistical inference but through careful specification of problem domains, deliberate artifact construction, and controlled evaluation protocols. This paradigm is particularly appropriate for investigating how explicit knowledge structures can constrain generative AI systems, as it permits isolation of specific behavioral mechanisms while maintaining scholarly depth.

The research problem is precisely scoped: Given that ungrounded LLMs exhibit systematic tendencies toward over-prescriptive recommendations and hallucinated information in pedagogical contexts, can explicit knowledge graph structures restrict the LLM action space such that outputs become logically valid under defined instructional rules?

This problem sits at the intersection of artificial intelligence system design and educational pedagogy. A junction where established design-science methods have proven effective. Recent comprehensive reviews of LLM applications in industrial domains (Yu et al., 2026) employ similar methodological approaches, systematically comparing task performance across adapted LLM variants with and without domain knowledge integration.

Primary Research Question (PRQ): How can Knowledge Graphs be used to ground Large Language Models to produce verifiable, culturally-aware, and pedagogically adaptive outputs in gamified learning environments?

To systematically evaluate the PRQ, the study breaks down the investigation into dual operational sub-questions, tested through a comparative analysis between a baseline ungrounded LLM and a KG-grounded prototype:

RQ1 (Verifiability & Constraint-induced prevention of logically invalid pedagogical actions): To what extent does grounding an LLM in a Knowledge Graph reduce pedagogical hallucinations compared to an unguided baseline model when identifying prerequisite learning paths?

RQ2 (Pedagogical Adaptivity): How effectively can a KG-grounded LLM utilize dynamic access control logic to recommend precise, individualized interventions based on a learner's specific capability deficits and target goals

3.1. Literature Extraction and Artifact Foundations

The foundational rules, pedagogical constraints, and technical parameters of the proposed artifact were derived through a systematic retrieval of high-cited literature. A targeted extraction was conducted on the ScienceDirect database to capture the intersection of graph-based reasoning, educational gamification, and AI personalization. The exact Boolean search string utilized was:

("knowledge graph" OR "knowledge graphs") AND ("education" OR "learning system" OR "e-learning" OR "gamification") AND ("artificial intelligence" OR "AI" OR "adaptive learning") AND ("framework" OR "model" OR "representation") AND ("personalization" OR "learning analytics") AND PUBYEAR > 2020 AND PUBYEAR < 2027 AND (LIMIT-TO (DOCTYPE , "ar")) AND (LIMIT-TO (LANGUAGE , "English"))

To ensure the artifact was built upon the most validated paradigms, the corpus was limited to the top 50 most-cited articles fitting these criteria. Subsequently, a critical appraisal process was performed to evaluate the relevance of each article based on title alignment, study focus, and contribution to the objectives of the present study. Following this screening process, 20 articles with moderate to high relevance were selected for inclusion in the review. In addition, two foundational references on Design Science Research (DSR) and two previous published works authored by the researcher were included as foundational and supporting references due to their direct relevance to the conceptual framework and research context of the study. This synthesis directly informed the logical boundaries of the experimental prototype.

3.2. Artifact Design: Baseline vs. Prototype Architecture

The core objective of the DSR artifact is to restrict the generative capabilities of a Large Language Model (LLM) within the hard logical boundaries of a Knowledge Graph (KG). To isolate and measure this effect, two parallel models were established for testing:

1. The Baseline Model (Ungrounded LLM): The baseline represents a standard generative AI acting as a gamified tutor. It is provided with a student's goal and current capability but lacks structural knowledge constraints. For example, the Baseline is simply prompted with: "Goal: Become Data Analyst. Student English Level: B1." This serves as the control to measure the frequency of generic advice or algorithmic hallucination [3].

2. The Prototype Model (KG-Grounded LLM): The proposed artifact injects an explicit semantic ontology into the LLM's prompt window. The Knowledge Graph is constructed using semantic triples (Subject, Predicate, Object) that map the educational ecosystem across three entity classes: Learners, Target Occupations, and Educational Interventions.

The Prototype encompasses the following rigid ontological triples:

Learner States:

(Student_A, hasLevel, B1_English)

(Student_B, hasLevel, A1_English)

(Student_C, hasLevel, A2_English)

Occupational Prerequisites (Target Goals):

(Job_Data_Analyst, requiresLevel, B2_English)

(Job_Data_Entry, requiresLevel, B1_English)

(Job_Machine_Operator, requiresLevel, A2_English)

(Job_CleaningStaff, requiresLevel, A1_English)

Educational Interventions (The "Corrective" Path):

(Course_Intermediate, deliversLevel, B2_English)

(Course_Grounded, deliversLevel, A2_English)

(Course_Hoop, deliversLevel, B1_English)

By structuring the data in this manner, the system operationalizes the dynamic access control concept. The LLM is instructed to act as the reasoning engine, but its outputs are strictly engineered to force semantic grounding.

It is important to note that the baseline models were not provided with explicit sufficiency or stopping condition (e.g., instruction to recommend no action when requirements are already met). This reflects common real-world LLM tutoring deployments, which are typically optimized for helpfulness and content elaboration. In contrast, the KG-grounded artifact implicitly encodes sufficiency as a logical constraint, enforcing minimal intervention without requiring explicit natural-language instruction.

3.3. Evaluation Protocol: Quasi-Experimental Comparative Testing

To validate the verifiability and pedagogical adaptivity of the artifact, a quasi-experimental, within-model comparative evaluation was executed. This approach isolates the LLM's cognitive behavior by feeding both the Baseline and the Prototype varying learner states and career goals through structured prompt scenarios.

The evaluation tests the system's ability to act as an accurate pedagogical engine without hallucinating non-existent courses or misinterpreting prerequisites. Three distinct scenarios were executed:

Scenario 1 (The Upskilling Path):

Prompt: "Question: What should Student_A learn to qualify for Job_Data_Analyst? Answer must: - reference the triples - explain reasoning step-by-step."

Testing Objective: Evaluates the system's ability to calculate the gap between current state (B1) and required state (B2), and accurately recommend the exact intervention (Course_Intermediate).

Scenario 2 (The Direct Match):

Prompt: "Question: What should Student_B learn to qualify for Job_CleaningStaff? Answer must: - reference the triples - explain reasoning step-by-step."

Testing Objective: Tests the system's ability to recognize when a student's current level (A1) exactly meets the baseline requirement (A1), thereby preventing unnecessary gamified "grinding" or redundant course recommendations.

Scenario 3 (The Downward/Lateral Match):

Prompt: "Question: What should Student_C learn to qualify for Job_CleaningStaff? Answer must: - reference the triples - explain reasoning step-by-step."

Testing Objective: Assesses the model's logical consistency when a student's level (A2) already exceeds the baseline requirement (A1). A trustworthy model must deduce that no further coursework is required.

The responses generated by both the Baseline and the Prototype are evaluated against a deterministic logical truth derived mathematically from the KG triples. This setup allows for a precise measurement of how effectively the Knowledge Graph constrains generative AI, answering the primary research question by providing empirical indication whether the resulting learning paths are structurally sound, personalized, and logically verifiable.

Results

The evaluation phase assessed the ability of five leading LLMs: OpenAI GPT-5.3, Anthropic Sonnet 4.6, Gemini 3.1 Pro, Mistral Large 2, and DeepSeek to navigate complex pedagogical decision-making. By comparing the Baseline (Ungrounded) performance against the Prototype (KG-Grounded) architecture, we isolate the specific corrective role that structured knowledge plays in constraint-induced prevention of logically invalid pedagogical actions and ensuring personalized learning path accuracy. Importantly, the KG-grounded architecture does not detect hallucinations post-generation. Instead, it prevents the generation of logically invalid actions by proactively restricting the LLM’s output space. This distinction reframes hallucination mitigation from a reactive filtering problem into a structural design, aligning symbolic knowledge representation with generative reasoning.

4.1. Evaluation Metrics

Because this study focuses on logical compliance rather than open-ended correctness, standard predictive accuracy metrics are not suitable. Instead, system behavior was evaluated using three logic-aware design-science metrics, defined as follows:

M1. Logical Action Validity (LAV)

The system produces exactly one logically valid instructional action derivable from the Knowledge Graph under the given learner–goal state.

M2. Minimal Intervention Compliance (MIC)

The system recommends no instructional action when the learner state meets or exceeds the goal requirement.

M3. Explicit Traceability (ET)

The system explicitly cites Knowledge Graph triples used in the reasoning process.

Quantitative Performance and Logical Accuracy

The contrast between ungrounded generation and KG-grounded reasoning is stark. As shown in Table 1, the Prototype architecture achieved a near-perfect success rate across all scenarios, while the Baseline models consistently struggled with logical consistency and educational over-recommendation.

Table 1: Logical Compliance Comparison Between Baseline and KG-Grounded Models

Metric	Baseline LLMs	KG-Grounded LLMs
Logical Action Validity (LAV)	Low (0–20%)	High (100%)
Minimal Intervention Compliance (MIC)	Not Observed	Consistently Observed
Over-Prescriptive Behavior	Systematic	Eliminated
Explicit Traceability (ET)	Absent	Present (Triples Cited)

Scenario-Based Analysis

Across all tested scenarios, unconstrained baseline models consistently produced over-prescriptive recommendations, even when the learner’s current proficiency exactly met or exceeded job requirements. This behavior indicates a systematic bias toward action generation, rather than a failure of factual memory. In contrast, KG-grounded models exhibited deterministic logical compliance, reliably terminating the instructional process when no pedagogical gap existed.

Qualitative Model Characterization and Stability

The transition from Baseline to KG-grounded mode revealed distinct behavioral shifts across different model architectures. We categorized these shifts to understand the "stability" of the AI Corrective Role (see Table 2).

Table 2: Qualitative Behavioral Characterization Under KG-Grounding

Model	Symbolic Explicitness	Narrative Style	Constraint Compliance
DeepSeek	High	Low	Strict
Sonnet 4.6	Moderate	High	Strict
Gemini 3.1 Pro	Low	High	Moderate
GPT-5.3	Moderate	Moderate	Strict
Mistral Large 2	Low	Low	Moderate

Most Symbolic Reasoner: DeepSeek exhibited the highest fidelity to formal logic, utilizing code-block notations for triples and explicitly defining the CEFR hierarchy ($A1 < A2 < B1$) required for deduction.

Most Adaptable Interface: Sonnet 4.6 demonstrated superior formatting, mapping graph triples into readable tables while maintaining strict logical boundaries.

Most Narrative Logic: Gemini 3.1 Pro maintained a Human-Centric tone, integrating the logic into conversational paragraphs rather than rigid symbolic proofs.

Traceability and the "Explainable" Effect

Explainability is considered an important property of trustworthy AI systems [17]. The Baseline outputs were predominantly "Black-Box" in nature, providing advice based on statistical probability. In contrast, the Prototype models utilized Triple-referenced reasoning. By explicitly citing the KG triples (e.g., Job_CleaningStaff, requiresLevel, A1_English), the models transformed the reasoning process into explainable audit trail. This transparency allows both the educator and the student to verify that the recommendation is rooted in the system's "Ground Truth" rather than an algorithmic hallucination.

Discussion

5.1. Validity of Knowledge Graph Constraint Mechanisms

The results clearly demonstrate that integrating knowledge graphs significantly constrains the outputs of large language models (LLMs) toward logically valid pedagogical decisions. The stark difference between Baseline performance, characterized by high rates of over-prescription and a systematic hallucination of non-existent interventions, and Prototype performance, which shows near-perfect logical compliance, provides empirical evidence for the core hypothesis: explicit knowledge structures can limit the generative action spaces of these models.

This finding aligns with and extends previous research that highlights the effectiveness of knowledge graphs in domain-specific AI systems. For instance, Ahmad et al. (2026) document similar constraint effects in IoT security systems, where graph-based representations of system states improve both detection accuracy and interpretability. Similarly, Ray (2026) identifies comparable mechanisms in lightweight LLM applications, noting that access to a knowledge base reduces computational demands and enhances reliability. However, the educational domain has unique characteristics that enhance the utility of knowledge graphs: educational constraints are formal, verifiable, and relatively stable, unlike the dynamic nature of security threats or manufacturing processes. This makes education an ideal domain for demonstrating the effectiveness of symbolic-neural integration.

The mechanism by which knowledge graphs constrain generative systems requires further exploration. Our findings suggest that this constraint does not operate through explicit negation (i.e., preventing the generation of certain tokens) but rather through the restriction of action space. By making only specific semantic compositions logically valid within the graph, the LLM is guided toward those compositions, even when prompted solely through engineering techniques. This approach differs from filtering, where incorrect outputs are generated and subsequently discarded. Instead, it appears that the LLM learns to restrict its own generation within a single inference pass when faced with an explicit semantic structure.

5.2. Behavioral Differentiation Across LLM Architectures

All tested models showed improvements when utilizing knowledge graph grounding, but differences in their reasoning styles became apparent. DeepSeek demonstrated the highest fidelity to formal logic by explicitly encoding prerequisite hierarchies ($A1 < A2 < B1$) and using code-block notation for symbolic

reasoning. Sonnet 4.6 excelled in formatting and readability while maintaining logical boundaries. Meanwhile, Gemini 3.1 Pro preserved a human-centered narrative tone while integrating logical constraints.

These variations indicate that knowledge graph grounding is not a one-size-fits-all approach; rather, different architectures assimilate symbolic constraints through distinct cognitive pathways.

This finding is significant for the design of educational systems. Educators might choose large language model (LLM) backends based not just on predictive accuracy, but also on reasoning transparency. DeepSeek would be preferred when explicit formal reasoning is important, Sonnet 4.6 when output readability is a priority, and Gemini 3.1 Pro when maintaining a natural conversational tone is essential for teaching.

This represents a shift in how educational AI systems are evaluated, moving from generic benchmarks to a more nuanced, architecture-specific assessment of how symbolic constraints are interpreted and expressed.

5.3. Explainability as an Inherent Property of Knowledge Graph Grounding

A significant secondary finding relates to explainability. The baseline models functioned as "black boxes," offering advice based solely on statistical probabilities without revealing the reasoning behind their decisions. In contrast, the prototype models explicitly referenced Knowledge Graph triples in their outputs, creating auditable trails of reasoning. This distinction was not explicitly prompted; rather, it emerged from the design of the constraint mechanism. When a language model must base its outputs on specific semantic triples, it naturally provides clear references to those triples.

This approach differs from post-hoc explainability techniques, such as SHAP, LIME, and attention visualization, which attempt to rationalize outputs after they have been generated. Gangwal and Lavecchia (2026) highlight the limitations of such methods: they offer correlations rather than causal explanations and often fail to maintain mechanistic plausibility. In contrast, grounding outputs in a Knowledge Graph produces inherently explainable results because the reasoning pathway is directly encoded within the graph structure itself. This signifies a shift from viewing explainability as a feature to understanding it as an architectural principle, with implications that extend far beyond the field of education.

5.4. Implications for Educational AI Trustworthiness

Educational AI systems are increasingly involved in important decisions, such as course recommendations, learning path designs, and identifying students who are struggling. Currently, these

systems tend to prioritize helpfulness and engagement over accuracy and logical validity. Our findings indicate that trust in educational AI could be improved not by making the models more complex, but by explicitly incorporating pedagogical constraints that exist outside the model.

Adi et al. (2026a) highlight that the success of AI-enhanced education relies on maintaining transparency, allowing learners to understand the basis for AI recommendations and ensuring they retain control over educational decision-making. Knowledge graph grounding offers a practical approach to achieving this transparency. By making educational logic clear and traceable, symbolic scaffolding enables educators to verify and, if necessary, override AI recommendations. This preserves the important distinction between AI functioning as a decision-maker and as a decision-support tool.

Moreover, knowledge graph grounding addresses equity concerns raised by Zadid et al. (2026) related to fairness in natural language processing (NLP). By explicitly encoding pedagogical principles, cultural education values, and equity constraints within graph structures, educators can guarantee that AI recommendations are fair across diverse learner populations. The graph itself becomes a platform for ethical deliberation, allowing educators to design constraints that reflect their pedagogical values and cultural commitments. This approach transforms knowledge graphs from mere technical infrastructure into tools for embedding educational philosophy directly into AI systems.

5.5. Toward Hybrid Symbolic-Neural Architectures in Educational AI

The success of knowledge graph grounding in shaping large language model (LLM) outputs raises important questions about the integration of symbolic and neural approaches in hybrid systems. Ray (2026) identifies this integration as a key future direction for LLM-Internet of Things (IoT) ecosystems, emphasizing that combining the flexibility of neural networks with the constraints of symbolic systems results in performance that exceeds what either approach can achieve on its own. Similarly, Saeidi et al. (2026) document how graph neural networks merge graph-structured data with neural learning, leading to improved interpretability and generalization.

For educational AI, this implies future research should focus on dynamic knowledge graphs that can evolve in real-time based on learner feedback and evidence of intervention effectiveness. Instead of relying on static constraint structures, adaptive graphs could represent learner proficiency along with uncertainty margins, depict prerequisite dependencies as probabilistic relationships, and include educational interventions with assessed efficacy ranges. This approach would integrate the explainability and constraint benefits of knowledge graphs with the adaptability and responsiveness of neural systems.

5.6. Critical Limitations and Scope Boundaries

It is essential to emphasize that this study is deliberately limited in scope. The Knowledge Graph employed is minimal and deterministic, which is designed to isolate logical constraint effects rather than model pedagogical richness. Real educational systems involve complex, multidimensional learner characteristics, nuanced prerequisite structures, and probabilistic learning pathways. Our simplified scenarios exclude ambiguity, conflicting prerequisites, and probabilistic progression—unrealistic simplifications compared to authentic educational contexts.

Moreover, this is a proof-of-concept demonstrating mechanism, not a system demonstrating learner outcome improvement. We do not measure whether knowledge-graph-grounded LLMs produce better learning outcomes, higher engagement, or more equitable educational experiences. Such questions require different research designs (randomized controlled trials with learner outcome measurements) and fall outside the scope of design-science mechanism validation.

Additionally, the evaluation protocol, while rigorous, operates at the model behavior level. Real deployed systems require careful attention to user interface design, educator training, and integration with existing educational information systems. These implementation challenges lie beyond this study's current scope but represent essential future work.

These limitations are not weaknesses but deliberate design choices consistent with design-science research principles. Hevner et al. (2004) emphasize that rigor in design science derives from carefully controlled artifact evaluation in bounded contexts, not from attempts to demonstrate external validity or generalization across heterogeneous populations.

5.7. Broader Implications for AI Trustworthiness Across Domains

While this study focuses on education, its implications extend across any high-stakes domain where AI recommendations must be both fluent and logically sound. Medical decision support systems (Warkiani et al., 2026; Sabit et al., 2026), pharmaceutical administration (Chen et al., 2026), industrial process optimization (Li et al., 2026), and supply chain management (Shishehgarkhaneh et al., 2026) all share the educational domain's core challenge: users must trust AI recommendations even when those recommendations are generated by opaque neural systems. Knowledge graph grounding offers a generalizable solution applicable across these domains.

Furthermore, our findings suggest that the current emphasis on larger, more capable LLMs may reflect a misplaced priority. Rather than endlessly scaling model parameters, domain-specific reliability might be achieved through integration with explicit domain knowledge. This represents a significant reorientation of AI development strategy, away from foundation models as universal solvers toward more modest, constrained models deployed within domains where symbolic structure is available.

Conclusion

This study demonstrates, through rigorous design-science evaluation, that symbolic scaffolding via pedagogical Knowledge Graphs can enforce minimal, logically valid instructional decisions in Large Language Models. By constraining the generative action space of neural systems through explicit semantic structures, the proposed approach eliminates over-prescriptive behavior while preserving explainability, determinism, and educational integrity.

The core contribution is precisely bounded: We provide empirical evidence that knowledge graph grounding prevents hallucinated pedagogical recommendations and enforces compliance with defined instructional rules across multiple state-of-the-art LLM architectures. We do not claim that this approach scales seamlessly to complex real-world educational ecosystems marked by ambiguity, conflicting constraints, and probabilistic outcomes. We do not demonstrate learner outcome improvements or claim pedagogical superiority over existing educational systems. Rather, we establish foundational proof-of-principle that symbolic knowledge representation can function as a corrective mechanism for unconstrained neural generation: a finding with significant implications for high-stakes AI deployment across multiple domains.

From a methodological perspective, this work contributes to design-science research methodology in information systems. By carefully isolating a specific research hypothesis: Can knowledge graphs constrain LLM outputs toward logical validity? Then designing controlled comparative evaluation, we demonstrate how design science can advance understanding of AI system behavior even within bounded, proof-of-concept contexts. This approach aligns with Hevner et al. (2004) principles, emphasizing that rigorous design science achieves depth through careful scoping rather than breadth through population-level inference.

The integration of Knowledge Graphs into educational AI systems offers promising directions for future research that move beyond static constraint structures toward adaptive knowledge representation. Real-world educational systems operate within dynamic contexts where learner proficiency evolves, prerequisite structures require updating as curricula change, and educational outcomes become clearer through empirical validation. Future work should explore how knowledge graphs can be made adaptive, updating in real-time as educators observe learning outcomes and as learners progress through educational pathways. This represents a significant technical challenge requiring integration of machine learning techniques for graph learning, symbolic reasoning engines for constraint satisfaction, and educational measurement methods for validation.

Equally important, future research must address how symbolic constraints can be made culturally responsive and equitable. Adi et al. (2024) document success in integrating indigenous pedagogical traditions (Javanese Wayang philosophy, Papuan cross-cultural learning) into digital systems. By incorporating cultural and philosophical ontologies directly into knowledge graph architecture, researchers can ensure that AI's corrective feedback remains sensitive to the diverse lived experiences and learning traditions of global learner populations. This represents not merely technical advancement

but a fundamental commitment to equitable AI, where diverse ways of knowing are encoded into the systems that guide learning.

Broader implications extend to trustworthiness in AI systems across domains. As documented across medical decision support (Warkiani et al., 2026), pharmaceutical AI (Chen et al., 2026), and industrial applications (Li et al., 2026), domain-specific reliability increasingly depends on explicit knowledge integration. Yet these integrations remain ad hoc and domain-specific. Future research should develop generalizable frameworks for knowledge graph construction and LLM integration applicable across medicine, education, manufacturing, and governance. Such frameworks must address not only technical challenges but also governance questions: Who decides what constraints are encoded in graphs? How are trade-offs between different values adjudicated when constraints conflict? How can knowledge graphs remain transparent to end-users while incorporating complex domain expertise?

From the perspective of fairness and equity, our findings suggest that trust in AI systems may be supported through explicit encoding of ethical principles as graph constraints. Zadid et al. (2026) identify that fairness in NLP requires socio-technical approaches integrating community participation and governance. Knowledge graphs provide concrete mechanisms for such participation: community stakeholders can collaboratively design graph structures encoding their values, constraints can be adjusted as communities identify unintended consequences, and the graph itself becomes a site of democratic deliberation about AI behavior.

Looking toward emerging research frontiers, Ray (2026) identifies integration of LLMs with edge computing, federated learning, and multimodal sensing as key challenges. Our findings suggest that knowledge graph grounding becomes even more valuable in these contexts: lightweight, resource-constrained LLM deployments benefit significantly from explicit constraint structures that reduce computational demands while improving reliability. Federated learning scenarios, where knowledge graphs might be partitioned across institutional boundaries to preserve privacy, represent another promising direction.

Additionally, the combination of symbolic and neural approaches within a scalable computing infrastructure represents a critical area for further research. Current systems typically implement symbolic constraints through prompt engineering that embeds graph triples into text prompts. More sophisticated approaches might integrate symbolic constraints at the model architecture level, through specialized layers that enforce graph-based constraints during neural computation. This would represent true symbolic-neural integration rather than surface-level constraint injection.

Ultimately, this research contributes to a broader reorientation in how artificial intelligence systems are designed and evaluated in high-stakes domains. Rather than pursuing ever-larger, more capable foundation models optimized for general-purpose reasoning, domain-specific trustworthiness may be achieved through explicit grounding in domain knowledge. Rather than attempting to retrofit interpretability onto opaque neural systems through post-hoc explanation, we might design systems where interpretability is architectural, where constraint structures themselves provide explainability.

Rather than treating fairness and equity as downstream concerns addressed through bias detection and mitigation, we might encode ethical principles and cultural values directly into system design.

The preservation of learner trust through verifiable and equitable educational systems must remain the research priority as AI tools proliferate in educational settings. This study provides one pathway toward that goal: explicit symbolic scaffolding that constrains neural generation toward logically valid, educationally grounded, culturally responsive pedagogical recommendations. Future work building on this foundation: expanding scope, testing real learners, integrating adaptive mechanisms, and addressing ethical governance, will be essential to realizing the promise of trustworthy, intelligent educational systems.

Declarations

Consent to Participate:

Not applicable.

Consent to Publish:

Not applicable.

Ethics Approval:

Not applicable.

Clinical Trial Registration:

Not applicable.

Data availability statement

The data supporting the findings of this study are available within the paper and its Supplementary Information.

Funding

The authors gratefully acknowledge the support of Indonesian Education Scholarship (BPI), Center for Higher Education Funding and Assessment/Pusat Pembiayaan dan Asesmen Pendidikan Tinggi (PPAPT Kemendikisaintek) and the Indonesia Endowment Fund for Education (LPDP), Republic of Indonesia.

Acknowledgement

This research was supported by Indonesian Education Scholarship (BPI), Center for Higher Education Funding and Assessment/Pusat Pembiayaan dan Asesmen Pendidikan Tinggi (PPAPT Kemendikisaintek), Graduate Academy – TU Dresden, The Saxon State and University Library Dresden

(SLUB) Dresden and financed by the Indonesia Endowment Fund for Education (LPDP), Republic of Indonesia, under grant identification number 202209091599. Open access funding was enabled and organized by Projekt DEAL through TU Dresden.

References

1. Adi, P. N., Köhler, T., Azizi-Ghanbari, S., Triyono, M. B., Priyanto, P., & Handayani, S. A. (2024). Digital Interaction Experience Assessment of Papuan Tribe Cross-Cultural Learning Using Javanese Tribe Wayang Philosophy. *Gemeinschaften in Neuen Medien. Vertrauen in Der Krise: Kommunikationsmodelle in Digitalen Gemeinschaften*. 27. Workshop GeNeMe'24 Gemeinschaften in Neuen Medien. <https://doi.org/10.25368/2025.121>
2. Adi, P. N., Köhler, T., Triyono, M. B., Priyanto, Handayani, S. A., & Maruanaya, R. F. (2026). AI-enhanced gamification in education: An integrative review of trends, impacts, and corrective role potential. *Journal of Computers in Education*. <https://doi.org/10.1007/s40692-026-00387-0>
3. Adouni, O., Bessadok, A., Hamroun, M., Ejali, R., & Sakli, H. (2026). Multimodal data integration for early autism detection and LLM-driven personalized intervention: A review. *Computerized Medical Imaging and Graphics*, 130, 102743. <https://doi.org/10.1016/j.compmedimag.2026.102743>
4. Ahmad, H. B., Gao, H., & Latif, N. (2026). Graph intelligence for IoT and CPS security: From connectivity patterns to cyber-physical reasoning. *Computer Science Review*, 61, 100978. <https://doi.org/10.1016/j.cosrev.2026.100978>
5. Al-Amin, M., Hossain, M. E., Ali, M. S., Salam, A., Ali, A., Ullah, A., Alam, N., Akhter, M. N., & Mahmood, S. A. (2026). History of chatbots and their emerging applications in textiles, polymers, and sustainability. *Sustainable Materials and Technologies*, 48, e01967. <https://doi.org/10.1016/j.susmat.2026.e01967>
6. Alansari, A., & Luqman, H. (2026). Large language models hallucination: A comprehensive survey. *Computer Science Review*, 61, 100970. <https://doi.org/10.1016/j.cosrev.2026.100970>
7. Arslan, M., & Munawar, S. (2026). Large language models for sustainable water management: A review. *Sustainable Cities and Society: Advances*, 2(2), 100044. <https://doi.org/10.1016/j.scsadv.2026.100044>
8. Asimopoulos, D. C., Radoglou-Grammatikis, P., Papadopoulos, G. T., & Sarigiannidis, P. (2026). Beyond vulnerabilities: A comprehensive survey of adversarial attacks across domains. *Computer Science Review*, 61, 100963. <https://doi.org/10.1016/j.cosrev.2026.100963>
9. Chen, X., Jin, P., Mai, H., Liang, S., Ping, Y., Wang, C., & Zhao, Y. (2026). Artificial intelligence in pharmaceutical administration and clinical pharmacy: A comprehensive review of advancements, challenges, and future directions. *Intelligent Pharmacy*. <https://doi.org/10.1016/j.ipha.2026.04.003>
10. Gangwal, A., & Lavecchia, A. (2026). Explainable AI methods for drug discovery: A survey of interpretability, metrics and mechanistic insight. *Computer Science Review*, 61, 100943. <https://doi.org/10.1016/j.cosrev.2026.100943>

11. Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design Science in Information Systems Research1. *MIS Quarterly*, 28(1), 75–106. <https://doi.org/10.2307/25148625>
12. Huang, Y., Huo, P., Fan, W., Wang, J., Xu, X., Lu, X., & He, Y. (2026). Empowering Traditional Chinese Medicine with Artificial Intelligence: Innovations in Quality Control, Therapeutic Applications, and Industrialization. *TrAC Trends in Analytical Chemistry*, 118856. <https://doi.org/10.1016/j.trac.2026.118856>
13. Li, Y., Zhao, H., Jiang, H., Pan, Y., Liu, Z., Wu, Z., Shu, P., Tian, J., Yang, T., Xu, S., Lyu, Y., Blenk, P., Pence, J., Rupram, J., Banu, E., Song, K., Zhu, D., Wang, X., & Liu, T. (2026). Large Language Models for manufacturing. *Journal of Manufacturing Systems*, 86, 516–545. <https://doi.org/10.1016/j.jmsy.2026.02.014>
14. Park, I., Kim, Y., Kim, K., & Chin, S. (2026). BIM–LLM integration in AECO workflows: Applications, validation, and future directions. *Automation in Construction*, 187, 106905. <https://doi.org/10.1016/j.autcon.2026.106905>
15. Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems*, 24(3), 45–77. <https://doi.org/10.2753/MIS0742-1222240302>
16. Ray, P. P. (2026). A Review on LLMs for IoT Ecosystem: State-of-the-art, Lightweight Models, Use Cases, Key Challenges, Future Directions. *Internet of Things and Cyber-Physical Systems*. <https://doi.org/10.1016/j.iotcps.2026.04.002>
17. Sabit, H., Yadav, A. K., Salimy, S., Sakr, A., Abdel-Ghany, S., Wadan, A. S., Alqosaibi, A. I., Rashwan, R., AlGosaibi, Y. S., Alnamshan, M. M., Almulhim, J., Alaqeel, N. K., & Arneeth, B. (2026). Integrating multi-omics and artificial intelligence for personalized breast cancer management: A guide to clinicians. *Cancer Letters*, 649, 218468. <https://doi.org/10.1016/j.canlet.2026.218468>
18. Saeidi, N., Hajipour, V., Hekmat, S., & Shoushtarian, B. (2026). Unfolding graph convolutional networks: Architectures, use cases, and trends. *Applied Soft Computing*, 196, 115122. <https://doi.org/10.1016/j.asoc.2026.115122>
19. Shahnazinia, S., Tavasoli, M., Sarrafzadeh, A., & Karimoddini, A. (2026). Healthcare digital twins: A methodological literature review on integrating iot and AI for personalized medicine and predictive care. *Telematics and Informatics Reports*, 22, 100325. <https://doi.org/10.1016/j.teler.2026.100325>
20. Shishehgarkhaneh, M. B., Moehler, R., Fang, Y., Hijazi, A. A., Aboutorab, H., & Dhamankar, A. J. (2026). Application of recommender systems in supply chain management: A state-of-the-art review. *Intelligent Systems with Applications*, 30, 200661. <https://doi.org/10.1016/j.iswa.2026.200661>
21. Warkiani, M. E., Safari, B., Moattar, M. H., & Hosseinzadeh, M. (2026). A systematic review of multimodal large language models for medical decision support: Architectures, fusion methods, and real-world applications. *Applied Soft Computing*, 196, 115051. <https://doi.org/10.1016/j.asoc.2026.115051>
22. Xu, H., Chen, Q., He, W., Chen, X., & Gao, H. (2026). Task-specific models vs. Large vision-language models in medical visual question answering: A survey. *Expert Systems with Applications*, 318,

132008. <https://doi.org/10.1016/j.eswa.2026.132008>

23. Yu, G., Wang, Y., Wang, Z., Chen, L., Zheng, Y., & Liu, Y. (2026). LLMs in industrial domains: A systematic review of adaptation techniques and applications from the product lifecycle perspective. *Advanced Engineering Informatics*, 74, 104655. <https://doi.org/10.1016/j.aei.2026.104655>
24. Zadid, T. H., Jahin, S. M., Dhruba, A. R., Tanim, S. A., Hamja, A. M., Hasan, M. R., Miah, M. S. U., Islam, M. I., & Mansur, A. A. (2026). A review of fairness challenges in natural language processing. *Array*, 30, 100773. <https://doi.org/10.1016/j.array.2026.100773>

Supplementary Files

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

- [LLMExperimentData.csv](#)