Computer Systems
Science and Engineering

Tech Science Press

https://doi.org/10.32604/csse.2026.089267

ARTICLE

A Proposal for the Automatic Curation of Subject Matter Questions in Online Course Forums

Antonio Leandro Martins Candido[1,2,*,#] **and José Everardo Bessa Maia**[1,#]

[1]Programa de Pós-graduação em Ciência da Computação, Universidade Estadual do Ceará—UECE, Campus Itaperi, Fortaleza, Ceará, Brazil

[2]Departamento de Ensino, Instituto Federal de Educação, Ciência e Tecnologia do Ceará—IFCE, Campus Caucaia, Caucaia, Ceará, Brazil

*Corresponding Author: Antonio Leandro Martins Candido. Email: leandro.candido@aluno.uece.br

[#]These authors contributed equally to this work

Received: 20 July 2026; Accepted: 23 September 2026; Published: 27 October 2026

ABSTRACT: Online course forums possess significant pedagogical potential but remain underutilized due to the difficulty and expense of their maintenance. Without meaningful, thought-provoking, and timely responses, student interest wanes, negatively impacting the effectiveness of the learning environment for both teaching staff and students. To address this challenge, this study introduces a framework for the automatic curation of subject matter questions in online course forums. The proposed framework consists of three macro functions: a curation engine, post classification, and an instructional Retrieval-Augmented Generation (RAG) module. The automatic curation engine is designed as a reflex agent with an internal state. The post classification process is multidimensional and employs domain adaptation to mitigate distributional deviations. Additionally, the RAG module leverages carefully curated pedagogical resources specific to the course to adjust the focus and grounding of Large Language Models (LLMs). These components were evaluated both independently and comprehensively as an integrated system. An end-to-end experiment conducted with a university technology course demonstrated broadly satisfactory performance for practical application. Quantitatively, the domain-adapted classifier achieved an 87.50% overall accuracy in question detection. Furthermore, the instructional RAG module attained a mean Context Recall of 0.988 and a Faithfulness score ranging from 0.62 to 0.95 across five evaluated LLMs. These findings indicate that the proposed automated curation framework can effectively identify and address student inquiries, significantly reducing the manual burden on educators. Ultimately, this approach enhances the viability of online course forums as active, responsive, and scalable pedagogical tools.

KEYWORDS: Discussion forum curation; curation engine; post-classification; domain adaptation; instructional RAG

1 Introduction

In online learning settings, discussion boards are essential for enabling communication between learners and instructors. Yet, if they aren't thoughtfully managed, these spaces can become cluttered and less successful at supporting learning. Strong curation includes ongoing moderation, encouraging focused conversations, and incorporating useful instructional materials. For instance, the MOOC Uniform Identifier for Resources (MUIR) framework suggests cross-linking MOOC resources to enhance forum discussions and improve participants' overall learning experience [1].

Online pedagogical forum curation is the practice of choosing, structuring, and providing context for content and exchanges in virtual learning settings, in order to foster substantive dialogue, encourage

Copyright © 2026 The Authors. Published by Tech Science Press. This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

active participation, and support collaborative knowledge building. This work is crucial for turning digital spaces into effective learning communities, particularly within distance education and Massive Open Online Courses (MOOCs).

Moreover, curation in online teaching forums entails fostering conditions that sustain a community of inquiry, in which participants take part in thoughtful, reflective dialogue. Recommendations for nurturing a community of inquiry in online discussion forums emphasize the value of approaches that encourage students' critical thinking and collaborative engagement [2–4].

Digital curation is widely viewed as a key practice within media and digital literacy education. Treating curation as a central competency in these fields highlights the importance of teaching students how to curate information so they can effectively navigate and critically assess the enormous volume of online content [5].

In sum, curation within online teaching forums is a complex activity that entails deliberately guiding content and participation to foster meaningful, collaborative learning. It calls for a blend of instructional, technical, and moderation approaches to build online learning spaces that are both effective and engaging.

Online learning environments increasingly rely on discussion forums as the primary channel through which learners raise doubts about course content, yet managing these learner questions manually poses substantial challenges: the volume of posts grows faster than instructor availability, response latency discourages engagement, and the cognitive effort required to triage urgent, confused, or off-topic messages is rarely accounted for in course workload. These challenges in managing learner questions motivate the need for automated curation mechanisms that can identify, prioritize, and respond to learner questions without compromising pedagogical quality.

In this work, we adopt the notion of a *subject-matter question* as a forum post in which a student expresses doubt, confusion, or a request for clarification about course content, as opposed to posts that are purely social, administrative, or unrelated to the subject matter, following the question and confusion dimensions defined by [6]. Distinguishing these posts from the broader stream of forum activity is the first step toward any automated curation strategy.

Courses are pedagogical events with carefully planned objectives and resources, losing effectiveness when discussions lose focus or deviate from these goals. The introduction of resources beyond the planned scope can also act as a dispersive factor. Previous efforts to automate course forum curation have relied on the capabilities of LLMs; however, even some RAG-based proposals employ open-web resource searches, generating an overwhelming diversity of information that makes it difficult for students to keep up.

In this research, we designed a framework for the automatic curation of subject matter questions in online course forums, which consists of three macro functions: a curation engine, post classification, and instructional RAG. These components were evaluated both separately and as an integrated system. The end-to-end experiment, conducted with a university technology course class, demonstrated broadly satisfactory performance for practical application.

The contributions of this paper are:

- We designed a framework for the automatic curation of subject matter questions in online course forums, which consists of three macro functions: a curation engine, post classification, and instructional RAG.
- We adopted a novel Markovian approximation to represent both the mental state of the active forum participant and the overall forum state, factored by individual participants.
- We implemented a data schema for unsupervised domain adaptation tailored to our data-deficient scenario. This approach improves the classification performance of posts, which typically consist of short texts conveying minimal semantic information regarding the target class;

- We evaluated the framework components separately on public datasets, achieving state-of-the-art performance for the targeted tasks;
- We evaluated the end-to-end system on subject-matter questions in a real-world experiment, with results indicating satisfactory performance for practical deployment.

Paper organization: Section 2 reviews the main works and datasets related to this research. Section 3 presents the methods that comprise the proposed and evaluated curation engine. Section 4 presents and analyzes the main results of this paper. The paper concludes in Section 5.

2 Related Works

Since our proposed framework is composed of three macro functions—the curation engine, the post classification function, and the RAG module—this section reviews relevant prior work in each of these areas, focusing on online forum performance. An additional subsection details the datasets utilized in this study.

2.1 Online Forum Automation

The automation of student support in educational discussion forums has been the subject of increasing interest in the literature, evolving from traditional recommendation systems to complex architectures based on LLMs. In initial studies focused on this domain, ref. [7] investigated the fundamentals for building support systems in MOOCs, proposing a framework aimed at the automatic recommendation of answers. The central objective of this approach was to mitigate the information overload inherent in massive forums, assisting students in the rapid identification of existing relevant solutions in the discussions.

With the advancement and popularization of generative models, the research focus expanded to strategies that combine the ability to generate text with the retrieval of contextual information. In this sense, ref. [8] developed RetLLMe, a question-answering strategy for student forums. This work introduces a technique that combines context retrieval and prompt engineering (retrieval-prompt strategy), demonstrating the effectiveness of anchoring Large Language Model (LLM) generation in previously validated information to improve the accuracy of automated support.

As the use of LLMs consolidates as a support tool in these online learning environments, new demands arise related to the quality, ethics, and transparency of interactions. Ref. [9] critically explored these dimensions, analyzing the support generated by LLMs from the perspective of fairness and explainability. The research highlights that, in addition to providing correct answers, it is imperative that online support systems operate in an equitable and transparent manner, ensuring that the reasoning behind the generated guidance is understandable and unbiased for students.

Seeking to further improve the fidelity and structuring of knowledge in support systems, very recent research has adopted more sophisticated information retrieval architectures. Ref. [10] proposed the application of Graph RAG in the context of MOOCs. By leveraging the representation of educational knowledge in graph format, the approach aims to overcome the limitations of purely semantic retrieval, offering more robust and interconnected support to help learners understand complex concepts.

The role of LLMs such as Chat Generative Pre-trained Transformer (ChatGPT) in learning has also been examined from a broader educational-theory perspective. Ref. [11] showed that combining activity-based learning strategies with ChatGPT support improves student performance, reinforcing that the pedagogical value of LLM-based assistance depends not only on model capability but also on how it is integrated with established learning theories. Complementing this, ref. [12] synthesized educational theories with virtual reality technologies to enhance educational experiences, illustrating a parallel trend of grounding emerging

educational technologies in established pedagogical frameworks, a principle that also underlies our RAG design choice to constrain generation to curated course material.

2.2 Post Classification

Using the identical MOOC dataset as this study, ref. [13] investigated the feasibility of developing cross-domain classification models for forum posts. The author categorizes posts into functional taxonomies (e.g., questions, answers, opinions) to create domain-independent classifiers. The study highlights the challenges of transferring knowledge across disciplines like Education, Medicine, and Humanities, employing feature engineering and machine learning to enhance model robustness. Findings indicate that while linguistic and structural features support cross-domain classification, model performance remains sensitive to the specificities of distinct educational contexts.

Ref. [14] presented a comprehensive text classification process using state-of-the-art Natural Language Processing (NLP) models. The authors describe an end-to-end architecture that integrates modern natural language processing techniques, including preprocessing, feature extraction, and deep learning algorithms for text classification. The work provides a comprehensive overview of the steps required to implement robust classification systems, demonstrating the applicability of contemporary NLP models in automatic text categorization tasks.

Ref. [15] proposed a method for classifying question texts in the tourism domain using deep learning models. Their work applies deep learning techniques to automatically categorize questions related to the tourism sector, demonstrating the effectiveness of neural networks in classifying domain-specific texts.

Driven by the rapid evolution of LLMs, recent research has actively explored the integration of these advanced tools into educational text classification. For instance, ref. [16] proposed the Annotation Guidelines-Based Knowledge Augmentation (AGKA) approach, which leverages GPT-4 to distill insights directly from coding guidelines. This methodology significantly enhances the categorization of student engagement in educational corpora while circumventing the computational overhead typically associated with extensive model fine-tuning.

MOOC forums are dynamic environments where new courses and student profiles continually emerge, often lacking exact historical data for supervised classifier training. This reality motivates recent efforts to improve the generalization of classifier performance. For instance, ref. [17] sought transferability based on semantic similarity; ref. [18] explored neural attention mechanisms; ref. [19] investigated weighted Bidirectional Encoder Representations from Transformers (BERT) features and multi Convolutional Neural Network (multi-CNN) models for this purpose; and ref. [20] utilized linguistic modeling to identify content-related MOOC discussion threads across domains.

2.3 Retrieval-Augmented Generation (RAG)

The underlying architecture of much of contemporary automated support solutions is based on RAG, fundamentally introduced by [21]. The proposal to integrate pre-trained generative models with document retrieval mechanisms established a new paradigm for handling knowledge-intensive NLP tasks, mitigating hallucinations and ensuring that responses are based on reliable textual sources.

Building on this technological foundation, ref. [22] improved the application of RAG architectures specifically for course forums. The authors demonstrated that fine-tuning LLMs combined with the use of multiple chain-of-thought reasoning strategies substantially increases the system's ability to process complex reasoning and generate more detailed and assertive answers to student questions.

Alongside strictly semantic vector retrieval, information representation through Knowledge Graphs has emerged as a powerful alternative for modeling the complexity of the educational domain. Ref. [23] proposed a multi-view knowledge graph with the aim of offering explainable content recommendations based on forum discussions. The use of this relational topology allowed mapping explicit connections between course concepts and student questions, ensuring greater transparency in the system.

Taking this structuring to the state of the art of generative architectures, ref. [24] conducted a rigorous evaluative study comparing the performance of traditional RAG systems with graph-based approaches (GraphRAG). The study, focused on question-and-answer tasks with page-level retrieval using a mathematics textbook, outlined the trade-offs between the strategies, highlighting scenarios where relational and topological retrieval of dense academic material outperforms purely semantic methods.

Although the presented works demonstrate significant advances in the automation of student support, ranging from recommendation systems to sophisticated architectures based on LLMs and GraphRAG, most assume that the analyzed posts have already been identified as questions or requests for clarification. This work contributes to this field by proposing a two-step architecture that integrates classification and response generation: first, a Domain-Adversarial Neural Network (DANN) is used to classify whether a post is indeed a question, and then RAG is used to provide contextualized and accurate answers to the identified questions. This approach aligns with recent trends in combining information retrieval and text generation, while introducing a screening step that improves the system's robustness, avoiding unnecessary attempts to answer non-questions and allowing for a more efficient allocation of automated support resources.

2.4 Research Gaps and Limitations

Despite the breadth of prior work reviewed above, three gaps remain. First, most forum-automation systems (Section 2.1) assume that candidate posts have already been filtered as questions, leaving the upstream detection problem underexplored. Second, post-classification approaches (Section 2.2) are typically evaluated within a single course or domain, with limited attention to the unsupervised domain adaptation required when a system is deployed on a new course lacking labeled data. Third, RAG-based support systems (Section 2.3) are rarely evaluated end-to-end alongside the upstream classification step that decides which posts should trigger a response in the first place. This work addresses these gaps by combining domain-adapted question detection with instructional RAG within a single curation engine, evaluated both separately and end-to-end. More broadly, this work is also connected to the Educational Data Mining and Learning Analytics literature, which studies how behavioral and textual traces from online learning environments can inform pedagogical decision-making; the curation engine proposed here can be seen as an applied instance of this broader research program, focused specifically on forum question triage.

3 Method

3.1 The Curation Engine

The curation engine is built as a reflex agent with internal state [25]. Fig. 1 presents our proposed framework for online discussion forum curation. A curation engine, inserted as a forum member, interacts with the forum by reading posts and reacting to the forum's state. The engine responds with resource or action recommendations, or even direct responses via dialogue, based on the combination of two sources of information: local knowledge bases (resources), specific to the course and its context, and support from pre-trained LLMs, via a RAG procedure. From this figure, it can be seen that if the engine's confidence does not exceed a threshold th, it forwards the situational state to a human tutor for resolution. The design goal is to refer less than 2% of intervention situations to humans.

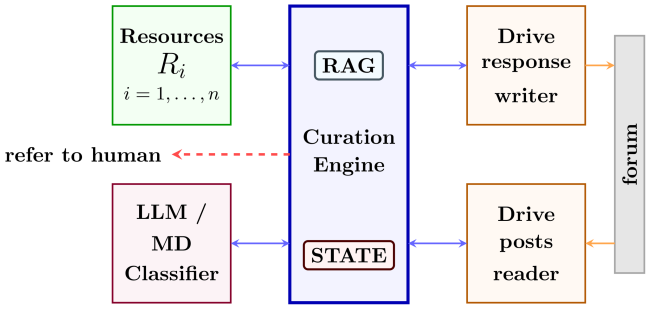


Figure 1: Functional structure of the online course forum curation agent software framework.

Before presenting a formal and abstract definition of the functions required to build the curation engine, it is instructive to provide a high-level overview of its operation.

The pedagogical objective of the forum is to promote the exchange of information and knowledge among participants, monitored by a tutor, and thereby foster the group's intellectual growth. Therefore, it is not desirable for the curation engine to immediately respond to every published post for which the sender expects a response. It is necessary to give participants the opportunity to discuss among themselves. With this objective in mind, the engine examines and acts on the forum periodically, once every τ hours (in this implementation, τ = 24 h, see Section 4.3), and in each time slot, it selects only f% of eligible learners to receive a response. To allow other participants time to express their opinions, the engine bases its intervention decisions solely on the states from the previous time slot.

Learner questions in course forums typically share a set of characteristics that shape the design requirements of a curation system: they are short, informal texts with limited semantic context; they may express varying degrees of urgency, confusion, or sentiment; and they arrive continuously and asynchronously, with no guarantee of a labeled history for the specific course or student population. Automatic curation of such questions must therefore operate under real-time constraints (bounded response latency), data scarcity (little to no course-specific labeled data), and a human-in-the-loop safety requirement (escalation to a tutor when the system's confidence is insufficient). These requirements motivate the formal model presented next.

Formalization of the problem: A model is an approximation of reality. We hypothesize that Markovian dynamics can serve as a good approximation of forum dynamics, provided that appropriate state and transition functions are defined. The letters i, j, k, and t index active learners, messages (posts) in a sequence, messages in a priority list for responding, and time slots, respectively. We define the "state" of an active learner in the forum as their "expectation state" regarding the set of posts they have published. Let s_i^a be the expectation state of learner i with respect to aspect a. In our framework, aspect $a \in \{urgent, confusion, sentiment, engagement\}$. Here we adopt the definitions of [6], where $urgent$ refers to messages reporting technical problems that require immediate attention, and $confusion$ refers to posts reporting doubts or questions about the subject matter.

Forum configuration at time t, $\mathcal{C}_t$, refers to the message configuration, that is, the set of post sequences associated with each forum participant ID. The configuration can be thought of as a list of lists: a list of IDs, each mapped to a chronological sequence of messages with associated timestamps. The **forum state** at time t, denoted as S_t, refers to a compact, internal representation developed in this project that stores all information critical for curating the forum. The concepts of configuration and state are designed to encapsulate the essential data needed for effective forum curation.

Let $\mathcal{C}_t = \{\mathcal{T}_1, \ldots, \mathcal{T}_j, \ldots, \mathcal{T}_N\}$ be a forum configuration for time slot t, where N is the number of active learners (or IDs) in the forum and each $\mathcal{T}_j$ is the string of messages (posts) posted in that forum by the

j-learner, and where each message is composed of text, a timestamp, and possibly associated meta-symbols, such as emojis, likes, or other indicators from a predefined vocabulary.

The state of each active ID in the forum is represented internally by the agent with a 4-tuple: urgent (u), confusion (c), valence sentiment (v), and engagement (e). The overall forum state utilizes a factored representation, where each component corresponds to the state of an active ID.

Let $\mathbf{s}_i = [u_i, c_i, v_i, e_i]$ be the state of the i-th active learner, and let $\mathbf{S} = [\mathbf{s}_1, \ldots, \mathbf{s}_i, \ldots, \mathbf{s}_N]$ be the forum state, where $\{u_i, c_i, v_i, e_i\} \in [0, 1]$. Note that it is optimal for u_i and c_i to be close to zero, and for v_i and e_i to be close to 1.

Let $\mathcal{R}$ also be the set of available pedagogical resources, a subset of which, R_i^j, must be selected as the context for the response to post p_j of learner i during the RAG procedure. With the notation already established, forum curation requires constructing the following sequence of processing functions.

- A post classification function $\omega_i^j = f_c(post_i^j)$ which assigns a class probability ω to post j of learner i. Alternatively, f_c can be a binary classifier with hard decision. This classifier is trained offline on a labeled course post database, such as the one in [6].
- A function that selects and prioritizes the list of posts to be addressed in the current time slot, denoted as $L = \{L_1, L_2, \ldots, L_k, \ldots\} = f_p(S, C)$.
- A function to select a subset of the resources (retrieval) from $\mathcal{R}$, R_k, to be used as context for the response to post L_k.
- A function to generate the response to the k-th post in list L, $Y_k = f_r(S, C, R_k)$.
- A state update function S using reply posts represented by $S = f_{sr}(S, \{\omega_i^j\})$;
- A state update function S using query posts (confusion) represented by $S = f_{sq}(S, \{\omega_i^j\})$.

In the implemented prototype, the binary output of the question-detection classifier f_c (Section 3.2) is used as an operational proxy for the confusion component c_i of the state update f_{sq}.

States and state transition functions: The volume of forum configuration data (reality) encoded within the state representation dictates the fidelity of the modeled environment. This fidelity, in turn, determines the complexity of both the transition function and the message classification function. For example, for a binary state representation (ID with a pending or not pending response), the post analysis function is a simple classifier with three classes: question, response, and neutral (with respect to being a question). Conversely, if the state incorporates richer information—such as multiple pending questions for a single learner ID—semantic post analysis becomes necessary to accurately associate responses with their corresponding questions, thereby increasing the complexity of the state transition function.

Algorithm 1 shows the macro steps of the framework's operation. In each action cycle with period τ, the engine reads new forum posts, categorizes and prioritizes a fraction f of posts to answer, retrieves context resources, and generates and publishes responses. Note that the forum status update occurs after the actions of responding to the forum, which means that τ is the waiting time for responses from active participants.

Algorithm 1: Pseudocode for the curation agent

Require: X-Current forum configuration
Require: S-Forum state (Permanent)
Require: th-Confidence threshold
Ensure: Y-Posts for intervention

(Continued)

Algorithm 1 (continued)

```
function CURATION_AGENT(S, X)
    X ← Acquires_current_configuration( forum)
    C ← Retrieves_post_contexts(X)
    ω ← Classifies_new_posts(X, C)
    S ← Updates_the_forum_status_for_reply_posts(S, ω)
    L ← Selects_and_sorts_to_respond(ω, S)
    for post = 1 to L do
        conf ← Evaluates_confidence(post, ω)
        if conf ≤ th then
            Escalates_to_human_tutor(post)
        else
            C_RAG ← Retrieves_RAG_context_resources(post, ω)
            P ← Builds_prompt(post, ω)
            Y(post) ← Generates_response(LLM, P, C_RAG)
        end if
    end for
    S ← Updates_the_forum_status_for_query_posts(S, ω)
    forum ← Publishes_to_forum(Y)
    return Y
end function
```

The functionalities of Algorithm 1 were implemented as Python procedures. The RAG retrieval function is simplified in our tests because of the small-scale nature of the evaluation. The scope of the curation engine, as implemented and evaluated in this work, is limited to detecting, prioritizing, and responding to individual learner questions; duplicate-question identification, topic modeling, automated curation-repository generation, an instructor recommendation module, and a continuous-learning feedback mechanism are not part of the current framework and are discussed as future extensions in the Conclusion. In the following subsections, we delve deeper into the fundamentals of post classification and RAG, the performance of which is essential for the curation engine's performance.

3.2 Post Classification

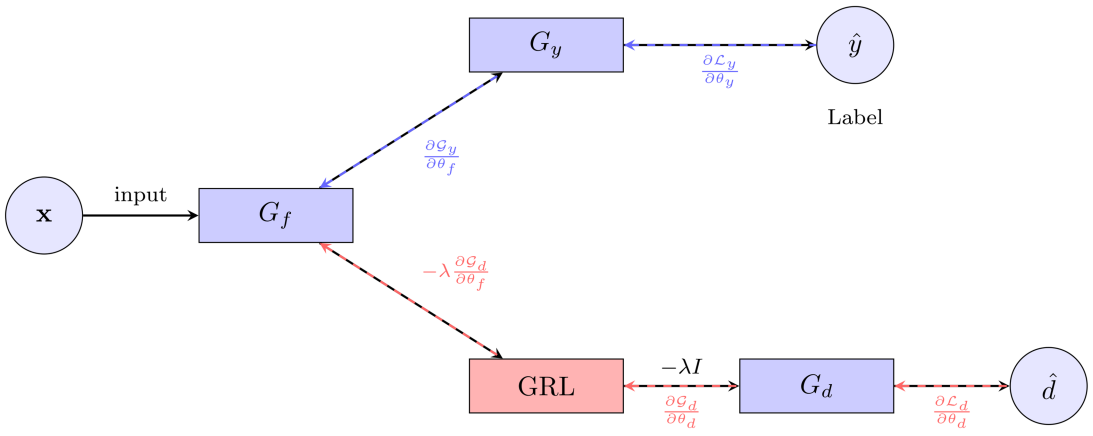
Classifying MOOC forum posts to determine whether tutor intervention is necessary is a well-recognized challenge. Domain adaptation is a central issue in this context because MOOC platforms are highly dynamic environments, continuously offering courses in new domains and modalities for which labeled training data is often unavailable. This challenge is particularly pronounced in natural language text classification, where linguistic ambiguity translates into feature ambiguity, playing a central role in the accurate discernment of text categories. Such scenarios are typically multidimensional, as is the case in our framework where the aspect $a \in \{urgent, confusion, sentiment, engagement\}$.

Prior to classification, forum posts are collected from the course's discussion platform, stripped of non-informative artifacts (e.g., automatically generated signatures, quoted replies, and exclusively numeric or emoji-only content), and normalized (lowercasing, whitespace and punctuation normalization) before being encoded as BERT embeddings for the classifier.

Multidimensional (or multiple outputs) classification [26] is characterized by considering the correlation between classes (labels) in addition to the correlation between features and classes present in standard one-dimensional classification. Building an independent classifier for each class is a viable option [27], but it fails to leverage inter-class correlations, which can often yield significant improvements. Incorporating this information can refine the decision boundary in regions that are ambiguous for independent classifiers.

3.2.1 Domain Adaptation

An alternative and influential approach to domain adaptation is the DANN, which applies adversarial training to learn domain-invariant features. DANN uses a domain classifier trained to distinguish between source and target domain samples, and a feature extractor trained to fool this domain classifier through a gradient reversal layer (GRL). The result is a minimax game in which the network learns features that are both discriminative for the main task and indistinguishable across domains [28]. The architecture of this network is shown in Fig. 2.



Forward pass (solid): data flow
Backward pass (dashed): gradient flow
Blue: label prediction gradients
Red: domain classification gradients (reversed by GRL)
Solid arrows show forward data flow from input $\mathbf{x}$ through feature extractor G_f to predictions. Dashed arrows show backward gradient flow: blue for label prediction (standard backpropagation) and red for domain classification (reversed by gradient reversal layer). The GRL multiplies domain gradients by $-\lambda$ during backpropagation, encouraging domain-invariant features.

Figure 2: DANN architecture with forward and backward passes.

In DANN, the total loss function is composed of two parts:

- A label predictor loss (L_y) computed on labeled source domain samples.
- A domain classification loss (L_d) computed on both source and target samples.

The feature extractor is optimized to minimize L_y while maximizing L_d (via gradient reversal), ensuring the learned features are informative for classification but uninformative for domain discrimination. The overall training objective is:

$$\underbrace{E(\theta_f,\theta_y,\theta_d)}_{L_{total}} = \underbrace{\frac{1}{n}\sum_{i=1}^{n} L_y^i(\theta_f,\theta_y)}_{L_y} - \lambda\underbrace{\left(\frac{1}{n}\sum_{i=1}^{n} L_d^i(\theta_f,\theta_d)+\frac{1}{n'}\sum_{i=1}^{n'} L_d^i(\theta_f,\theta_d)\right)}_{L_d}. \tag{1}$$

In this expression:

- θ_f are the parameters of the feature extractor,
- θ_y are the parameters of the label classifier,
- θ_d are the parameters of the domain classifier,
- n and n' are the number of samples from the source and target domains, respectively,
- λ is a weighting factor controlling the trade-off between classification and domain confusion.

3.3 Retrieval-Augmented Generation (RAG)

RAG is an AI technique that improves the accuracy of LLMs by retrieving relevant information from external databases before generating a response. This reduces hallucinations (fabrications) and allows the use of private, current, or specific data that the model did not see during training.

The proposed method follows a standard RAG pipeline adapted to educational support tasks. First, instructional materials are segmented into semantically coherent text units (contexts). These units are indexed as a knowledge base. For each forum question, the retriever selects the most relevant context candidates. Then, the generator receives a prompt containing the question and retrieved evidence, producing the final answer. Fig. 3 illustrates the complete workflow, from data ingestion through evaluation.

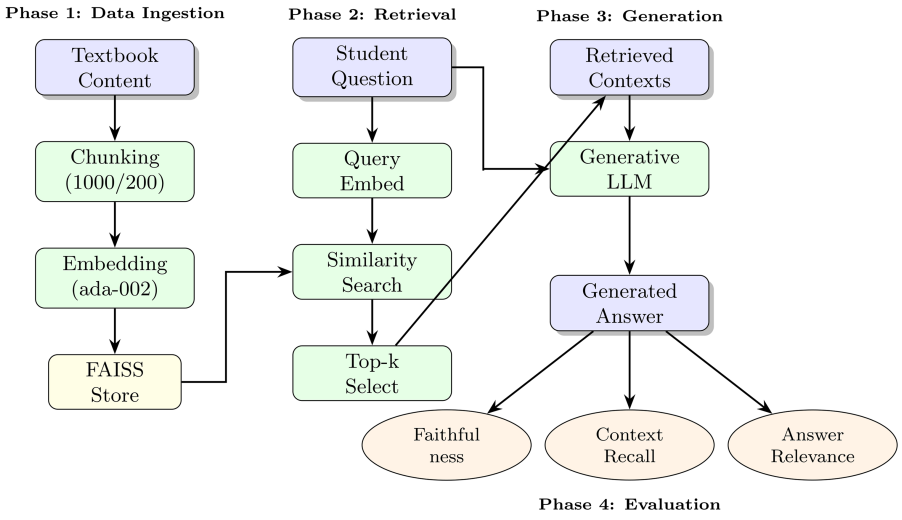


Figure 3: Complete RAG pipeline workflow. The system processes textbook content through a chunking preprocessing step (window 1000/200) and embedding, stores vectors in Facebook AI Similarity Search (FAISS), retrieves relevant contexts for student questions via similarity search, generates answers using LLMs, and evaluates outputs with Retrieval-Augmented Generation Assessment Suite (RAGAS) metrics.

3.3.1 Resources and Retrieval

Context Retrieval: This component implements the function $\mathcal{R} : Q \to \mathcal{P}(\mathcal{D})$, where $\mathcal{P}(\mathcal{D})$ is the power set of the documents. The process involves:

1. **Indexing:** Preprocessing the knowledge base $\mathcal{D}$ into embeddings using an encoder model (e.g., text-embedding-ada-002):

$$\mathbf{E}_{\mathcal{D}} = \{\mathbf{e}(d_1), \mathbf{e}(d_2), \ldots, \mathbf{e}(d_n)\} \tag{2}$$

2. **Semantic Search:** For each query q, $\mathbf{e}(q)$ is calculated and a cosine similarity search is performed:

$$\text{score}(q, d_i) = \cos(\mathbf{e}(q), \mathbf{e}(d_i)) \tag{3}$$

3. **Top-k Selection:** Returns the k documents with the highest score:

$$C = \text{Top-k}(\{\text{score}(q, d_i)\}_{i=1}^{n}) \tag{4}$$

3.3.2 Context-Supported Generation

Generative LLM: A pre-trained language model adapted to the specific domain is employed. The model receives the following input sequence:

$$\text{prompt} = [q] \oplus [\text{SEP}] \oplus [C] \oplus [\text{SEP}] \oplus [\text{instruction}] \tag{5}$$

where $\oplus$ denotes concatenation and [SEP] are separator tokens. The response is generated through autoregressive decoding with beam search or nucleus sampling [29].

4 Experiments and Discussion

4.1 Datasets

We utilized two datasets in this study. The first is a publicly available benchmark used to validate the post classification step. The second is a proprietary dataset, constructed during a real-world experiment, designed to validate the end-to-end process.

4.1.1 Dataset for Training Classifier

In this study, we employed the Stanford MOOC forum post dataset[1] [6], consisting of 29,604 de-identified messages collected from 11 different online courses. This corpus is made available to the scientific community upon request. The included courses encompass three major academic domains: education, medicine, and humanities/sciences. To establish a robust gold-standard reference, every entry was manually coded by three independent annotators [6].

The annotation process evaluated each post across six analytical dimensions. The categories of opinion, question, and answer were classified using binary labels. Conversely, the dimensions of sentiment, confusion, and urgency were measured on a 1–7 scale, providing the quantitative basis required to address the following research questions:

- Is this post a question? (Yes/No)
- Is this post an answer? (Yes/No)
- Is this post an opinion? (Yes/No)

[1] http://datastage.stanford.edu/StanfordMoocPosts

- What sentiment does this post express? (1–7)
- How much confusion does this post express? (1–7)
- How much urgent intervention in this post? (1–7)

Regarding the urgency scale, a score of 1 denotes a non-urgent post that requires no intervention, whereas a score of 7 represents a critical message demanding immediate instructor response. To facilitate binary classification, this 1–7 scale was binarized: posts scoring below 4 were assigned to class 0 (non-urgent), and those scoring 4 or above were assigned to class 1 (urgent). After filtering out entries consisting exclusively of numeric characters, the refined dataset comprised 29,604 posts. Of these, 23,186 instances were labeled as non-urgent and 6418 as urgent. The same binarization threshold (scores of 4 and above assigned to class 1, on the same 1–7 scale) was applied to sentiment and confusion. A comprehensive breakdown of the dataset distribution is provided in Table 1.

Table 1: Sample distribution across distinct domains within the StanfordMOOC dataset.

Areas/Sets	No/Yes	Opinion	Question	Answer	Confusion	Sentiment	Urgency
DS1	no	938	9168	9690	6715	1690	9418
	yes	8941	711	189	3164	8189	461
DS2	no	2782	3574	3179	481	1277	3314
	yes	2402	1610	2005	4703	3907	1870
DS3	no	2471	2232	2244	490	196	2053
	yes	559	798	786	2540	2834	977
Education	no	938	9168	9690	6715	1690	9418
	yes	8941	711	189	3164	8189	461
Humanities/Science	no	6016	7671	8002	1358	846	7247
	yes	3707	2052	1721	8365	8877	2476
Medicine	no	6181	6806	5857	1581	1851	6521
	yes	3821	3196	4145	8421	8151	3481

Note: The data is partitioned following two organizational schemes. The first scheme focuses on individual courses: DS1 (Education—EDUC115N: How to Learn Math), DS2 (Medicine—SciWrite, Fall 2013), and DS3 (Humanities/Science—Statistical Learning, Winter 2014). The second scheme classifies the posts by broader learner areas: Education (EDU), Medicine (MED), and Humanities & Sciences (H&S). The Education area contains a single course (DS1); consequently, the Education row is identical to the DS1 row, and the two partitioning schemes coincide for this area.

4.1.2 Source and Dataset for RAG Experiment

The experiments employ a curated knowledge base derived from a foundational computer architecture textbook (Hennessy, J. L. & Patterson, D. A. Computer Architecture: A Quantitative Approach). This material, adopted for a course at the State University of Ceara, encompasses core topics such as instruction set architecture, pipelining, cache memory, and memory hierarchy. The evaluation dataset comprises 300 instances of student posts—150 of which are questions—constructed from real student inquiries that arose during individual study and discussions in Brazilian Portuguese. The experiment was conducted via Google Classroom with a class of 50 students (aged 18 or older). The data collection was carried out retrospectively and comprised strictly anonymized textual questions derived from standard pedagogical activities, without involving sensitive personal data. In accordance with Brazilian regulations (National Health Council Resolution No. 510/2016), this type of non-interventional educational data analysis is exempt from formal

Institutional Review Board (CEP) evaluation. Informed consent was obtained from all participants, and the study complies with the Declaration of Helsinki. The 150 evaluation questions were originally formulated and collected in Brazilian Portuguese; the textbook knowledge base is in English, and, since this manuscript is written in English, the illustrative question examples shown in this section are presented in English translation for readability. This constitutes a cross-lingual retrieval setting (Portuguese queries against an English knowledge base), which may affect Context Recall relative to a monolingual configuration and is noted here as a limitation.

For each instance identified as a question, a structured dataset was developed to evaluate the response of LLMs with the following attributes: *page*, which denotes the identifier of the source page within the teaching material; *contexts*, which represent the excerpts retrieved from the textbook (covering one to three paragraphs); *question*, which contains the question in natural language formulated in a style characteristic of the student; *ground_truth*, which provides the reference answer extracted or paraphrased directly from the source; and *answer*, which corresponds to the answer generated by the model conditioned on the question and the contexts provided.

To guide the generative LLM, the following system prompt was utilized: "*You are a helpful assistant that answers questions in absolute mode based on the following context:* {*context*}. *Answer clearly and concisely in up to 500 tokens, without additional definitions, introductions, conclusions, or examples. If the information is not in context, say you don't know.*"

Although the dataset is modest in scale (150 instances), it reflects the practical constraints of domain-specific educational deployments, where expert manual annotation is highly resource-intensive. The formulated questions span various cognitive difficulty levels, ranging from definitional queries (e.g., "What is RAM memory?") and conceptual explanations (e.g., "Why does increasing associativity reduce conflict misses?") to quantitative reasoning problems (e.g., "How do we calculate effective memory access time?").

Recent frameworks such as Automated Evaluation Framework for Retrieval-Augmented Generation Systems (ARES) [30], benchmark suites [31], and RAGAS [32] provide structured evaluation protocols that decompose system quality into interpretable metrics. To validate our experiment, we employed RAGAS due to its LLM-as-a-judge methodology, which allows for the independent assessment of both retrieval and generation components; note that, since our evaluation set includes *ground_truth* reference answers, Context Recall is computed against these references and the protocol is not fully reference-free. By measuring specific dimensions, such as faithfulness to check context adherence, answer relevance to ensure the query was directly addressed, and context recall to validate the retriever's accuracy, RAGAS offers a comprehensive diagnostic tool that pinpoints exact areas of performance degradation within the RAG architecture.

4.2 Results for Question Detection in Course Forums

Prior to adopting DANN, we also evaluated Deep Correlation Alignment (Deep CORAL) as an alternative unsupervised domain-adaptation architecture; DANN achieved higher accuracy in our preliminary tests and was therefore selected for all experiments reported below. The DANN approach, as shown in Table 2, demonstrated relevant gains, but with sensitivity to semantic variations between domains. When comparing a subset Hum3000 (3000 posts of Humanities/Science—Statistical Learning, Winter 2014) as the source and subset Hum2000 (2000 posts of Humanities & Sciences) as the target, DANN achieved an accuracy of 81% and an F1-score of 90%. However, in scenarios involving more heterogeneous domains, such as Med5000 (5000 posts of Medicine), DANN's performance showed greater variation, revealing reduced stability in this more heterogeneous, genuinely cross-domain setting.

After verifying the good performance of DANN across different domains, we utilized the PyTorch library to conduct training on a real-world dataset comprising student posts. Training was conducted for 300 epochs using the Adam optimizer with an initial learning rate of 0.001, weight decay of 10^{-5}, and a learning rate scheduler with patience of 5 epochs. Cross-entropy loss was employed for both classification tasks, and we applied 5 iterations to compute statistical metrics. BERT was used to extract text embeddings, as shown in Table 3. The training labels adhered to the annotation strategy of the source corpus, where each post was manually coded by three independent annotators across the six dimensions described in Section 4.1.1.

Table 2: Unsupervised domain adaptation (UDA)—DANN.

(Source/Target) → Val	Hum2000		Hum3000		Med5000	
	Acc	F1	Acc	F1	Acc	F1
Hum3000/Hum2000	0.81	0.90	0.94	0.97	0.76	0.86

Note: Hum2000 posts of Humanities & Science. Hum3000 posts of Humanities/Science—Statistical Learning, Winter 2014. Med5000 posts of Medicine. The Hum3000 column reports validation on the source domain itself and is therefore not informative for adaptation; Hum2000 and Med5000 are the genuine target-domain evaluations, with Hum3000→Hum2000 representing a within-domain shift and Hum3000→Med5000 the only genuinely heterogeneous (cross-domain) pair.

Table 3: Training configuration and hyperparameters for the DANN classifier.

Parameter	Value
Epochs	300
Optimizer	Adam
Initial learning rate	0.001
Weight decay	10^{-5}
Learning rate scheduler patience	5 epochs
Loss function	Cross-entropy
Text embeddings	BERT
Evaluation runs	5 (different seeds)

Table 4 presents the classification performance of the DANN model on the test set, evaluated both overall and separately for each domain. In this experiment, we used the Hum3000 subset as the source domain for DANN adaptation, and the student posts as the target domain.

Table 4: DANN classification performance on question detection.

Metric	Value	Standard Deviation
Overall Accuracy	87.50%	±1.2%
Source Domain Accuracy	88.33%	±1.5%
Target Domain Accuracy	86.67%	±1.8%
Domain Accuracy Gap	1.66%	–
Domain Classifier Accuracy	52.08%	±2.1%

Note: Source: Hum3000 posts of Humanities/Science—Statistical Learning, Winter 2014. Target: 300 posts of 50 students of discipline.

Table 5 provides detailed classification metrics—including precision, recall, and F1-score—for both classes within the student dataset.

Table 5: Detailed classification metrics by class.

Class	Precision	Recall	F1-Score
Non-Question	0.86	0.88	0.87
Question	0.89	0.87	0.88
Macro Average	0.875	0.875	0.875

Table 6 and Fig. 4 compare the DANN performance across domains with a baseline model without domain adaptation.

Table 6: Comparison of DANN vs. baseline model across domains (logistic regression).

Model	Source Domain	Target Domain
Baseline (No Adaptation)	89.17%	78.33%
DANN (With Adaptation)	88.33%	86.67%
Performance Gap	−0.84%	+8.34%

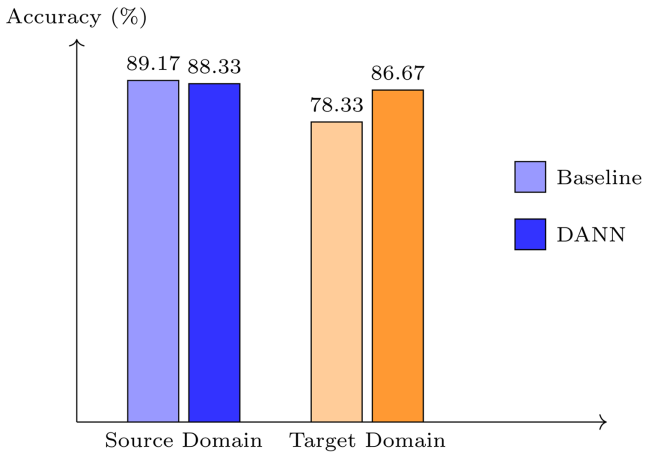


Figure 4: Baseline vs. DANN accuracy across source and target domains (data from Table 6).

The experimental results demonstrate the effectiveness of DANN for cross-domain question classification. Several key observations emerge from the analysis:

Domain Adaptation Success: The small accuracy gap between source and target domains (1.66%) indicates successful domain adaptation. This contrasts sharply with the baseline model, which exhibited a 10.84% performance degradation on the target domain.

Domain Confusion: The domain classifier achieved an accuracy of 52.08%, which is close to random guessing (50%). This near-chance performance is a positive indicator that the feature extractor successfully learned domain-invariant representations, as intended by the adversarial training mechanism.

Balanced Performance: The model maintained high and balanced precision and recall for both question and non-question classes (F1-scores of 0.88 and 0.87, respectively), suggesting that the domain adaptation did not compromise the discriminative power of the learned features for the primary classification task.

Trade-Off Analysis: While DANN showed a marginal decrease in source domain performance (−0.84% compared to the baseline), this trade-off was more than compensated by the substantial improvement in target domain performance (+8.34%). This demonstrates the model's enhanced generalization capability across domains.

4.3 Results—End-to-End Tests

In this implementation, the interval between curation engine actions was set to $\tau = 24$ h. The fraction of eligible messages addressed during each action cycle was determined using a baseline rate of $f = 50\%$. Specifically, letting n_c represent the total number of detected messages requiring a response, and n_a represent the actual number of messages addressed, the rule applied was: if $n_c < 5$, then $n_a = n_c$; otherwise, $n_a = \max(5, \lceil f \cdot n_c \rceil)$. In this implementation, simple state representation and transition function models were used. In the simplest model, each active learner in the forum can occupy one of two states: 1 (waiting for an answer to their question) and 0 (not waiting for an answer). The state update function for learner i, denoted as $s_i = f_s(s_i, m)$—where m indicates a reply message directed at learner i and '–' indicates the absence of a reply—is defined as $f_s(0, -/m) = 0$, $f_s(1, m) = 0$, and $f_s(1, -) = 1$. Here, m may be a reply generated either by a peer or by the curation engine.

Table 7 and Fig. 5 summarize the RAGAS metric scores for all five models evaluated. Scores are normalized to the range [0, 1], where higher values indicate better performance.

Table 7: RAG evaluation scores across five state-of-the-art language models (scale: 0 to 1).

Model	Faithfulness	Answer Relevance	Context Recall
GPT-4o	0.9267	0.9481	1.0000
Claude 3.5 Sonnet	0.9500	0.9221	0.9800
Gemini 2.5 Pro	0.8500	0.8800	0.9800
Llama 3.3 70B	0.8000	0.8500	0.9800
DeepSeek-V3	0.6215	0.8242	1.0000
Mean	0.8296	0.8849	0.9880

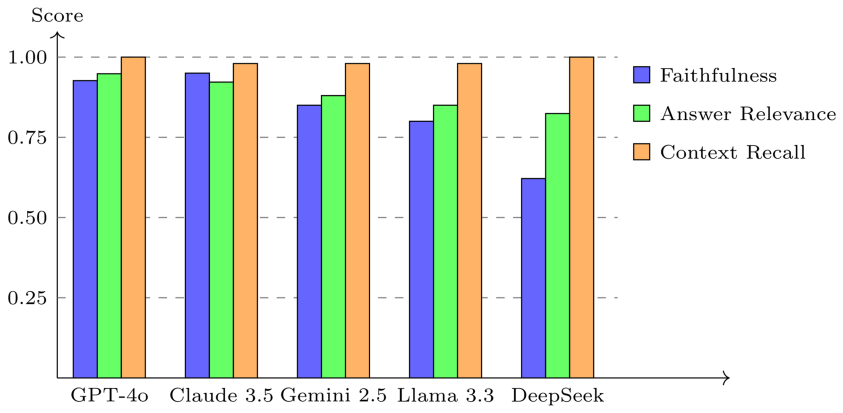


Figure 5: RAGAS metric scores across the five evaluated language models (data from Table 7).

Retrieval Quality (Context Recall): The retriever (*text-embedding-ada-002*) demonstrates consistently high performance across all models, with a mean Context Recall of 0.9880. GPT-4o and DeepSeek-V3 achieve perfect retrieval coverage (1.0000), while Claude 3.5 Sonnet, Gemini 2.5 Pro, and Llama 3.3 70B score 0.9800. This minor variance likely reflects edge cases where ground-truth evidence spans multiple non-contiguous text segments. Since the retriever, queries, and reference answers are identical across all five rows of Table 7, this small variation is attributed to scoring stochasticity in the RAGAS LLM-judge rather than an actual difference in retrieval behavior across generators. Importantly, retrieval quality is largely independent of the generative model, confirming that differences in other metrics stem from generation behavior.

Generation Fidelity (Faithfulness): Faithfulness scores exhibit substantial variation, ranging from 0.6215 (DeepSeek-V3) to 0.9500 (Claude 3.5 Sonnet). Claude 3.5 Sonnet and GPT-4o obtain the highest scores (0.9500 and 0.9267, respectively); given the modest sample size (n = 150) and the absence of confidence intervals or a significance test across repeated runs, this difference is not shown to be statistically significant and should not be read as a definitive ranking. Gemini 2.5 Pro and Llama 3.3 70B occupy a middle tier (0.8500 and 0.8000, respectively), while DeepSeek-V3 lags significantly. This pattern suggests that frontier proprietary models (Claude, GPT-4o) have stronger instruction-following and context-adherence capabilities compared to open-weight alternatives. The low faithfulness of DeepSeek-V3 indicates a tendency to generate plausible-sounding but unverified claims, which is particularly problematic in educational settings.

Response Utility (Answer Relevance): Answer Relevance scores are generally high (mean: 0.8849), indicating that all models produce responses that are semantically aligned with student questions. GPT-4o and Claude 3.5 Sonnet obtain the highest relevance scores (0.9481 and 0.9221, respectively); as with Faithfulness, this difference is not supported by a formal significance test. The narrower spread in this metric compared to Faithfulness suggests that modern LLMs are competent at addressing question intent, but differ more in their ability to ground answers in provided evidence.

Error Analysis. Beyond aggregate metrics, manual inspection of model outputs reveals systematic patterns:

- **GPT-4o** produces concise, well-structured answers with clear references to retrieved contexts. Occasionally, it omits minor details present in the source material, prioritizing readability over exhaustiveness.
- **Claude 3.5 Sonnet** demonstrates exceptional grounding: answers closely paraphrase retrieved evidence and avoid extrapolation. However, responses can be verbose, sometimes restating context verbatim rather than synthesizing information.
- **Gemini 2.5 Pro** balances detail and conciseness effectively but occasionally introduces minor terminological variations not present in the source (e.g., using "primary cache" instead of "L1 cache"), which affects faithfulness scores.
- **Llama 3.3 70B** provides generally accurate answers but exhibits occasional factual drift, particularly on quantitative questions requiring precise formula application.
- **DeepSeek-V3** frequently generates contextually plausible but unsupported elaborations. For example, when asked about cache replacement policies, it might mention additional policies (e.g., MRU, FIFO variants) not discussed in the retrieved context, leading to low faithfulness despite high retrieval coverage.

Case Studies. To illustrate these patterns concretely, Table 8 revisits three representative student questions introduced in Section 4.1.2, together with the qualitative behavior observed across models for each cognitive difficulty level.

Table 8: Illustrative case studies of model behavior across question types.

Question	Type	Observed Behavior
"What is RAM memory?"	Definitional	Concise, well-grounded answers across all models (low difficulty, consistently high faithfulness).
"Why does increasing associativity reduce conflict misses?"	Conceptual	GPT-4o and Claude 3.5 Sonnet provide accurate, source-grounded explanations; Gemini 2.5 Pro occasionally introduces non-source terminology (e.g., "primary cache" for "L1 cache").
"How do we calculate effective memory access time?"	Quantitative	Llama 3.3 70B shows occasional factual drift on formula application; DeepSeek-V3 elaborates with unsupported details (e.g., mentioning cache replacement policies not present in the retrieved context).

4.4 Discussion

Interpretation of Experimental Results. DANN, while effective, demonstrated variability in performance, especially when applied to more heterogeneous or semantically distant domains, such as the Med5000 dataset. These variations can be attributed to the adversarial nature of DANN's training process, which seeks to learn domain-invariant features through a minimax game between the feature extractor and a domain discriminator. While powerful, this approach tends to be more sensitive to semantic shifts, out-of-distribution features, and variations in textual length or vocabulary richness—factors frequently encountered in real-world short text classification tasks.

The results confirm a nuanced relationship between retrieval quality and generation fidelity, consistent with prior RAG benchmarking literature [30,31]. High Context Recall (≈0.99 across models) demonstrates that the retriever successfully identifies relevant textbook passages. However, this retrieval success does not automatically translate to faithful generation: DeepSeek-V3 achieves perfect retrieval coverage yet produces the least grounded answers (Faithfulness: 0.6215).

This decoupling underscores a critical pedagogical implication: *retrieval adequacy is necessary but insufficient for trustworthy educational assistance*. A system that retrieves correct evidence but generates unfaithful answers may be more harmful than one that admits ignorance, as students are unlikely to cross-verify LLM outputs against source material [33].

For deployment in educational contexts, different models present distinct trade-offs:

- **Claude 3.5 Sonnet** emerges as the most pedagogically reliable choice, balancing high faithfulness (0.9500) with strong relevance (0.9221). Its tendency toward verbosity may actually benefit learners who prefer detailed explanations with explicit source attribution.
- **GPT-4o** offers the best overall user experience (highest Answer Relevance: 0.9481) while maintaining strong faithfulness (0.9267). Its concise response style suits quick clarifications during problem-solving sessions.
- **Gemini 2.5 Pro** and **Llama 3.3 70B** represent viable middle-tier options, particularly if cost or open-weight accessibility are constraints. However, their lower faithfulness scores (0.8500 and 0.8000) necessitate additional verification mechanisms or confidence thresholds.

- **DeepSeek-V3**, despite strong retrieval and reasonable relevance, exhibits unacceptable hallucination risks (Faithfulness: 0.6215) for unsupervised educational deployment. Its outputs would require extensive human oversight.

4.4.1 Limitations and Threats to Validity

Several limitations warrant discussion:

- **Dataset Size:** With 150 instances, statistical power for detecting small performance differences is limited. However, the observed faithfulness gaps (e.g., 0.6215 vs. 0.9500) are substantial and unlikely to be artifacts of sampling variance.
- **Domain Specificity:** Results are derived from a single textbook on computer architecture. Generalization to other technical domains (e.g., operating systems, networking, databases) or non-STEM learners requires further validation.
- **LLM-as-a-Judge Bias:** RAGAS metrics rely on LLM-based evaluation, which may favor models stylistically similar to the judge [34]. We mitigate this by using a separate evaluator model (GPT-3.5 for judgment), but residual bias cannot be eliminated. GPT-3.5 was deliberately chosen as a model from a different generation than the LLMs being judged (GPT-4o, Claude 3.5 Sonnet, Gemini 2.5 Pro, Llama 3.3 70B, DeepSeek-V3), reducing the risk of a judge favoring outputs from its own model family; a more capable judge (e.g., GPT-5) was considered but not adopted due to its substantially higher API cost at the evaluation scale used here. We acknowledge, following [34] and the broader LLM-as-a-judge literature, that a judge weaker than some of the evaluated candidates may still introduce reliability limitations not fully eliminated by this choice.
- **Temporal Validity:** LLM capabilities evolve rapidly. Models evaluated here (as of early 2026) may be superseded by newer versions with improved grounding mechanisms.
- **Prompt Engineering:** Performance is sensitive to prompt design. We used a standard RAG prompt template; domain-specific prompt optimization might alter relative model rankings.
- **Scope of Evaluation:** This study does not include an ablation study over the classifier's state representation, a computational efficiency analysis (training/inference time, memory footprint), nor duplicate-question detection or topic-discovery evaluation, since these functionalities are outside the scope of the implemented curation engine. These are identified as directions for future work.
- **Undisclosed Deployment Parameters:** The confidence threshold th, the exact number of posts processed, the observation window, and the counting rubric (including inter-rater agreement, if any) behind the reported 2% clarification rate and 15% supplement rate are not currently available and are not disclosed in this revision; reporting them with full methodological detail is left for a future revision.

Nonetheless, the methodology provides a reproducible baseline for educators and system designers exploring RAG deployment in technical education.

4.4.2 Pedagogical Aspects

Educational Impact. Assessing the impact of automated and timely responses on student learning and dropout rates requires creating a control and test experimental setup, which is tremendously difficult and beyond the scope of this research. However, the literature indicates that this impact is strong in both reducing dropout [35] and promoting student success [36]. Thus, we sought indirect evidence of the positive impact on learning.

Benefits for Learners and Instructors. Beyond automated metric evaluation, we collected qualitative evidence from student and instructor perspectives to assess the educational value of the RAG assistant.

The deployed generative LLM was GPT-4o. During the semester-long deployment, students also utilized the system for exam preparation, particularly during the week preceding each assessment period. Access logs indicate concentrated usage patterns aligned with the academic calendar, with students formulating clarification questions about textbook content while reviewing course material. The immediacy of responses emerged as a critical affordance: whereas forum posts to human instructors typically required several hours to receive attention, the RAG assistant provided contextualized, rapid answers. This temporal advantage proved especially valuable during asynchronous study sessions outside regular office hours, when instructor availability was limited.

Student feedback, gathered through reviews and informal instructor observations, indicated predominantly positive reception of the assistant. When asked to rate answer quality and pedagogical usefulness, students reported satisfaction with the conciseness and textbook alignment of generated responses. Importantly, the system succeeded in addressing the majority of student information needs without instructor intervention: quantitative analysis of forum activity revealed that only 2% of questions required subsequent clarification or correction by teaching staff. This low intervention rate suggests that the high faithfulness scores observed in automated evaluation translated into practical reliability in authentic educational contexts. The remaining 98% of interactions were resolved entirely through the RAG assistant, demonstrating scalability potential for large-enrollment courses.

Scalability and Practical Deployment. From the instructor perspective, the deployment yielded measurable reductions in routine support workload. Teaching assistants reported that the assistant effectively handled definitional queries, procedural clarifications, and concept review questions that previously consumed substantial forum moderation time. This delegation allowed instructional staff to reallocate effort toward higher-order pedagogical activities such as addressing misconceptions, facilitating conceptual discussions, and providing personalized feedback on student assignments. Post-deployment analysis indicated that 15% of assistant-generated responses received minor supplementary explanations from instructors, typically to provide worked examples or connect concepts to broader course themes. Crucially, these supplements represented refinements rather than corrections, indicating that the assistant established a reliable foundation upon which instructors could build, rather than requiring remediation of inaccurate information. This human-AI complementarity suggests a viable operational model wherein RAG systems handle first-tier informational support, while human expertise remains accessible for complex or ambiguous queries.

5 Conclusion

This paper proposed and evaluated a curation engine for the automatic triage and response to subject-matter questions in online course forums. The system was built upon two evaluated components—domain-adapted question detection and instructional RAG—along with a formal design proposal for Markovian forum-state modeling. We summarize the findings for each evaluated component below before discussing the engine's broader implications and directions for future work.

The application of DANN for question classification across heterogeneous educational datasets proved highly effective. The model successfully detected student questions with 87.50% overall accuracy while maintaining robust performance across both source and target domains. The negligible domain accuracy gap (1.66%) and the near-random domain classifier performance (52.08%) confirm that the DANN architecture succeeded in learning domain-invariant features without sacrificing task-specific discriminative power.

These results suggest that DANN is a promising approach for building question detection systems that can generalize across diverse educational contexts, writing styles, and course materials.

Furthermore, this study presented a comprehensive evaluation of RAG for supporting computer architecture education through textbook-grounded question answering. We addressed the dual challenges of timely student support and pedagogical alignment by constraining LLM generation to authoritative instructional material. Our empirical study compared five state-of-the-art language models (GPT-4o, Claude 3.5 Sonnet, Gemini 2.5 Pro, Llama 3.3 70B, DeepSeek-V3) using RAGAS metrics: Context Recall, Faithfulness, and Answer Relevance.

Retrieval-Augmented Generation offers a promising pathway for scalable, pedagogically aligned student support in technical education. By grounding generation in curated instructional content, RAG systems can mitigate hallucination risks inherent in pure LLMs while preserving conversational flexibility. Our multi-model evaluation demonstrates that careful model selection and continuous quality monitoring are essential for educational deployment. As LLM capabilities continue to advance, the integration of RAG-based assistants into computer architecture and broader STEM education represents an actionable strategy to enhance learning outcomes while respecting curriculum integrity.

Future Work. Several extensions naturally follow from the scope boundaries identified throughout this study. First, integrating the curation engine directly with Learning Management Systems (LMS) beyond the Google Classroom deployment reported here would broaden applicability across institutional contexts. Second, incorporating explainability mechanisms into the post-classification and prioritization decisions could increase instructor trust and facilitate auditing of the engine's actions. Third, extending the use of LLMs beyond response generation to forum moderation tasks, such as duplicate-question identification and topic discovery, represents a promising direction not covered by the present framework. Finally, evaluating the full 4-tuple Markovian state representation (urgent, confusion, sentiment, engagement) against the simplified binary state used in our end-to-end experiment, together with a sensitivity analysis of the operating parameters τ and f, remains an open avenue for future research. Personalized learning support, wherein the curation engine adapts its prioritization and response style to individual learner profiles, is a further direction we consider promising for large-enrollment courses.

Acknowledgement: We thank all the volunteers who participated in this research and the university for providing the space and data for this experiment.

Funding Statement: This work was carried out with the support of the Coordination for the Improvement of Higher Education Personnel—Brazil (CAPES)—Funding Code 001. Process Number 88887.006841/2024-00.

Author Contributions: The authors confirm contribution to the paper as follows: conceptualization, Antonio Leandro Martins Candido and José Everardo Bessa Maia; methodology, Antonio Leandro Martins Candido and José Everardo Bessa Maia; software, Antonio Leandro Martins Candido; validation, José Everardo Bessa Maia; formal analysis, José Everardo Bessa Maia; investigation, Antonio Leandro Martins Candido and José Everardo Bessa Maia; resources, Antonio Leandro Martins Candido and José Everardo Bessa Maia; data curation, Antonio Leandro Martins Candido and José Everardo Bessa Maia; writing—original draft preparation, Antonio Leandro Martins Candido and José Everardo Bessa Maia; writing—review and editing, Antonio Leandro Martins Candido and José Everardo Bessa Maia; visualization, Antonio Leandro Martins Candido and José Everardo Bessa Maia; supervision, José Everardo Bessa Maia; project administration, Antonio Leandro Martins Candido and José Everardo Bessa Maia; funding acquisition, Antonio Leandro Martins Candido. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The research complied with the General Data Protection Regulation (GDPR). All collected data was securely stored on protected servers, with access restricted to authorized members of the research team. The data can be made available upon request to the corresponding author.

Ethics Approval: The experiment was conducted via Google Classroom with a class of 50 students (aged 18 or older). The data collection was carried out retrospectively and comprised strictly anonymized textual questions derived from standard pedagogical activities, without involving sensitive personal data. In accordance with Brazilian regulations (National Health Council Resolution No. 510/2016), this type of non-interventional educational data analysis is exempt from formal Institutional Review Board (CEP) evaluation. Informed consent was obtained from all participants, and the study complies with the Declaration of Helsinki.

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

References

1. An Y, Chandresekaran MK, Kan M, Fu Y. The MUIR framework: cross-linking MOOC resources to enhance discussion forums. arXiv:2005.07382. 2020.
2. Mudau PK, den Berg GV. Guidelines for supporting a community of inquiry through graded online discussion forums in higher education. Educ Sci. 2023;13(9):963. doi:10.3390/educsci13090963.
3. de Lima DP, Gerosa MA, Conte TU, de M Netto JF. What to expect, and how to improve online discussion forums: the instructors' perspective. J Internet Serv Appl. 2019;10(1):22. doi:10.1186/s13174-019-0120-0.
4. Garg A, Wadhwa D, Gomes NM, Sahasrabudhe S. Effects of intervention implemented in MOOC discussion forums: contributing factors and participation analysis. Educ Inf Technol. 2025;30(18):25753–88. doi:10.1007/s10639-025-13754-8.
5. Mihailidis P, Cohen JN. Exploring curation as a core competency in digital and media literacy education. J Interact Media Educ. 2013;2013(1):1–19. doi:10.5334/2013-02.
6. Agrawal A, Venkatraman J, Leonard S, Paepcke A. YouEDU: addressing confusion in MOOC discussion forums by recommending instructional video clips. In: Proceedings of the International Conference on Educational Data Mining 2015; 2015 Jun 26–29; Madrid, Spain.
7. Rahma BD, Koutheair KM. Towards a framework for building automatic recommendations of answers in MOOCs' discussion forums. In: Proceedings of the 2019 7th International Conference on ICT & Accessibility (ICTA); 2019 Dec 13–15; Hammamet, Tunisia. p. 1–6. doi:10.1109/ICTA49490.2019.9144850.
8. Mitra C, Miroyan M, Jain R, Kumud V, Ranade G, Norouzi N. RetLLM-E: retrieval-prompt strategy for question-answering on student discussion forums. Proc AAAI Conf Artif Intell. 2024;38(21):23215–23. doi:10.1609/aaai.v38i21.30368.
9. Liu Z, Xing W, Jiao X, Li C. Exploring fairness and explainability in LLM-generated support for online learning discussion forums. J Learn Anal. 2025;12(3):8–33. doi:10.18608/jla.2025.8885.
10. Abdelmagied M, Chatti MA, Joarder S, Ain QU, Alatrash R. Leveraging graph retrieval-augmented generation to support learners' understanding of knowledge concepts in MOOCs. arXiv:2406.00954. 2025.
11. Al Shloul T, Mazhar T, Abbas Q, Iqbal M, Ghadi YY, Shahzad T, et al. Role of activity-based learning and ChatGPT on students' performance in education. Comput Educ Artif Intell. 2024;6(3):100219. doi:10.1016/j.caeai.2024.100219.
12. Mallek F, Mazhar T, Shah SFA, Ghadi YY, Hamam H. A review on cultivating effective learning: synthesizing educational theories and virtual reality for enhanced educational experiences. PeerJ Comput Sci. 2024;10(3):e2000. doi:10.7717/peerj-cs.2000.
13. Bakharia A. Towards cross-domain MOOC forum post classification. In: Proceedings of the Third ACM Conference on Learning@ Scale; 2016 Apr 25–26; Scotland, UK. p. 253–6.
14. Dogra V, Verma S, Kavita, Chatterjee P, Shafi J, Choi J, et al. A complete process of text classification system using state-of-the-art NLP models. Comput Intell Neurosci. 2022;2022(1):1883. doi:10.1155/2022/1883698.
15. Luo W, Zhang L. Question text classification method of tourism based on deep learning model. Wirel Commun Mob Comput. 2022;2022(1):4330701. doi:10.1155/2022/4330701.
16. Liu S, Liu S, Sha L, Zeng Z, Gašević D, Liu Z. Annotation guidelines-based knowledge augmentation: towards enhancing large language models for educational text classification. IEEE Trans Learn Technol. 2025;18:619–34. doi:10.1109/TLT.2025.3570775.

17. Ntourmas A, Daskalaki S, Dimitriadis Y, Avouris N. Classifying MOOC forum posts using corpora semantic similarities: a study on transferability across different courses. Neural Comput Appl. 2023;35(1):161–75. doi:10.1007/s00521-021-05750-z.
18. Capuano N, Caballé S, Conesa J, Greco A. Attention-based hierarchical recurrent neural networks for MOOC forum posts analysis. J Ambient Intell Humaniz Comput. 2021;12(11):9977–89. doi:10.1007/s12652-020-02747-9.
19. El-Rashidy MA, Farouk A, El-Fishawy NA, Aslan HK, Khodeir NA. New weighted BERT features and multi-CNN models to enhance the performance of MOOC posts classification. Neural Comput Appl. 2023;35(24):18019–33. doi:10.1007/s00521-023-08673-z.
20. Wise AF, Cui Y, Jin W, Vytasek J. Mining for gold: identifying content-related MOOC discussion threads across domains through linguistic modeling. Internet High Educ. 2017;32(4):11–28. doi:10.1016/j.iheduc.2016.08.001.
21. Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In: Proceedings of the 34th International Conference on Neural Information Processing Systems; 2020 Dec 6–12; Vancouver, BC, Canada. p. 9459–74.
22. Wang N, Singh S. Answering students' questions on course forums using multiple chain-of-thought reasoning and finetuning RAG-enabled LLM. arXiv:2511.09831. 2025.
23. Bhattacharjee SD, Gokaraju JSAV, Yuan J, Kalwa A. Multi-view knowledge graph for explainable course content recommendation in course discussion posts. In: Proceedings of the 2022 26th International Conference on Pattern Recognition (ICPR); 2022 Aug 21–25; Montréal, QC, Canada. p. 2785–91. doi:10.1109/ICPR56361.2022.9956098.
24. Chen E, Li C, Li S, Xiao Z, Lin J, Koedinger KR. Comparing RAG and GraphRAG for page-level retrieval question answering on math textbook. arXiv:2509.16780. 2025.
25. Russell S, Norvig P. Artificial intelligence: a modern approach. 4th ed. Harlow, UK: Pearson Education Limited; 2022.
26. Jia B, Zhang ML. Multi-dimensional classification: paradigm, algorithms and beyond. Vicinagearth. 2024;1(3):1–27. doi:10.1007/s44336-024-00004-7.
27. Candido ALM, Maia JEB. Detecting urgent instructor intervention need in learning forums with a domain adaptation. In: Intelligent systems design and applications. Berlin/Heidelberg, Germany: Springer; 2023. p. 502–12. doi:10.1007/978-3-031-35507-3_49.
28. Ganin Y, Ustinova E, Ajakan H, Germain P, Larochelle H, Laviolette F, et al. Domain-adversarial training of neural networks. J Mach Learn Res. 2016;17(59):1–35. doi:10.1007/978-3-319-58347-1_10.
29. Holtzman A, Buys J, Du L, Forbes M, Choi Y. The curious case of neural text degeneration. arXiv:1904.09751. 2019.
30. Saad-Falcon J, Khattab O, Potts C, Zaharia M. ARES: an automated evaluation framework for retrieval-augmented generation systems. In: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 2024 Jun 16–21; Mexico City, Mexico. p. 338–54. doi:10.18653/v1/2024.naacl-long.20.
31. Chen J, Lin H, Han X, Sun L. Benchmarking large language models in retrieval-augmented generation. Proc AAAI Conf Artif Intell. 2024;38(16):17754–62. doi:10.1609/aaai.v38i16.29728.
32. Es S, James J, Anke LE, Schockaert S. RAGFAS: automated evaluation of retrieval augmented generation. In: Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations; 2024 Mar 17–22; St. Julian's, Malta. p. 150–8. doi:10.18653/v1/2024.eacl-demo.16.
33. Yan L, Sha L, Zhao L, Li Y, Martinez-Maldonado R, Chen G, et al. Practical and ethical challenges of large language models in education: a systematic scoping review. Br J Educ Technol. 2024;55(1):90–112. doi:10.1111/bjet.13370.
34. He X, Lin Z, Gong Y, Jin AL, Zhang H, Lin C, et al. AnnoLLM: making large language models to be better crowdsourced annotators. In: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 2024 Jun 16–21; Mexico City, Mexico. p. 165–90. doi:10.18653/v1/2024.naacl-industry.15.
35. Hone KS, El Said GR. Exploring the factors affecting MOOC retention: a survey study. Comput Educ. 2016;98(2):157–68. doi:10.1016/j.compedu.2016.03.016.
36. Van der Kleij FM, Feskens RC, Eggen TJ. Effects of feedback in a computer-based learning environment on students' learning outcomes: a meta-analysis. Rev Educ Res. 2015;85(4):475–511. doi:10.3102/0034654314564881.