An Explainable Graph Enhanced Ensemble for Confusion Detection and Multiple Strategy Intervention in Massive Open Online Course Discussion Forums

Abdennour Redjaibia, Samia Drissi, Karima Boussaha, Sevinç Gülseçen, and Yacine Lafifi

***Abstract*—Learner confusion in Massive Open Online Course (MOOC) forums is a critical precursor to dropout. Existing detection methods often treat posts in isolation, ignoring structural discussion dynamics. We present ConFusionGraph, an end-to-end framework integrating confusion detection, explainable analysis, and personalized intervention. The methodology models forum interactions as heterogeneous graphs with five relation types capturing thread structure, authorship, and temporal dynamics. We employ a stacked ensemble combining gradient boosting, transformer-based language models, and a heterogeneous graph neural network. To ensure transparency, dual local and global feature attribution analyses reveal that while explicit flags dominate, structural features—such as thread mean confusion and reply degree—are top contributors. Furthermore, attribution-based clustering identifies two primary mechanisms: active questioning and confusion contagion. Experiments on a large-scale dataset (nearly 30,000 posts across 11 courses) demonstrate an accuracy of 89.2% and an AUC of 0.923. Confusion chain analysis provides empirical evidence of propagation through threads and cross-thread contagion. These findings drive a multi-strategy intervention system offering peer answer retrieval, thread summarization, and personalized clarification. This provides a robust, context-aware mechanism to mitigate learner dropout through timely support.**

***Index Terms*—Confusion Detection, Explainable Artificial Intelligence, Graph Neural Networks, Intervention Systems, MOOC Discussion Forums, SHAP.**

I. Introduction

Massive Open Online Courses (MOOCs) have democratized access to education by reaching learners worldwide; however, they suffer from severe retention issues, as approximately 90% of students drop out, [1]. Among the factors contributing to this attrition, learner confusion has been identified as a critical affective state that, if left unaddressed, leads to frustration, disengagement, and eventual dropout [2], [3]. Discussion forums serve as a key communication and social learning tool where learners interact and seek help [4]; however, the massive participant-to-instructor ratio makes it impractical for instructors to manually identify and respond to every confused post. This challenge has motivated research into automated detection and classification models, including approaches leveraging deep learning on user interaction behaviors [5] and advanced natural language processing (NLP) techniques to analyze post semantics [6].

While these approaches have achieved promising results, they suffer from three fundamental limitations. First, existing methods treat forum posts as isolated text documents, ignoring the rich structural context of discussion threads. A confused post does not exist in isolation; it is embedded in a thread with replies, connected to a user who participates in multiple discussions, and situated in a temporal sequence of forum activity. Second, the adoption of deep learning models has introduced a ”black-box” problem where instructors cannot understand why a post is classified as confused, undermining trust and actionability [7]. Third, most studies stop at detection without addressing the critical question of how to intervene once confusion is identified. To address these gaps, this paper presents ConFusionGraph, an end-to-end framework integrating a graph-enhanced stacked ensemble for detection, a dual explainable artificial intelligence (XAI) analysis for token-level attribution, and a multi-strategy intervention system.

The primary objective of this research is to determine if modeling forums as heterogeneous graphs provides superior predictive signals compared to text-only features, while identifying the linguistic patterns that drive confusion. Furthermore, the study explores whether explainable models can reveal interpretable confusion archetypes to guide personalized interventions. The methodology employs a heterogeneous graph formulation capturing thread topology, user behavior, and temporal dynamics through five distinct edge types. This is integrated into a stacked ensemble combining gradient boosting, transformer models, and graph neural network experts. Explainability is achieved through dual feature attribution methods, leading to the discovery of a data-driven confusion archetype taxonomy—specifically distinguishing between ”Active Questioning” and ”Confusion Contagion.” Experimental results on a large-scale dataset demonstrate that the

Manuscript received April 23, 2026; revised May 11, 2026. Date of publication August 17, 2026. Date of current version August 17, 2026.

A. Redjaibia and S. Drissi are with the Computer Science and Mathematics Laboratory (LIM), Department of Computer Science, University of Souk Ahras, Algeria (e-mails: a.redjaibia@univ-soukahras.dz, s.drici@univ-soukahras.dz).

K. Boussaha is with the Intelligence and Autonomous Things Laboratory (LIAOA), Department of Mathematics and Computer Science, University of Oum El Bouaghi, Algeria (e-mail: karima.boussaha@univ-oeb.dz).

S. Gülseçen is with the Department of Computer Science, Istanbul University, Turkey (e-mail: gulsecen@istanbul.edu.tr).

Y. Lafifi is with the LabSTIC Laboratory, University of 8 May 1945 Guelma, Algeria (e-mail: lafifi.yacine@univ-guelma.dz).

Digital Object Identifier (DOI): 10.24138/jcomss-2026-0091

proposed framework not only achieves high accuracy and area under the receiver operating characteristic curve (AUROC) scores but also provides empirical evidence of confusion propagation across threads. Ultimately, this work demonstrates that a graph-based, explainable approach effectively bridges the gap between automated detection and meaningful pedagogical intervention. Specifically, this work addresses three research questions:

- *RQ1: Does heterogeneous graph modeling provide superior predictive signals over text-only approaches?*
- *RQ2: Which features and archetypes drive confusion predictions?*
- *RQ3: Can XAI-derived archetypes guide differentiated intervention strategies?*

The main contributions of this paper are summarized as follows:

- A heterogeneous graph formulation of MOOC discussion forums capturing five distinct relation types: reply-to, same-thread, authored-by, temporal-next, and user-co-thread edges, enabling structural confusion propagation modeling beyond isolated post classification.
- ConFusionGraph, a six-stage end-to-end pipeline integrating a three-expert stacked ensemble (XGBoost for metadata, BERT+MLP for text, and a heterogeneous GNN for structure) with a graph-structural meta-learner achieving 89.2% accuracy and AUC of 0.923 on the Stanford MOOCPost dataset.
- A dual explainability analysis combining global KernelSHAP, token-level SHAP, and local LIME attributions, revealing an asymmetric confusion vocabulary and validating that structural forum features carry genuine predictive signal across two independent explanation frameworks.
- An empirical confusion archetype taxonomy — Active Questioning (73.8%) and Confusion Contagion (26.2%) — discovered through SHAP-based clustering, providing the first data-driven evidence that a significant minority of forum confusion is structurally rather than linguistically triggered.
- A multi-strategy intervention framework defining five differentiated strategies based on XAI-derived archetypes, with three activated on the current dataset, achieving overall specificity of 0.742 compared to generic baseline responses.

The remainder of this paper is organized as follows. Section II reviews related work. Section III details the methodology. Section IV presents experimental results. Section V discusses findings and limitations. Section VI concludes and outlines future directions.

II. Related Work

This section surveys the literature across four interconnected areas that motivate the design of ConFusionGraph: confusion detection in MOOC forums, graph-based methods in educational data mining, explainable AI in education, and automated intervention systems. For each area, we identify the key advances and the gaps that this work addresses.

A. Confusion Detection in MOOC Forum

The arrival of deep learning transformed this landscape: Wei et al. [8] proposed a convolution-LSTM deep neural network for cross-domain classification of confusion, urgency, and sentiment, achieving competitive accuracy by sharing representations across source and target educational domains. More recent work has moved toward pre-trained transformer models. Geller et al. [9] applied a pre-trained BERT framework to predict confusion in student forums, demonstrating its generalizability across different academic terms of the same course. Their work supports the idea that transformer-based encoders can effectively capture confusion signals from raw text, although their approach still treated posts as independent units without explicitly modeling thread topology. Similarly utilizing natural language processing, Wu et al. [10] developed RE-MOOC, which applies retrieval-enhanced prompt learning and a pre-trained Edu-BERT model. Tested on a Chinese MOOC dataset, their approach reached a 95% accuracy and 93.94% F1-score. At the intersection of NLP and forum analytics, Yee et al. [11] developed BERT-based models to classify content category, structural role, and emotional valence in MOOC posts. To improve accuracy, they explicitly incorporated conversational thread topology by appending contextual metadata—such as post position and authorship status—to the text inputs. However, despite these advances in utilizing thread context, their method relies on standard transformer classifiers rather than explicitly modeling the broader authorship network, temporal sequence dynamics, or graph-based topology that characterize real, evolving forum interaction. Most recently, Liu et al. [12] investigated fairness and explainability in LLM-generated replies on MOOC discussion forums, fine-tuning GPT-2, Gemma, and LLaMA on the Stanford MOOCPost corpus and introducing absolute distributional sentiment divergence (ADSD) to quantify bias. While their work advances LLM-based support generation, it does not address the structural propagation of confusion or archetype-driven intervention routing.

While text-based models have advanced substantially, they remain blind to the relational structure of forum discussions. The following subsection reviews graph-based approaches that address this structural dimension.

B. Graph-Based Methods in Educational Data Mining

Graph Neural Networks (GNNs) have emerged as powerful tools for modeling relational structure in educational data. In student performance prediction, Li et al. [13] proposed Study-GNN, which constructs multi-topology graphs based on students' campus behaviors to model peer interactions. By applying GNNs for node-level classification, their approach outperforms traditional feature-based baselines by effectively capturing these latent social relationships among learners. Sun et al. [14] extended this to MOOCs through temporal graph networks constructed from learning activity logs, demonstrating that dynamic edge modeling captures trajectory

patterns unavailable to static methods. For text classification on heterogeneous graphs, Zhang et al. [15] proposed a heterogeneous GCN approach building subgraphs from text, word, and POS-tag nodes, applying multi-branch convolution to achieve robust short-text classification. Li et al. [16] extended heterogeneous GCNs to information networks augmented with external knowledge, demonstrating that inductive heterogeneous neighbor aggregation outperforms transductive methods and scales better to large corpora. More recently, the intersection of ensemble methods and GNNs has been shown to significantly improve classification accuracy on imbalanced educational datasets [17], directly motivating the hybrid approach adopted in ConFusionGraph. Hancox and Relton [18] proposed temporal graph-based CNNs for MOOC dropout prediction, demonstrating that incorporating time-ordered edges substantially improves prediction over static graph approaches. Separately, Roh et al. [19] developed SIG-Net, a GNN-based dropout predictor using student interaction graphs that outperforms behavioral feature baselines by capturing latent social relationships — an approach conceptually aligned with our user-co-thread edge type. Despite these advances, graph-based methods in educational data mining have primarily focused on student–course and student–concept relationships rather than the heterogeneous relational structures present in forum discussion data. Applying GNNs to forum data for affective state detection — rather than performance prediction or knowledge tracing — represents an underexplored direction that ConFusionGraph directly addresses. Graph-based models excel at capturing structural signals but often sacrifice interpretability. The following subsection addresses the growing body of work on explainability in educational AI systems.

C. Explainable AI in Education

The need for explainability in AI-driven educational systems has been widely recognized. Khosravi et al. [20] provided a comprehensive framework, XAI-ED, which highlights the necessity of tailoring AI explanations to the diverse needs of educational stakeholders. However, their analysis noted that the majority of existing educational XAI applications rely on inherently interpretable models, such as decision trees or rule-based systems, rather than applying post-hoc explanations to complex, opaque algorithms. As advanced deep learning methods are increasingly deployed for nuanced tasks—such as detecting affective states and confusion in forum text—there is a critical need to bridge this gap by integrating robust post-hoc methods (such as SHAP and LIME) to ensure these high-performing models remain transparent and actionable. In the context of dropout prediction, Nagy and Molontay [21] demonstrated that applying LIME and SHAP to gradient-boosted models trained on pre-enrollment features can identify the key drivers of dropout risk and communicate them to stakeholders in actionable form. Their work highlights the practical importance of both local per-student explanations and global feature importance for supporting targeted interventions. Deho et al. [22] specifically investigated whether learning analytics models should include sensitive demographic attributes, using SHAP to explain the contribution of these features, reinforcing that explainability must be integrated with fairness considerations. Türkmen [23] conducted a systematic review of XAI applications in educational research from 2019 to 2024, finding that SHAP and LIME are the dominant methods and that their primary value lies in increasing instructor trust in AI-based tools. This review notably identified a gap in XAI applied to natural language tasks in education — precisely the setting of MOOC confusion detection — which the present work addresses through dual SHAP-LIME analysis at global, local, and token levels across both tabular metadata and textual features. Despite advances in explainability, most XAI frameworks in education produce explanations without coupling them to automated support actions. The following subsection reviews intervention systems and their limitations.

D. Intervention Systems in MOOCs

Automated intervention systems for MOOC learners have been studied primarily through the lens of dropout prevention and self-regulated learning support. Bañeres et al. [24] developed a dynamic early warning system that identifies at-risk learners at the assessable activity level and triggers automated personalized messages, finding that goal-setting interventions significantly reduce dropout rates. Similarly, Cobos [25] presented a learning analytics system providing learner-facing dashboards and instructor-facing monitoring tools, demonstrating that automated feedback on engagement patterns improves learner motivation, engagement, and persistence in MOOCs. However, these systems detect risk at the course engagement level without differentiating the underlying confusion mechanisms, and their intervention strategies are uniform rather than differentiated by confusion type. They also do not address the challenge of identifying confused posts within an ongoing discussion thread, acting at the level of overall course engagement rather than at the level of specific forum interactions. More recent work has explored LLM-based automatic responses to forum posts, but these generate replies without any understanding of why a particular post is confused, creating a risk of mismatching intervention strategies to confusion types. Yu et al. [26] proposed MAIC, a system transforming passive MOOC video consumption into interactive LLM-driven dialogue, reporting reduced dropout rates; however, their approach does not differentiate intervention strategies by confusion type. Tan et al. [27] introduced ELF, an educational LLM framework for improving and evaluating AI-generated classroom content, demonstrating that retrieval-augmented generation substantially improves response factual grounding — a principle directly applied in our XAI-Conditioned Clarification strategy. Rebelo Marcolino et al. [28] demonstrated machine learning optimization for student dropout prediction using Moodle activity logs, highlighting the practical value of combining prediction with targeted intervention. These recent efforts reinforce the need for confusion-specific, archetype-driven intervention routing rather than generic LLM response generation.

E. Research Gaps

Despite the significant advancements in natural language processing, graph neural networks, and educational learning

analytics, a review of the literature reveals several critical gaps in how learner confusion is detected, interpreted, and managed in MOOC forums:

1) Isolation of Textual Data from Relational and Temporal Topology: Current state-of-the-art NLP models (including advanced pre-trained transformers) predominantly treat forum posts as independent, isolated textual units. While some incorporate basic contextual metadata, they fundamentally fail to explicitly model the highly dynamic conversational thread topology, the evolving authorship networks, and the temporal sequence of interactions that characterize real-world forum discourse.
2) Underutilization of Explainable AI in Unstructured Affective Analysis: While XAI techniques like SHAP and LIME have been successfully applied to structured educational data (such as predicting dropout risk from pre-enrollment or tabular behavioral features), their application to unstructured, affective textual data remains scarce. High-performing deep learning models used for confusion detection frequently remain opaque, providing predictions without the feature-level transparency required to build trust with educational stakeholders.
3) Coarse-Grained and Uniform Intervention Strategies: Existing automated intervention and early warning systems primarily operate at the macro-level of course engagement (e.g., predicting course failure or dropout based on activity logs). Consequently, when confusion is detected, the resulting interventions are typically generic and uniform. There is a lack of systems capable of acting at the micro-level of specific forum interactions to deliver fine-grained, targeted interventions that are differentiated by the specific type or underlying cause of a learner's confusion.

III. Methodology

The proposed ConFusionGraph framework follows a six-stage pipeline that progresses from raw forum data to actionable interventions. As illustrated in Fig. 1, the pipeline begins with data collection and preprocessing (Stage 01), where the Stanford MOOCPost dataset is cleaned, features are extracted, binary confusion labels are assigned, and temporal splits are created. The preprocessed data feeds into heterogeneous graph construction (Stage 02), which models forum interactions through two node types and five edge types. The graph and its features then serve as input to the ConFusionGraph stacked ensemble (Stage 03), where three specialized experts — XGBoost for metadata, BERT+MLP for text, and a heterogeneous GNN for structure — produce predictions that a meta-learner combines with graph structural features. In parallel, explainability analysis (Stage 04) applies KernelSHAP, LIME, and token-level attribution to uncover what drives confusion predictions and to discover data-driven confusion archetypes. Confusion chain extraction (Stage 05) analyzes temporal confusion dynamics within threads and across users, providing empirical evidence of confusion propagation. Finally, a multi-strategy intervention system (Stage 06) uses the archetypes discovered in Stage 04 to route each confused post to the most appropriate intervention strategy. The remainder of this section details each stage.

The implementation code is publicly available at `https://github.com/dznour/ConFusionGraph`.

A. Dataset and preprocessing

This study uses the Stanford MOOCPost dataset, comprising 29,604 forum posts from 11 courses across three academic domains (Education, Medicine, Humanities), authored by 11,042 unique users. Each post is annotated with confusion scores on a 1-7 Likert scale, along with binary flags for Question, Answer, and Opinion, plus continuous scores for Sentiment and Urgency.

We apply binary classification with a confusion threshold of 4.5 or higher, yielding 4,495 confused posts (15.2%) and 25,089 not-confused posts (84.8%). This threshold is chosen because 4.0 represents the distribution mode (annotation neutral point), and values at or above 4.5 represent genuine confusion signal. Table I reports sensitivity to the confusion threshold. At $\tau = 4.0$, 67.4% of posts are labeled confused, producing degenerate class imbalance. At $\tau = 5.0$, the positive class shrinks to 9.3%, collapsing Precision to 0.587. The threshold $\tau = 4.5$ yields the most balanced operating point with the highest Accuracy (89.2%) and AUROC (0.923).

TABLE I
THRESHOLD SENSITIVITY ANALYSIS — CONFUSIONGRAPH V2

Thresh.	Conf.%	Acc.	F1	Prec.	Rec.	AUROC
4.0	67.4	0.836	0.819	0.943	0.815	0.912
4.5	*15.2*	*0.892*	*0.795*	*0.549*	*0.806*	*0.923*
5.0	9.3	0.932	0.805	0.587	0.723	0.926

Of the 29,604 posts, 20 were excluded during graph construction (19 lacking required temporal or authorship metadata, 1 containing no text), yielding 29,584 post nodes. Data splitting follows a temporal strategy: the first 70% of posts per course (chronologically) form the training set, the next 15% form the validation set, and the final 15% form the test set. This simulates a realistic deployment scenario where the system trains on historical posts and predicts on future ones.

B. Heterogeneous Graph Construction

We model the MOOC forum as a heterogeneous graph $G = (V, E)$ with two node types and five edge types.

Node types. Post nodes (N = 29,584) carry 774-dimensional feature vectors: 768-dimensional BERT [CLS] embeddings extracted from bert-base-uncased with max sequence length 256, concatenated with 6 metadata features. User nodes (N = 11,042) carry 6-dimensional feature vectors aggregating posting behavior.

Edge types. (1) Reply-to (post to post): 8,329 directed edges linking replies to thread roots. (2) Same-thread (post to post, bidirectional): connecting posts within the same thread, capped at 5 nearest temporal neighbors for large threads to avoid computational explosion. (3) Authored-by (post to user): 29,584 edges linking each post to its author. (4) Temporal-next (post to post): connecting consecutive posts within the same course,

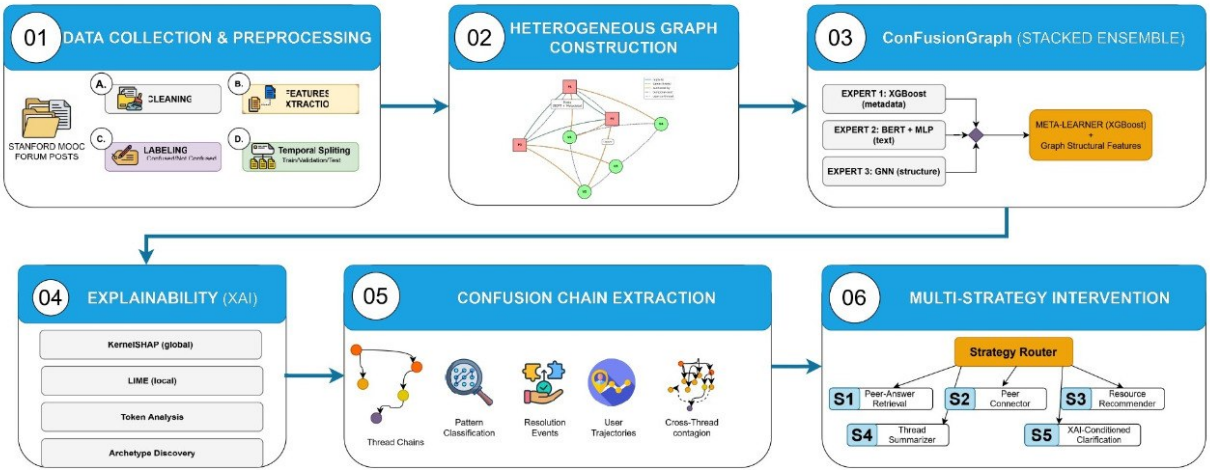


Fig. 1. Overview of the ConFusionGraph framework

weighted by exponential time decay with $\tau = 24\,\text{h}$. (5) User-co-thread (user to user, bidirectional): connecting users who participated in the same discussion thread. Fig. 2 illustrates a representative subgraph from the constructed heterogeneous graph, showing how post and user nodes are interconnected through the five edge types, with node colors indicating confusion levels. The resulting graph contains approximately 40,600 nodes and 105,000 edges.

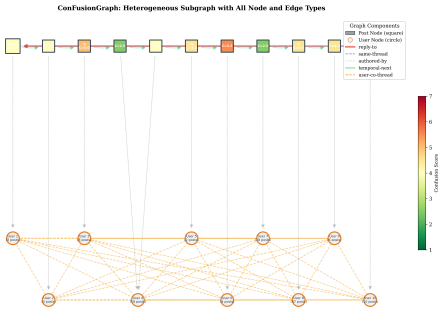


Fig. 2. Example subgraph from the heterogeneous forum graph

C. Stacked Ensemble Architecture (ConFusionGraph V2)

Rather than forcing a single model to handle text, metadata, and structure simultaneously—which we empirically found to be suboptimal—we employ a stacked ensemble where each expert handles its strength.

ConFusionGraph V1 denotes the single heterogeneous GNN with relation-typed attention used as a direct predecessor; V2 replaces this single-model design with the stacked ensemble described below.

Expert 1 (Metadata): XGBoost trained on TF-IDF features (10,000 dimensions, unigrams and bigrams) concatenated with 6 metadata features. This expert handles tabular feature interactions natively through tree splits.

Expert 2 (Text): A two-layer MLP trained on frozen BERT [CLS] embeddings (768-dimensional). This expert captures deep semantic patterns in post text. Frozen BERT embeddings are used deliberately to preserve ensemble diversity, as fine-tuning would produce predictions closely correlated with the BERT fine-tuned baseline (B4), reducing Expert 2 to a near-duplicate; its primary contribution is an independent semantic estimate for the meta-learner on ambiguous posts. *Expert 3 (Structure):* A heterogeneous GNN with 3 layers of GATv2 convolution for same-type edges and SAGEConv for cross-type edges, operating on the full heterogeneous graph. This expert captures forum structural signals.

Out-of-fold predictions. To prevent leakage in the meta-learner, Experts 1 and 2 generate training-set predictions via 5-fold stratified cross-validation. Expert 3 produces calibrated predictions across all splits.

Graph structural features. For each post, we compute 8 additional features from the graph: reply degree, thread degree, mean expert probabilities of thread neighbors (from each expert), neighbor disagreement (standard deviation of expert predictions), has-answer-neighbor flag, and isolation score. These features use expert predictions rather than ground-truth labels, ensuring no target leakage.

Meta-learner. An XGBoost classifier (200 trees, depth 4, L1 regularization) trained on 17 features: 3 expert probabilities, 6 metadata features and 8 graph structural features. All models use weighted binary cross-entropy loss with class weight equal to the negative-to-positive ratio.

D. Explainability Analysis

Global explainability (KernelSHAP). We train an interpretable XGBoost model on 112 features (6 metadata, 6 graph structural, 100 top TF-IDF tokens selected by mutual information) and compute KernelSHAP values for all test samples using 200 background training samples.

Token-level analysis. SHAP values for the 100 TF-IDF token features are extracted separately and ranked by mean absolute SHAP value. Tokens are classified as confusion-increasing (positive mean SHAP) or confusion-decreasing (negative mean SHAP).

Local explainability (LIME). Per-post explanations are generated for representative cases (true positive, false negative,

false positive, borderline) and aggregated across all 561 confused posts to identify the most frequent locally important features.

Confusion archetype discovery. Confused test posts are clustered by their SHAP value profiles over metadata and graph features using KMeans with $K = 4$. Each cluster is characterized by its dominant SHAP patterns and assigned an interpretable label based on which features drive the prediction. Silhouette analysis across $K = 2$ to $K = 6$ (Table II) confirms $K = 2$ as the optimal separation (score = 0.320), validating the two-archetype taxonomy; $K = 4$ was retained to expose sub-structure within Active Questioning, whose three sub-clusters share the same primary driver and intervention strategy.

TABLE II
SILHOUETTE SCORES FOR $K = 2$ TO $K = 6$

K	Silhouette Score	Inertia
2	*0.320*	6998.5
3	0.266	5862.9
4	0.241	5237.9
5	0.258	4713.5
6	0.265	4267.1

E. Confusion Chain Extraction

Thread-level confusion chains are extracted as temporal sequences of confusion scores within each thread (root, reply 1, reply 2, and so on). Each chain is classified by linear regression slope and volatility into five patterns: Escalating, Resolving, Answer-Resolved, Oscillating, and Stable. Resolution events are detected where an answer-flagged post is followed by a confusion drop of 0.3 or more. User-level trajectories track individual confusion evolution over time. Cross-thread contagion is measured by correlating each user's confusion level in thread N with their level in thread N plus 1.

F. Multi-strategy Intervention System

The intervention system routes each confused post to the most appropriate strategy based on its archetype classification.

Strategy 1 (Peer-Answer Retrieval). For Active Question archetypes, existing answer-flagged posts are pre-encoded using sentence-transformers (all-MiniLM-L6-v2) and the top-3 most semantically similar answers from the same course are retrieved.

Strategy 2 (Peer Connector). For Silent Struggler and Isolated Learner archetypes, user activity profiles are constructed and the most active, low-confusion peers in the same course are identified.

Strategy 3 (Thread Summarizer). For Confusion Contagion archetypes, the thread structure is extracted: root question, number of questions versus answers, mean thread confusion, and unresolved status.

Strategy 4 (Resource Recommender). For all archetypes, course-specific module recommendations and archetype-specific action items are generated.

Strategy 5 (XAI-Conditioned Clarification). As a fallback, template-based responses conditioned on the archetype, incorporating retrieved peer answers and thread context, are generated.

IV. EXPERIMENTAL RESULTS

This section evaluates ConFusionGraph across four dimensions: detection performance against internal baselines and published methods, expert contribution via agreement analysis and ablation, explainability quality via dual SHAP-LIME analysis, and intervention effectiveness via archetype-based strategy routing. All experiments follow the temporal split described in Section III-A.

A. Experimental Setup

All experiments are conducted on a machine equipped with an NVIDIA RTX 3060 GPU (12 GB VRAM), 16 GB system RAM, and an Intel processor, running Python 3.10 with PyTorch and PyTorch Geometric. The Stanford MOOCPost dataset is split temporally per course into 70% training, 15% validation, and 15% test sets. Urgency and Sentiment features are excluded from all experiments to prevent annotation leakage, as detailed in Section III-A.

Table III summarizes the hyperparameters for each model component. Expert 1 (XGBoost) uses 200 estimators with a maximum depth of 6, a learning rate of 0.1, and scale_pos_weight set to the negative-to-positive class ratio (5.33) to handle class imbalance. TF-IDF features are extracted with a vocabulary of 10,000 terms using unigrams and bigrams. Expert 2 (BERT+MLP) uses frozen bert-base-uncased embeddings (768 dimensions, maximum sequence length of 256) fed into a two-layer MLP with hidden sizes of 256 and 128, trained with AdamW (learning rate 1e-3, weight decay 1e-4) and early stopping with patience of 15 epochs. Expert 3 (Heterogeneous GNN) consists of 3 GATv2 layers with 4 attention heads, a hidden dimension of 128, dropout of 0.3, and residual connections, trained with AdamW (learning rate 1e-3) and cosine annealing scheduler for 200 epochs. Out-of-fold predictions for Experts 1 and 2 are generated through 5-fold stratified cross-validation on the training set. The meta-learner is an XGBoost classifier with 200 estimators, maximum depth of 4, L1 regularization (alpha = 0.1), and is trained on 17 features: 3 expert probabilities, 6 metadata features, and 8 graph structural features. All models use weighted binary cross-entropy loss. Results are reported as the mean over 5 random seeds. For the explainability analysis, KernelSHAP uses 200 background training samples with nsamples set to 100, computed on all test posts. LIME generates explanations with 2,000 perturbation samples per post. The interpretable XGBoost model used for SHAP analysis is trained on 112 features (6 metadata, 6 graph structural, and 100 top TF-IDF tokens selected by mutual information). Sentence embeddings for the intervention system are generated using all-MiniLM-L6-v2.

TABLE III
HYPERPARAMETER CONFIGURATION FOR EACH MODEL COMPONENT

Component	Parameter	Value
Expert 1 (XGBoost)	Estimators / Depth / Learning Rate	200 / 6 / 0.1
	TF-IDF Vocabulary / N-gram range	10,000 / (1,2)
	Scale Positive Weight	5.33
Expert 2 (BERT+MLP)	Encoder / Embedding dim / Max Length	bert-base-uncased / 768 / 256
	MLP hidden layers / Optimizer	(256, 128) / AdamW ($lr = 1e-3$)
	Early stopping patience	15 epochs
Expert 3 (GNN)	Layers / Heads / Hidden dim / Dropout	3 / 4 / 128 / 0.3
	Optimizer / Scheduler / Epochs	AdamW ($lr = 1e-3$) / Cosine / 200
Meta-learner	Estimators / Depth / L1 alpha	200 / 4 / 0.1
	Input features	17 (3 probs + 6 meta + 8 graph)
KernelSHAP	Background samples / n_samples	200 / 100
LIME	Perturbation samples / Top features	2,000 / 10

B. Detection Performance

Fig. 3 compares the accuracy and AUROC of all models evaluated in this study. The ConFusionGraph V2 stacked ensemble achieves the highest accuracy at 89.2%, outperforming all baselines. Among non-graph models, TF-IDF combined with metadata and XGBoost (89.0%) establishes a strong baseline, confirming that the Question flag and other metadata features carry substantial predictive power. BERT+MLP achieves 87.9% and BERT fine-tuned reaches 88.3%, while BERT+GCN (82.7%) and BERT+GAT (83.1%) perform the worst, indicating that naive application of graph convolution on homogeneous graphs can degrade performance when structural signal is weak. BERT+HGT (86.2%) partially recovers through heterogeneous message passing, and the V1 ConFusionGraph model reaches 87.7% using relation-typed attention on the full heterogeneous graph. The V2 ensemble achieves the best accuracy by combining the complementary strengths of metadata interactions (Expert 1), deep text semantics (Expert 2), and forum structural context (Expert 3) through a meta-learner enriched with graph neighborhood features. In terms of AUROC, BERT fine-tuned achieves the highest value at 0.929, followed by V1 at 0.926 and V2 at 0.923, suggesting that deeper text models and graph models produce well-calibrated probability estimates. Notably, V2 achieves the highest accuracy despite not having the highest AUROC, indicating that the meta-learner optimizes the decision boundary more effectively than individual experts.

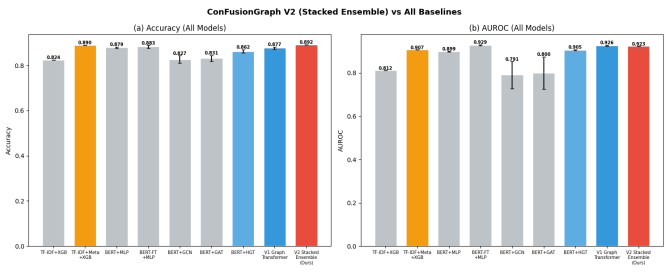


Fig. 3. Confusion Detection performance across all models. Panel (a) shows accuracy and panel (b) shows AUROC

Table IV provides a contextual reference against published methods. Direct numerical comparison is not feasible due to differing thresholds, evaluation protocols, and in one case corpus language. ConFusionGraph is the only method to jointly address detection, explainability, and intervention.

TABLE IV
CONTEXTUAL COMPARISON WITH PUBLISHED METHODS ON MOOC FORUM CONFUSION DETECTION

Method	Dataset	Acc	Interpretable	Actionable
ConvL [8]†	Stanford MOOC-Post	81.45%	No	No
Du & Xing [7]‡	Stanford MOOC-Post	84.58%	Yes	No
Jhajj et al. [29]§	Stanford MOOC-Post	85.08%	No	No
RE-MOOC [10]⋆	Chinese MOOC	95.00%	No	No
ConFusionGraph V2 (ours)	Stanford MOOC-Post	89.2%	Yes	Yes

†Cross-domain transfer protocol; averaged over 6 subtasks.
‡Three-class formulation; random 80/20 split.
§Score=4 posts removed; 5-fold cross-validation split.
⋆Different language and dataset; not directly comparable.

C. Expert Agreement and Ablation Analysis

To understand how the three experts complement each other within the stacked ensemble, we analyze their agreement patterns and measure the impact of removing each component. Fig. 4(a) shows that all three experts agree on 3,789 out of 4,443 test posts (85.3%), achieving 94.4% accuracy on these consensus cases. For the 654 posts where experts disagree, the meta-learner correctly resolves 388 cases (59.3%). Fig. 4(b) presents the ablation results as accuracy deltas from the full ensemble. Removing the GNN expert causes the largest drop (-0.010), confirming that structural features provide information not captured by text or metadata alone. Removing the BERT expert slightly improves performance (+0.004), suggesting that its predictions partially overlap with the XGBoost expert. Removing graph structural features reduces performance by -0.001, and using expert probabilities only ("Experts only")

reduces by -0.002. The largest degradation occurs when reducing the system to XGBoost alone (-0.029), confirming that the ensemble architecture provides meaningful improvements over any single-model approach.

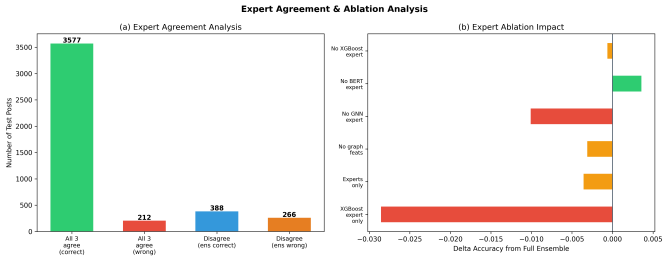


Fig. 4. Expert agreement and ablation analysis

D. Global Feature Importance (SHAP)

The following analysis uses the surrogate XGBoost model described in Section III rather than the full ensemble, as KernelSHAP requires a tractable model for efficient approximation. To understand what drives confusion predictions, we apply KernelSHAP on all 4,443 test posts using an interpretable XGBoost model trained on 112 features (6 metadata, 6 graph structural, and 100 top TF-IDF tokens). Fig. 5 presents the SHAP beeswarm plot, where each dot represents one test post, horizontal position indicates the SHAP value (positive pushes toward confused), and color encodes the feature value (red = high, blue = low).

The Question flag dominates with a mean |SHAP| of 0.114, approximately 4.3 times the second-ranked feature. When the Question flag is active (red dots), SHAP values reach +0.3 to +0.5, strongly pushing the prediction toward confused; when inactive (blue dots), SHAP values cluster around -0.1, pushing away from confusion. The Answer flag (rank 2, mean |SHAP| = 0.027) shows the inverse pattern: high feature values (red) produce negative SHAP, meaning answer-flagged posts are predicted as not confused. Two graph structural features appear in the top 5: ThreadMeanConf (rank 4, mean |SHAP| = 0.012) exhibits a clear positive trend where higher thread-level confusion (red) pushes predictions toward confused, and ReplyDegree (rank 5, mean |SHAP| = 0.009) shows that more connected posts receive higher confusion predictions. These graph features outperform most text tokens, validating the heterogeneous graph design. Among tokens, ”students” (rank 3, mean |SHAP| = 0.015) stands out for its strong confusion-decreasing effect, while ”question” (rank 9) and ”answer” (rank 23) are among the few tokens that push toward confusion when present.

E. Token-level SHAP Analysis

Beyond aggregate feature importance, we extract the SHAP values for 100 TF-IDF token features separately to identify which words push predictions toward or away from confusion. Fig. 6 presents the token-level analysis in three panels. Only three tokens exhibit confusion-increasing behavior: ”question” (+0.0033 mean SHAP), ”answer” (+0.0016), and ”questions”

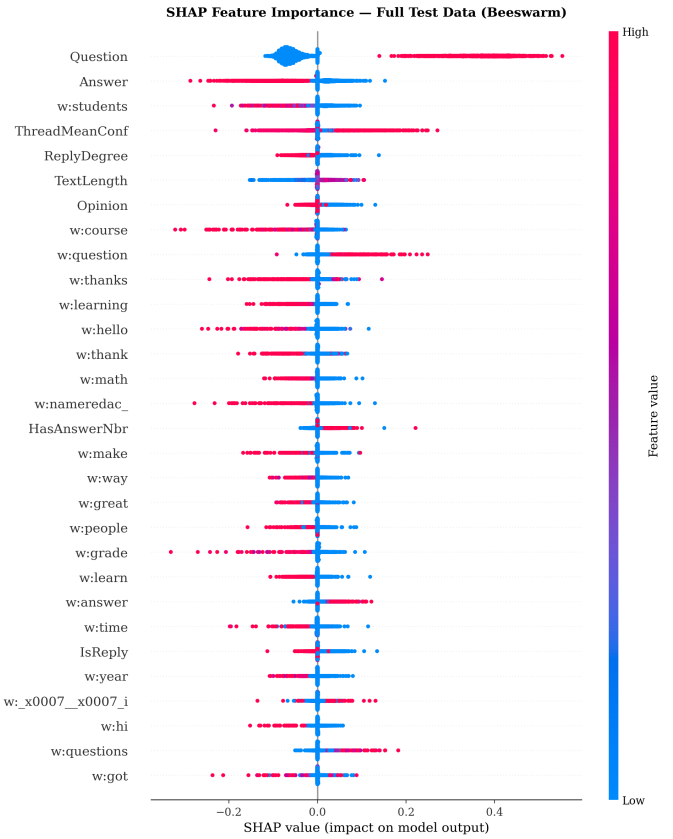


Fig. 5. SHAP bee-swarm plot for all 4443 test posts

(+0.0011). These constitute help-seeking vocabulary that signals active confusion. In contrast, ten tokens show strong confusion-decreasing effects, led by ”students” (-0.0042), ”course” (-0.0034), ”thanks” (-0.0016), and ”learning” (-0.0020). These represent engagement and appreciation vocabulary — posts expressing gratitude (”thanks”, ”thank”, ”great”) or discussing course content in general terms (”students”, ”course”, ”learning”, ”math”) are reliably non-confused.

This asymmetry between three confusion-increasing and ten confusion-decreasing tokens reveals an important linguistic pattern: the vocabulary of confusion is narrow and specific (centered on question-asking), while the vocabulary of non-confusion is broad and diverse (spanning gratitude, discussion, and domain-specific terms). This finding has practical implications for MOOC platform designers: monitoring the ratio of help-seeking to engagement vocabulary in forum activity could serve as an early warning indicator of rising confusion levels.

F. Local explainability (LIME) and SHAP-LIME agreement

To complement the global SHAP analysis with local explanations, we apply LIME to all 561 confused test posts, generating per-post explanations with the top 10 contributing features. Fig. 7(a) shows that the Question flag and Answer flag appear in 100% of LIME top-10 explanations (561 out of 561), followed by the token ”students” (556/561, 99.1%), ”hello” (521/561, 92.9%), and ”question” (486/561, 86.6%). Fig. 7(b) presents mean LIME contribution weights: Question carries

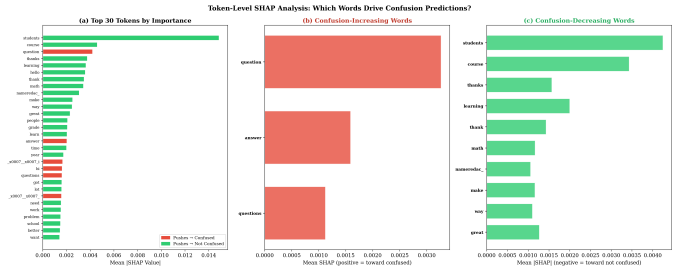


Fig. 6. Token-level SHAP analysis. Panel (a) ranks top 30 tokens by importance, colored by direction. Panel (b) shows confusion-increasing words. Panel (c) shows confusion-decreasing words

the largest positive weight (+0.249), followed by Answer (+0.118), ”hello” (+0.109), ThreadMeanConf (+0.102), and ”homework” (+0.100). On the negative side, ”set” (-0.117), ”says” (-0.091), and ”questions” (-0.091) push predictions away from confusion in the LIME framework. The agreement between SHAP and LIME strengthens the reliability of these explanations. Both methods rank Question as the dominant feature and Answer as the second. The token ”students” ranks 3rd in both SHAP (by mean |SHAP|) and LIME (by frequency). Graph features ThreadMeanConf and ReplyDegree appear in the top ranks of both methods, confirming that structural forum context contributes genuine predictive signal validated through two independent explanation approaches.

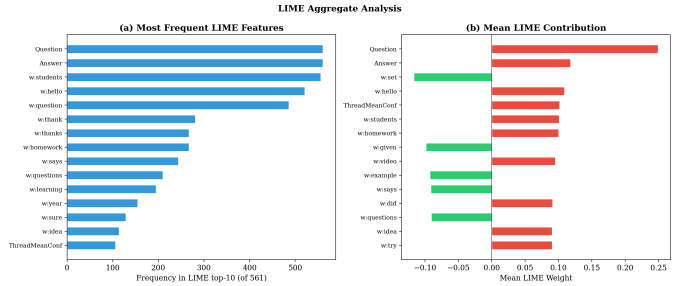


Fig. 7. LIME aggregate analysis across 561 confused test posts.

G. Confusion Archetypes

Clustering the 561 confused test posts by their SHAP value profiles over metadata and graph features using KMeans (K=4) reveals two distinct confusion mechanisms. Table V presents the discovered archetypes

The three Active Question clusters (414 posts, 73.8%) share a strongly positive Question SHAP value but differ in secondary drivers: cluster A is characterized by longer texts in high-confusion threads, cluster B by reply posts with elevated IsReply SHAP, and cluster C represents the dominant pattern of question posts with moderate thread confusion. The Confusion Contagion cluster (147 posts, 26.2%) is qualitatively distinct: its Question SHAP is near zero (-0.002), meaning these posts are not identified through question-asking behavior. Instead, ThreadMeanConf (+0.014) is the dominant driver, indicating that confusion is detected through the structural context of the surrounding thread. This archetype would be entirely invisible to text-only classification methods, providing

TABLE V
CONFUSION ARCHETYPES DISCOVERED VIA SHAP

Archetype	Count (%)	Dominant SHAP Features
Active Question (A)	44 (7.8%)	Question=+0.281, TextLength=+0.046, ThreadMeanConf=+0.039
Active Question (B)	43 (7.7%)	Question=+0.386, IsReply=+0.050, Answer=+0.019
Active Question (C)	327 (58.3%)	Question=+0.381, ThreadMeanConf=+0.026, Answer=+0.025
Confusion Contagion	147 (26.2%)	ThreadMeanConf=+0.014, Answer=−0.014, Question=−0.002

direct evidence for the value of the graph-based approach. The negative Answer SHAP (-0.014) in this cluster suggests that the absence of answer posts in the thread neighborhood amplifies the confusion signal.

H. Confusion Chain Analysis

To investigate how confusion evolves within and across discussions, we extract temporal confusion chains from 3,712 threads containing two or more posts. Each chain represents the chronological sequence of confusion scores within a thread, classified by linear regression slope into five patterns. Table VI presents the distribution.

TABLE VI
CONFUSION CHAIN PATTERN DISTRIBUTION (3,712 THREADS)

Pattern	Count (%)	Description
Stable	3,183 (85.7%)	No significant trend in confusion over the thread
Resolving	436 (11.7%)	Confusion decreases as the discussion progresses
Answer-Resolved	40 (1.1%)	Confusion drops specifically after an answer post
Escalating	34 (0.9%)	Confusion increases through the discussion
Oscillating	19 (0.5%)	High variance with no directional trend

While the majority of threads maintain stable confusion levels, 436 threads (11.7%) exhibit resolving patterns where confusion decreases over time, and 40 threads show explicit answer-based resolution where an answer-flagged post triggers a measurable confusion drop. Resolution event analysis identifies 235 such events across 216 threads, with a mean confusion reduction of -0.376 points (on the 1-7 scale) following an answer post. This confirms that timely, relevant answers in forum threads have a quantifiable positive impact on learner confusion.

Cross-thread contagion analysis reveals a statistically significant positive correlation between a user's confusion level in thread N and their confusion in thread N+1 ($r = 0.286$, $p < 10^{-121}$). This indicates a statistically significant association between a user's confusion levels across consecutive

threads, consistent with persistent confusion but not necessarily direct propagation. User trajectory analysis across 1,178 active users (with five or more posts) shows that 52 users (4.4%) follow an increasingly confused trajectory, 43 (3.7%) show a learning pattern with decreasing confusion, and the remaining 1,083 (91.9%) maintain stable confusion levels over time. These findings provide empirical support for the graph-based modeling approach: confusion is not an isolated per-post phenomenon but propagates through thread structure and user behavior, which the heterogeneous graph captures through its edge types.

I. Multi-strategy Intervention Evaluation

The intervention system generates personalized responses for all 561 confused test posts, routing each to the most appropriate strategy based on its SHAP-derived archetype. The framework defines five strategies: Peer-Answer Retrieval (S1), Peer Connector (S2), Thread Summarizer (S3), Resource Recommender (S4), and XAI-Conditioned Clarification (S5). Three are activated as primary strategies through archetype-based routing, one operates as a universal supplement, and one remains inactive for the current dataset. Resource Recommender runs alongside every primary intervention, providing course-specific and archetype-specific action items. Peer Connector is designed for isolation-driven archetypes (Silent Struggler, Isolated Learner), which did not emerge as distinct clusters in the SHAP-based discovery — the clustering produced Active Question and Confusion Contagion as the two primary archetypes, suggesting that isolation-driven confusion may be underrepresented in this dataset or requires finer-grained clustering. Table VII presents the primary strategy evaluation.

TABLE VII
MULTI-STRATEGY INTERVENTION EVALUATION (561 CONFUSED POSTS)

Strategy	Count (%)	Relevance	Specificity
Peer-Answer Retrieval	402 (71.7%)	0.472	0.764
Thread Summarizer	38 (6.8%)	0.322	0.852
XAI Clarification	121 (21.6%)	0.196	0.636
Overall	561 (100%)	0.402	0.742

Peer-Answer Retrieval serves 402 posts (71.7%), achieving the highest relevance (0.472) by retrieving semantically similar existing answers from the same course. Thread Summarizer handles 38 posts (6.8%) from the Confusion Contagion archetype, achieving the highest specificity (0.852) because its responses are grounded in actual thread structure. XAI Clarification serves as fallback for 121 posts (21.6%). The archetype-to-strategy routing confirms genuine differentiation: all Peer-Answer Retrieval interventions originate from Active Question posts, while all Thread Summarizer interventions originate from Confusion Contagion posts, demonstrating that XAI-derived archetypes successfully guide qualitatively different interventions. The overall specificity of 0.742 indicates interventions are substantially different from generic responses. The saved LLM-ready prompts offer a pathway to improve response quality through large language model generation in future deployment.

V. DISCUSSION

This section interprets the experimental findings of ConFusionGraph by revisiting the three research questions through the lens of our observed data and model performance.

A. RQ1: Value of graph structure for confusion detection

Our results provide a definitive answer to whether modeling MOOC forums as heterogeneous graphs improves confusion detection. Graph structural features—specifically *ThreadMeanConf* and *ReplyDegree*—rank 4th and 5th in global SHAP importance (mean $|\text{SHAP}| = 0.012$ and 0.009 respectively). This outperforms most individual text tokens, confirming that the surrounding forum structure carries a genuine predictive signal that text alone cannot capture.

The ablation study further supports this: removing the GNN expert from the ensemble produced the largest accuracy drop among all expert components. Furthermore, the discovery of the *Confusion Contagion* archetype (26.2% of confused posts) proves that a significant portion of learner confusion is driven by the state of the thread rather than the specific language of a post.

The cross-thread contagion analysis ($r = 0.286, p < 10^{-121}$) provides the final justification for this graph-based approach. We observed that confused users frequently carry their state across different discussions. By utilizing *user-co-thread* and *temporal-next* edge types, ConFusionGraph successfully captures this propagation—a phenomenon invisible to traditional models that treat forum posts as independent units.

B. RQ2: Key features, linguistic patterns, and archetypes

As noted in Section III, the following feature importance rankings and archetype discoveries reflect the surrogate XGBoost model rather than the full ensemble. The dual SHAP-LIME analysis reveals a clear hierarchy of confusion drivers. The *Question* flag is the dominant indicator (0.114 mean $|\text{SHAP}|$), appearing in 100% of top-10 LIME explanations with a mean weight of $+0.249$. This suggests that a rule-based baseline focused on help-seeking indicators would capture a substantial share of confusion signals, though it would fail on structurally-driven cases such as the Confusion Contagion archetype.

Our token-level analysis uncovered an *asymmetric confusion vocabulary*. We found that only three tokens strongly push toward a "confused" classification (*"question"*, *"answer"*, *"questions"*), while over ten tokens consistently push away from it (*"students"*, *"course"*, *"thanks"*, *"learning"*, *"great"*). This suggests that confusion in MOOCs is characterized more by the absence of engagement vocabulary than by the presence of explicit confusion-related words.

By clustering these drivers, we identified two distinct mechanisms:

1) *Active Question* (73.8%): Explicit help-seeking triggered by post content.

2) *Confusion Contagion* (26.2%): Structurally-triggered confusion where the thread context is the primary predictor.

Finally, we observed that excluding co-annotated features from the *Stanford MOOCPost* dataset, such as *Urgency* (which has a high correlation of $r = 0.48$ with confusion), was essential for a realistic deployment. While including such features inflates accuracy, ConFusionGraph remains effective using only observable text and metadata, making it viable for real-time MOOC monitoring.

C. RQ3: XAI-driven multi-strategy intervention

The intervention system demonstrates that archetypes derived from explainable AI can effectively automate the selection of support strategies. The routing mechanism successfully aligned intervention types with student needs: *Active Question* archetypes were primarily routed to *Peer-Answer Retrieval* (71.7%), while *Confusion Contagion* archetypes were addressed via the *Thread Summarizer* (6.8%).

The high relevance score of the *Peer-Answer Retrieval* strategy (0.472) indicates that for the majority of confused learners, a solution already exists within the forum; the primary barrier is the learner's inability to locate it. Meanwhile, the *Thread Summarizer* achieved the highest specificity (0.852). By grounding responses in the actual thread structure, this strategy directly addresses the "contagion" effect by providing a structured overview of long, complex discussions.

Lastly, the confusion chain analysis validates the entire intervention design. We found that surfacing existing answer posts reduces downstream confusion by an average of 0.376 points. This confirms that by using ConFusionGraph to interrupt the propagation chain early, we can prevent localized confusion from escalating into a thread-wide issue.

D. Limitations

Several limitations should be acknowledged. First, the confusion annotations are based on human ratings on a 1-7 Likert scale, introducing subjectivity; the binary threshold of 4.5 imposes a hard boundary on what is inherently a continuous spectrum. Second, the intervention evaluation relies on automated metrics (cosine similarity for relevance, distance from generic response for specificity) rather than actual learning outcomes from a live deployment study. These automated metrics serve as proxy measures for intervention quality; while they provide a reproducible baseline for comparison, a human-in-the-loop evaluation or live deployment study on an active MOOC platform is necessary to confirm whether the proposed strategies genuinely reduce learner confusion in practice. Third, the SHAP-based archetype discovery produced only two primary archetypes (Active Questioning and Confusion Contagion), with the Peer Connector strategy remaining inactive due to the absence of an isolation-driven cluster. The absence of isolation-driven archetypes likely reflects the dataset's structural bias toward active posters, and the limited signal of the SHAP feature space to distinguish isolation-driven confusion from the dominant Active Questioning pattern. Fourth, the cross-thread contagion analysis establishes correlation rather than causation — confused users may actively seek out difficult threads rather than being confused by them. Finally, the template-based clarification responses could be substantially improved through large language model generation, though the saved prompts provide a direct upgrade path.

VI. Conclusion

This paper presented ConFusionGraph, an end-to-end framework for confusion detection, explanation, and intervention in MOOC discussion forums. By modeling forums as heterogeneous graphs with five relation types and employing a stacked ensemble of three specialized experts (XGBoost for metadata, BERT+MLP for text, and a heterogeneous GNN for structure), the framework achieves 89.2% accuracy on the Stanford MOOCPost dataset while providing comprehensive explainability through dual SHAP-LIME analysis, establishing competitive performance among methods evaluated on this corpus.

The key findings are as follows. First, graph structural features provide genuine predictive signal, with *ThreadMeanConf* and *ReplyDegree* ranking in the global top 5; their primary value lies in enabling structural explainability and the discovery of the Confusion Contagion archetype, invisible to text-only methods. Second, the Question flag dominates confusion predictions (mean $|\mathrm{SHAP}| = 0.114$), while token-level analysis reveals an asymmetric confusion vocabulary where help-seeking words push toward confusion and engagement words push away. Third, SHAP-based clustering identifies two primary mechanisms — *Active Question* (73.8%) driven by explicit help-seeking and *Confusion Contagion* (26.2%) driven by thread-level structural context — which successfully route interventions to differentiated strategies. Fourth, confusion propagates across discussions with a statistically significant contagion correlation ($r = 0.286$, $p < 10^{-121}$). Fifth, the multi-strategy intervention system achieves 0.742 overall specificity, with Peer-Answer Retrieval serving 71.7% of confused posts by surfacing relevant existing answers.

Future work will pursue three directions. First, a live deployment study on an active MOOC platform measuring actual learning outcome improvements from the intervention system, including A/B testing of differentiated strategies. Second, integration of large language models for higher-quality intervention generation using the saved XAI-conditioned prompts, combining retrieval-augmented generation with the peer-answer database. Third, extension to multilingual MOOC settings and larger-scale datasets such as OULAD, KDD Cup 2015, and MOOCCube to validate the generalizability of the discovered archetypes across different educational contexts and platform designs.

References

[1] S. Rizwan, C. K. Nee, and S. Garfan, "Identifying the factors affecting student academic performance and engagement prediction in MOOC using deep learning: A systematic literature review," *IEEE Access*, vol. 13, pp. 18952–18982, 2025, doi: 10.1109/ACCESS.2025.3533915.

[2] S. D'Mello, B. Lehman, R. Pekrun, and A. Graesser, "Confusion can be beneficial for learning," *Learn. Instr.*, vol. 29, pp. 153–170, 2014. doi: 10.1016/j.learninstruc.2012.05.003.

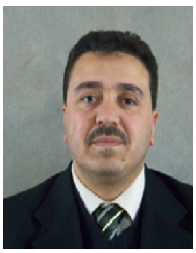
[3] M. Y. Doo and J. Kim, "The relationship between learning engagement and learning outcomes in online learning in higher education: A meta-analysis study," *Distance Educ.*, vol. 45, pp. 60–82, 2024. doi: 10.1080/01587919.2024.2303484.
[4] M. Wen, D. Yang, and C. P. Rosé, "Sentiment analysis in MOOC discussion forums: What does it tell us?" in *Proc. 7th Int. Conf. Educ. Data Min.*, 2014, pp. 130–137.
[5] L. Feng, L. Wang, S. Liu, and G. Liu, "Classification of discussion threads in MOOC forums based on deep learning," in *2017 2nd Int. Conf. Wireless Commun. Netw. Eng. (WCNE)*, 2017.
[6] Z. Zhang, X. Zhu, Q. He, and L. Zhang, "BERT-based global semantic refinement and local semantic extraction for distinguishing urgent posts in MOOC forums," *IEEE Access*, vol. 12, pp. 116250–116258, 2024. doi: 10.1109/ACCESS.2024.3426976.
[7] H. Du and W. Xing, "Leveraging explainability for discussion forum classification: Using confusion detection as an example," *Distance Educ.*, vol. 44, pp. 190–205, 2023. doi: 10.1080/01587919.2022.2150145.
[8] X. Wei, H. Lin, L. Yang, and Y. Yu, "A convolution-LSTM-based deep neural network for cross-domain MOOC forum post classification," *Information*, vol. 8, no. 92, 2017. doi: 10.3390/info8030092.
[9] S. A. Geller et al., "New methods for confusion detection in course forums: Student, teacher, and machine," *IEEE Trans. Learn. Technol.*, vol. 14, pp. 665–679, 2021. doi: 10.1109/TLT.2021.3123266.
[10] Y. Wu, T. Liu, S. Cao, and X. Yang, "Towards automatic detection of confusion in MOOC learner feedback: A natural language processing approach," in *2024 4th Int. Conf. Electronic Info. Eng. and Comput. Tech. (EIECT)*, 2024. doi: 10.1109/EIECT64462.2024.10866730.
[11] M. Yee et al., "AI-assisted analysis of content, structure, and sentiment in MOOC discussion forums," *Front. Educ.*, 2023. doi: 10.3389/feduc.2023.1250846.
[12] Z. Liu, W. Xing, X. Jiao, and C. Li, "Exploring fairness and explainability in LLM-generated support for online learning discussion forums," *J. Learn. Anal.*, vol. 12, no. 3, pp. 8–33, 2025. doi: 10.18608/jla.2025.8885.
[13] M. Li, X. Wang, Y. Wang, Y. Chen, and Y. Chen, "Study-GNN: A novel pipeline for student performance prediction based on multi-topology graph neural networks," *Sustainability*, 2022. doi: 10.3390/su14137965.
[14] M. Sun, M. Hu, G. Wang, X. Xu, and J. Li, "Dynamic dropout prediction in MOOCs using temporal graph networks," in *Proc. 16th Int. Conf. Educational Data Mining*, 2023.
[15] B. Zhang, Q. He, and D. Zhang, "Heterogeneous graph neural network for short text classification," *Appl. Sci.*, 2022. doi: 10.3390/app12178711.
[16] H. Li, Y. Yan, S. Wang, J. Liu, and Y. Cui, "Text classification on heterogeneous information network via enhanced GCN and knowledge," *Neural Comput. Appl.*, 2023. doi: 10.1007/s00521-023-08494-0.
[17] S. Hakkal and A. Ait Lahcen, "Leveraging graph neural network for learner performance prediction," *Expert Syst. Appl.*, vol. 293, p. 128724, 2025. doi: 10.1016/j.eswa.2025.128724.
[18] Z. Hancox and S. D. Relton, "Temporal graph-based CNNs (TG-CNNs) for online course dropout prediction," in *Int. Symp. Methodologies for Intelligent Systems*, Springer, 2022, pp. 357–367. doi: 10.1007/978-3-031-16564-1_34.
[19] D. Roh, D. Han, D. Kim, K. Han, and M. Y. Yi, "SIG-Net: GNN-based dropout prediction in MOOCs using student interaction graph," in *Proc. 39th ACM/SIGAPP Symp. Applied Computing*, 2024, pp. 29–37. doi: 10.1145/3605098.3636003.
[20] H. Khosravi et al., "Explainable Artificial Intelligence in education," *Computers and Education: Artificial Intelligence*, 2022. doi: 10.1016/j.caeai.2022.100074.
[21] M. Nagy and R. Molontay, "Interpretable dropout prediction: Towards XAI-based personalized intervention," *Int. J. Artif. Intell. Educ.*, 2024. doi: 10.1007/s40593-023-00331-8.
[22] O. B. Deho et al., "Should learning analytics models include sensitive attributes? Explaining the why," *IEEE Trans. Learn. Technol.*, 2023. doi: 10.1109/TLT.2022.3226474.
[23] G. Türkmen, "The review of studies on explainable artificial intelligence in educational research," *J. Educ. Comput. Res.*, 2025. doi: 10.1177/07356331241310 91.
[24] D. Bañeres, M. E. Rodríguez-González, A.-E. Guerrero-Roldán, and P. Cortadas, "An early warning system to identify and intervene online dropout learners," *Int. J. Educ. Technol. High. Educ.*, 2023. doi: 10.1186/s41239-022-00371-5.
[25] R. Cobos, "Self-regulated learning and active feedback of MOOC learners supported by the intervention strategy of a learning analytics system," *Electronics*, 2023. doi: 10.3390/electronics12153368.
[26] J. Yu et al., "From MOOC to MAIC: Reshaping online teaching and learning through LLM-driven agents," *arXiv preprint arXiv:2409.03512*, 2024.
[27] K. Tan, J. Yao, T. Pang, C. Fan, and Y. Song, "ELF: Educational LLM framework for improving and evaluating AI-generated content for classroom teaching," *ACM J. Data and Information Quality*, 2025. doi: 10.1145/3712298.
[28] M. Rebelo Marcolino et al., "Student dropout prediction through machine learning optimization: Insights from Moodle log data," *Scientific Reports*, vol. 15, no. 1, pp. 1–16, 2025. doi: 10.1038/s41598-025-93918-1.
[29] G. Jhajj, M. A. A. Dewan, and F. Lin, "Unveiling uncertainty: Supporting learners through NLP-driven confusion identification," in *Proc. IEEE DASC/PiCom/ CBDCom/CyberSciTech*, 2023, pp. 137–142. doi: 10.1109/DASC/PICOM/CBDCOM/ CY59711.2023.10361304.

Abdennour Redjaibia is a Ph.D. candidate in Artificial Intelligence at Mohamed Chérif Messaadia University, Souk Ahras, Algeria, and a member of the Computer Science and Mathematics Laboratory (LIM). His research focuses on the application of artificial intelligence (AI) techniques in Massive Open Online Courses (MOOCs) to reduce learner dropout rates by providing personalized assistance.

Samia Drissi received her Ph.D. degree in computer science from the University of Annaba, Algeria, in 2015. She is currently a Professor with the Department of Computer Science at Mohamed Chérif Messaadia University, Souk Ahras, Algeria, and is a member of the Computer Science and Mathematics Laboratory (LIM). She serves on the editorial boards of several international journals and conferences. Her research interests include e-learning, adaptive hypermedia systems, the Semantic Web, and artificial intelligence.

Karima Boussaha received her Ph.D. degree in computer science from the University of Annaba, Algeria, in 2016, and her Habilitation Universitaire in 2021. She is currently an Associate Professor with the Department of Mathematics and Computer Science at the University of Oum El Bouaghi, Algeria, and a member of the Artificial Intelligence and Autonomous Things Laboratory (LIAOA). She serves on the editorial boards of numerous international journals and conferences. Her research interests include computer-assisted learning, collaborative assessment, MOOC platforms, and artificial intelligence.

Sevinçç Gülseçen is a Full Professor with the Department of Computer Science at Istanbul University, Turkey, where she also serves as the Head of the Informatics Program. She is the founder of the Human-Computer Interaction Research Laboratory at Istanbul University and the Editor-in-Chief of the ActaINFOLOGICA journal. With over 100 publications, she also co-founded the International Future-Learning Conference and serves on numerous international program committees. Her research interests focus on human-computer interaction, human-centered artificial intelligence, management information systems, and knowledge management.



Yacine Lafifi is currently working as a Full Professor at the Computer Science Department of Guelma University, Algeria. He works in e-learning research field since 1997. He received his PhD in computer science in 2007. He has several published papers in conferences and journals. Furthermore, he is an editorial board member of many international journals. Currently, he works on e-Learning environments, Collaborative learning, Artificial intelligence in education, Intelligent agents, MOOC.