

Article

Evidence-Grounded LLM Summarization for Actionable Student Feedback Analysis

Zhanerke Baimukanova ¹ , Yerassyl Saparbekov ¹, Hyesong Ha ²  and Minhoo Lee^{1,*} 

¹ Department of Computer Science, Nazarbayev University, Astana 010000, Kazakhstan; zhanerke.gabidenkyzy@nu.edu.kz (Z.B.); yerassyl.saparbekov@nu.edu.kz (Y.S.)

² Graduate School of Public Policy, Nazarbayev University, Astana 010000, Kazakhstan; hyesong.ha@nu.edu.kz

* Correspondence: minho.lee@nu.edu.kz

Abstract

Analyzing large-scale student feedback is critical for higher education quality assurance, yet manual analysis is inefficient and subjective. This paper proposes an integrated framework that unifies supervised classification, unsupervised clustering, and retrieval-augmented generation (RAG) to produce evidence-grounded and actionable insights. Ensemble-based supervised models perform thematic classification, while multi-encoder embedding fusion enables unsupervised discovery of coherent feedback clusters. A multi-stage RAG module integrates category predictions and cluster structure to retrieve representative evidence and generate transparent summaries with citation traceability. The framework is evaluated on student feedback collected from a Central Asian university and two public benchmarks, EduRABSA and Coursera course reviews, covering seven thematic categories. The supervised ensemble achieves 83.0% accuracy and 0.829 Macro-F1 on the primary dataset, while unsupervised clustering attains a silhouette score of 0.271 under the best fusion strategy. Independent evaluation on external benchmarks yields ensemble accuracy of 81.1% on EduRABSA and 49.8% on Coursera, confirming the framework's adaptability across diverse educational contexts. By leveraging supervised labels and unsupervised structure, the proposed framework enables evidence-grounded, category-aware LLM-based summaries that faithfully reflect the diversity and distribution of student feedback and support actionable educational decision-making.

Keywords: student feedback analysis; retrieval-augmented generation; supervised learning; unsupervised clustering; higher education analytics



Academic Editors: Tianchong Wang and Eric C. K. Cheng

Received: 13 February 2026

Revised: 31 March 2026

Accepted: 3 April 2026

Published: 7 April 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Student course evaluations are widely used to monitor teaching quality and inform curriculum development across higher education institutions. While quantitative ratings provide a coarse overview, open-ended student comments offer richer insights into instructional practices, learning resources, and student experience. However, manual analysis of such feedback is time-consuming, inconsistent, and susceptible to subjective bias [1,2]. As the scale of educational data continues to grow, recent research has increasingly applied natural language processing (NLP) and machine learning techniques to automate student feedback analysis [3,4].

Supervised learning approaches have formed the backbone of many educational NLP systems. Classical pipelines based on TF-IDF with logistic regression or support vector machines remain competitive baselines for text categorization [5–7]. Transformer-based

architectures have further advanced performance, with models such as BERT and RoBERTa successfully applied to course evaluations and educational text analysis [1,3,8,9], alongside XLNet and ALBERT [10,11]. More recently, parameter-efficient fine-tuning methods, including LoRA and related approaches [12,13], and few-shot approaches such as SetFit [14], have demonstrated that strong performance can be achieved in low-resource educational settings. Ensemble fusion strategies that combine complementary models have shown additional gains by leveraging strengths across classical, few-shot, and contextualized representations [15,16].

In parallel, unsupervised learning methods have been explored to uncover latent themes beyond predefined taxonomies. Classical topic models such as Latent Dirichlet Allocation (LDA) remain common in educational analytics [2,17], but they often struggle with short, noisy student feedback. Embedding-based clustering and neural topic modeling approaches, including BERTopic, have demonstrated improved semantic coherence and interpretability for educational reviews [18,19]. Complementary research investigates dimensionality reduction and multi-embedding fusion techniques to enhance clustering robustness and interpretability [20–23].

At the system level, retrieval-augmented generation (RAG) has emerged as a promising paradigm for linking neural text generation with explicit evidence retrieval. Initially proposed for knowledge-intensive question answering [24], RAG has since been extended through retrieval-enhanced and self-reflective generation frameworks [25,26], including adaptive retrieval strategies that dynamically select evidence based on query complexity or on demand [25,27]. Prior studies and recent surveys highlight its potential to improve factuality and transparency in decision-support settings while also noting challenges related to evaluation and integration [28,29].

Despite these advances, several limitations remain. First, supervised and unsupervised approaches are typically investigated in isolation, limiting insights into their complementary roles when applied to the same feedback data. Second, many existing datasets and evaluation settings are sentiment-focused and Western-centric, which may overlook institutional and cultural variability in higher education contexts [1,2]. Third, cross-dataset generalization is rarely examined, with most studies validating performance on a single institutional corpus without assessing robustness across diverse educational environments [20,30].

To address these gaps, this paper proposes EduFeedback-RAG, an integrated framework that combines supervised classification, unsupervised clustering, and retrieval-augmented generation for student feedback analysis. By unifying these components within a single pipeline, the proposed approach aims to support both structured institutional reporting and exploratory discovery, while maintaining transparency through evidence-grounded summarization. The framework is evaluated across multiple datasets to assess performance, robustness, and cross-dataset generalization.

2. Literature Review

Research on student feedback analysis spans educational NLP, transformer-based classification, short-text clustering, and retrieval-augmented generation. However, these areas have largely evolved in isolation, resulting in a persistent gap between classification accuracy, thematic interpretability, and transparency of summarized insights, highlighting the need for integrated analytical pipelines.

Recent years have seen significant advances in the automated analysis of course evaluations, driven by deep learning and transformer-based approaches. Prior studies show that both transformer-based models [8–11] and classical machine learning pipelines [5,6] remain competitive for sentiment analysis and thematic extraction, while additional work demon-

strates the effectiveness of deep learning approaches in educational feedback analysis across diverse datasets [1–4]. However, most existing research continues to prioritize classification accuracy and descriptive reporting, without addressing challenges related to summarization and transparency. Moreover, benchmark datasets for educational feedback remain fragmented; recent releases such as EduRABSA [31] and Coursera course reviews [32] offer opportunities for cross-dataset validation that have yet to be systematically explored.

General-purpose transformers such as BERT [8], RoBERTa [9], XLNet [10], and ALBERT [11] have demonstrated strong performance across a wide range of NLP tasks. In educational feedback analysis, prior work also reports strong performance of deep learning and transformer-based methods for text classification and sentiment analysis [1–4]. Parameter-efficient fine-tuning techniques such as LoRA and related PEFT methods [12,13] enable effective model adaptation to student feedback datasets even under limited computational resources. Few-shot learning frameworks, including SetFit [14], further reduce annotation requirements, making them particularly attractive for institutions with constrained labeling capacity. Beyond individual architectures, ensemble fusion strategies have demonstrated substantial performance gains by combining classifiers with complementary strengths. Classical stacked generalization [16] and more recent ensemble deep learning approaches [15] exploit synergies between feature-based models, few-shot learners, and contextualized transformers, consistently yielding improvements over single-model baselines. However, these techniques typically consider feedback classification in isolation and do not integrate downstream summarization or evidence-grounding, limiting their ability to support fully transparent educational analytics pipelines.

Understanding student concerns requires uncovering latent themes in large collections of short and noisy feedback. Classical topic modeling approaches such as Latent Dirichlet Allocation (LDA) [17] often struggle in this setting, motivating the use of neural and transformer-based clustering methods that provide improved semantic coherence and robustness for educational reviews [18,19]. Sentence-level embedding models, particularly Sentence-BERT and its multilingual extensions [20,21], have therefore become standard baselines for short-text clustering, often combined with density-based methods such as HDBSCAN [33]. Despite these advances, most clustering-based studies focus on theme discovery alone and do not generate grounded, verifiable summaries, leaving interpretation largely manual.

3. Dataset Collection and Annotation

3.1. Primary Dataset Collection

The primary dataset was collected through an anonymous online questionnaire administered to first-year undergraduate students enrolled in an introductory computer science programming course. The survey was distributed immediately after the final course session to capture recent and authentic reflections while the learning experience was still salient. Participation was voluntary, and no identifying information was requested to encourage candid and unbiased responses. All participants were informed about the purpose of the study and provided informed consent prior to participation.

The questionnaire consisted exclusively of open-ended questions, allowing students to freely describe their experiences, challenges, and suggestions in their own words. The questions were informed by prior literature on higher education quality evaluation and designed to cover seven pedagogically relevant dimensions, including teaching quality, course organization, and learning resources.

Each question implicitly targeted one or more thematic categories to facilitate subsequent annotation. For example, prompts such as “*What aspects of the teaching process helped or hindered your learning?*” corresponded to the *Teaching Quality* category, whereas “*How*

could this course be improved for future students?” elicited responses mapped to *Improvement Suggestions*. This design balanced open expression with structured thematic coverage.

3.2. Student Responses and Annotation

A total of 114 students participated in the survey, submitting 1087 textual responses across all questions. After removing incomplete, duplicated, or off-topic entries, 959 unique feedback items were retained for analysis. Each item comprised a self-contained sentence or short paragraph expressing a specific opinion, observation, or recommendation. All retained feedback entries were manually annotated and assigned to one of seven predefined categories, as summarized in Table 1. Two annotators collaboratively developed annotation guidelines to ensure labeling consistency. In cases where a comment referenced multiple aspects (e.g., teaching and workload), the entry was assigned to the dominant theme based on semantic emphasis.

Table 1. Distribution of labeled comments across thematic categories ($n = 959$).

Category	Example	Count	Percentage
Teaching Quality	Instructor explained topics clearly and answered questions patiently.	105	10.9
General Comments & Satisfaction	Good course, but pacing could be adjusted for beginners.	210	21.9
Improvement Suggestions	More coding examples would make the lectures more engaging.	95	9.9
Facilities/Inclusivity	Computers in the lab were slow, and Wi-Fi often disconnected.	172	17.9
Career Preparation	The assignments helped me understand industry-relevant projects.	96	10.0
Support Services	Technical support responded quickly when I had setup issues.	89	9.3
Social & Extracurricular	Hackathons organized by the department motivated me to participate.	192	20.0
Total		959	100.0

3.3. Preprocessing and Data Cleaning

Preprocessing was applied to prepare the feedback for downstream analysis and model training. The pipeline included the following steps: (i) removal of duplicate or near-duplicate responses, (ii) filtering of empty or irrelevant entries, (iii) text normalization through lowercasing, punctuation standardization, tokenization, and automated spell correction, and (iv) content-level anonymization to remove references to instructors or courses. Each cleaned feedback item was assigned a unique identifier to ensure traceability and reproducibility throughout subsequent experiments.

3.4. Data Augmentation and Expansion

Although the manually collected corpus contains nearly one thousand authentic feedback samples, its size remains limited for training deep neural models. To increase data diversity and reduce overfitting, a three-stage data augmentation pipeline was applied: (1) synonym replacement, which introduces lexical variation while preserving semantic meaning; (2) back-translation, in which feedback is translated through an intermediate language to induce syntactic diversity; and (3) LLM-based paraphrasing, which generates semantically equivalent rewrites that reflect realistic student writing styles.

This process produced augmented datasets containing approximately 10k, 50k, and 100k samples. All augmented corpora preserved the original label distribution and category definitions to ensure compatibility across experiments. Importantly, augmentation was applied exclusively to the training split, while the test set remained fixed and untouched throughout all experiments, ensuring that no data leakage occurs during evaluation.

3.5. External Validation Datasets

To evaluate generalizability beyond the primary institutional corpus, experiments were conducted on two external benchmark datasets: EduRABSA [31] and Coursera course reviews [32].

EduRABSA [31] is an aspect-based sentiment analysis dataset comprising approximately 6500 student reviews collected from multiple universities and online learning platforms. The dataset includes three-class sentiment annotations (positive, neutral, and negative) across aspects such as teaching quality, course content, assessment, and learning resources. These characteristics make EduRABSA suitable for evaluating robustness across institutional and contextual variations. Table 2 summarizes key characteristics across all three corpora.

Table 2. Comparison of dataset characteristics across primary and external validation corpora.

Characteristic	Primary (Ours)	EduRABSA [31]	Coursera [32]
Sample size	959	6500	8200+
Average length (tokens)	87.3	42.3	67.8
Learning context	In-person	Mixed	Online
Geographic scope	Central Asia	US/EU	Global
Label structure	7 categories	Aspect-level	Course-level

The Coursera course reviews dataset [32] contains over 8200 reviews from more than 450 massive open online courses across diverse disciplines. The reviews reflect online learning contexts characterized by asynchronous delivery, self-paced progression, and globally distributed learners, providing complementary linguistic and contextual properties.

4. Methodology

This study proposes an integrated analytical framework that combines supervised classification, unsupervised clustering, and retrieval-augmented generation (RAG) to transform unstructured student feedback into evidence-grounded and institutionally actionable insights. As illustrated in Figure 1, the framework operates across three functional layers and executes three parallel analytical pipelines that jointly support structured monitoring, exploratory analysis, and transparent summarization.

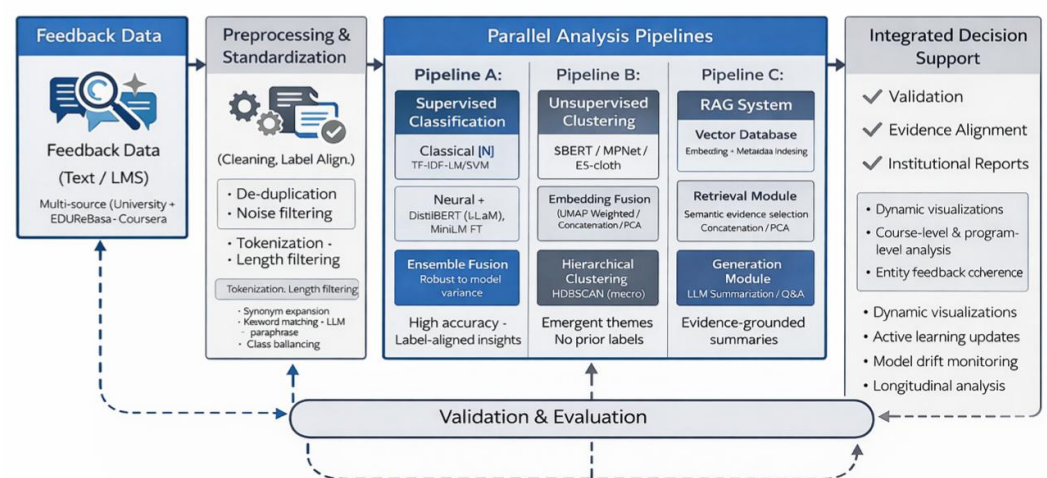


Figure 1. Integrated framework for student feedback analysis combining supervised classification, unsupervised clustering, and RAG-based retrieval and summarization.

4.1. Framework Overview

The framework is organized into three layers:

- **Layer 1: Data Collection and Preparation** constructs standardized, analysis-ready corpora from institutional and external datasets through cleaning, standardization, and harmonized preprocessing.
- **Layer 2: Parallel Analysis Pipelines** executes three complementary analytical modules: (A) supervised classification for category-aware feedback organization, (B) unsupervised clustering for latent theme discovery, and (C) retrieval-augmented generation for evidence-grounded summarization.
- **Layer 3: Integration and Decision Support** synthesizes outputs from all pipelines, cross-validates analytical signals, and produces interpretable summaries and diagnostic reports to support institutional decision-making.

The framework is modular: each pipeline can operate independently, while their integration enables richer interpretation, internal validation, and transparency.

4.2. Supervised Classification Module

The supervised classification module assigns each feedback instance to one of the seven thematic categories defined during dataset annotation (Section 3). These categories represent institutionally relevant dimensions of educational quality, including Teaching Quality, Improvement Suggestions, and Facilities and Inclusivity. Supervised predictions serve both as analytical outputs and as explicit routing constraints for downstream clustering analysis and RAG-based retrieval.

4.2.1. Supervised Classification Models

Seven supervised models spanning classical, neural, and transformer-based paradigms were implemented:

- TF-IDF + Logistic Regression (LR) [6].
- TF-IDF + Support Vector Machine (SVM) [5].
- Hybrid TF-IDF + MPNet [34].
- Deep Text-Level Processing (DTLP) [35].
- Multi-Head Attention Fusion (MHAF) [36].
- SetFit (contrastive few-shot learning) [14].
- DeBERTa-v3 with Low-Rank Adaptation (LoRA) [12].
- Weighted ensemble (TF-IDF + SVM, SetFit, DeBERTa-LoRA) (ours).

Classical TF-IDF models provide interpretable lexical baselines with low computational cost, making them suitable as strong reference methods for short-text classification [5,6]. Neural architectures such as DTLP and MHAF capture hierarchical and attention-based patterns in longer feedback sequences [35,37]. Transformer-based models leverage contextual semantic representations learned from large-scale pretraining, with SetFit enabling data-efficient adaptation through contrastive learning [14] and DeBERTa-v3 with Low-Rank Adaptation (LoRA) providing strong generalization under limited annotation budgets [12].

4.2.2. Ensemble Fusion Strategy

In this study, we adopt an ensemble strategy that combines models with complementary inductive biases to ensure robust performance across heterogeneous feedback styles. The final ensemble integrates TF-IDF+SVM, SetFit, and DeBERTa-LoRA using weighted late fusion.

Each base model produces a probability distribution over categories, and the ensemble prediction is computed as

$$\hat{y}^{\text{ens}} = \arg \max_{c \in \{1, \dots, K\}} \sum_{i=1}^3 w_i P_{M_i}(c | x), \quad (1)$$

where $w_i \geq 0$, $\sum_i w_i = 1$, and $\hat{y}^{\text{ens}}$ denotes the final supervised category predicted by the ensemble classifier.

4.2.3. Training and Evaluation

All supervised models are evaluated using stratified 5-fold cross-validation on the annotated dataset described in Section 3. Model performance is measured using accuracy and the Macro-F1 score. While accuracy reflects overall classification correctness, Macro-F1 is used as the primary evaluation metric to account for class imbalance by equally weighting performance across all categories. All models are optimized using AdamW and trained with 20 epochs. Experiments are conducted on a system equipped with two NVIDIA RTX 5090 GPUs.

Weighted ensemble fusion uses $w_{\text{TF-IDF}} = 0.35$, $w_{\text{DeBERTa-LoRA}} = 0.25$, and $w_{\text{SetFit}} = 0.40$, selected via grid search on the validation set optimizing Macro-F1. LoRA is applied to the query and value projection layers of DeBERTa-v3-base with rank $r = 16$, scaling factor $\alpha = 32$, and dropout 0.1. HDBSCAN is configured with `min_cluster_size= 15` and `min_samples= 5`. PCA reduces concatenated embeddings to 47 dimensions preserving approximately 66% variance. SetFit is trained for 1 epoch with 20 iterations using cosine contrastive loss.

4.3. Unsupervised Clustering Module

The unsupervised clustering module aims to uncover latent semantic structure in student feedback without relying on predefined institutional categories. While supervised classification enforces an explicit taxonomy, clustering provides a data-driven view of semantic organization, revealing sub-themes, emerging concerns, and intra-category heterogeneity.

4.3.1. Sentence Embedding Models

Each feedback instance f is mapped to dense semantic representations using multiple pretrained sentence encoders obtained from the Hugging Face model repository. Specifically, we employ (1) GTE-large [37], (2) BGE-large, (3) MXBAI-Embed-Large, (4) Instructor-XL [38], and (5) the MPNet-based sentence transformer [34]. For each encoder E_j , an embedding is produced as

$$\mathbf{e}^{(j)} = E_j(f), \quad j = 1, \dots, m,$$

where all embeddings are ℓ_2 -normalized prior to fusion.

4.3.2. Embedding Fusion

We adopt feature-level concatenation to preserve encoder-specific information without assuming cross-space alignment. The fused representation for feedback instance f_i is defined as

$$\mathbf{z} = [\mathbf{e}_i^{(1)} \parallel \mathbf{e}_i^{(2)} \parallel \dots \parallel \mathbf{e}_i^{(m)}],$$

where $\parallel$ denotes vector concatenation.

To mitigate the high dimensionality and dominance effects inherent to concatenation, principal component analysis (PCA) is applied as an unsupervised dimensionality reduction, yielding a compact and stable representation for subsequent clustering and retrieval.

4.3.3. Clustering and Pseudo-Label Induction

Clustering is performed on the fused embedding set $\{\mathbf{z}_i\}_{i=1}^N$ to induce unsupervised pseudo-labels that reflect latent semantic organization beyond predefined category boundaries.

HDBSCAN [33] is employed as the primary unsupervised clustering method, as it discovers intrinsic semantic structure without requiring the number of clusters to be specified in advance. Applied to $\{\mathbf{z}_i\}_{i=1}^N$, HDBSCAN produces cluster assignments

$$\hat{y}_i^{\text{hdb}} \in \{1, \dots, C\} \cup \{-1\},$$

where -1 denotes noise points or semantically ambiguous feedback instances. Given the resulting set of C clusters, each cluster is subsequently aligned with predefined educational feedback categories reported in the prior literature [1–3]. For each cluster, a cluster-level semantic representation is constructed by aggregating the embeddings of its constituent feedback instances. Similarity scores are then computed between these cluster representations and category-level semantic profiles derived from the predefined feedback taxonomy. Each cluster is associated with the category exhibiting the highest semantic alignment score, yielding a soft semantic correspondence rather than a strict categorical assignment.

4.3.4. Unsupervised Clustering Evaluation Metrics

To assess the quality of clustering structures in the absence of ground-truth labels, we adopt three intrinsic clustering metrics commonly used in unsupervised text analysis. The silhouette score evaluates cluster compactness and separation by comparing intra-cluster cohesion with inter-cluster distance. The Davies–Bouldin index measures the average similarity between each cluster and its most similar neighboring cluster. The Calinski–Harabasz index quantifies the ratio of between-cluster dispersion to within-cluster dispersion. In addition to these metrics, we report the proportion of noise points identified by HDBSCAN, representing feedback instances that are not confidently assigned to any cluster. The noise ratio reflects the degree of semantic ambiguity in the data and the conservativeness of the clustering process rather than a direct measure of clustering performance.

4.4. Retrieval-Augmented Generation (RAG) Module

The RAG module serves as the synthesis layer of the framework, integrating outputs from supervised classification and unsupervised clustering to generate structured, evidence-grounded summaries.

Each feedback instance f_i is represented as

$$f_i = \langle \text{text}_i, y_i^{\text{ens}}, \hat{y}_i^{\text{hdb}}, \mathbf{v}_i \rangle, \quad (2)$$

where y_i^{ens} denotes the ensemble-predicted supervised category, $\hat{y}_i^{\text{hdb}}$ is the HDBSCAN-induced cluster identifier, and $\mathbf{v}_i \in \mathbb{R}^d$ is the final fused embedding obtained after concatenation and PCA (Section 4.3.2).

Given a query q , the ensemble classifier first predicts a category y_q^{ens} . Retrieval candidates are then restricted to

$$\mathcal{F}_{y_q^{\text{ens}}} = \{f_i \mid y_i^{\text{ens}} = y_q^{\text{ens}}\}, \quad (3)$$

ensuring alignment with institutional taxonomy and preventing cross-category semantic drift. Within $\mathcal{F}_{y_q^{\text{ens}}}$, feedback instances are ranked by cosine similarity

$$s_i = \cos(\mathbf{v}_q, \mathbf{v}_i), \quad (4)$$

where $\mathbf{v}_q$ denotes the fused embedding of the query. HDBSCAN cluster identifiers are retained to preserve coverage across latent semantic sub-themes and to avoid over-representation of dominant patterns during retrieval. While more sophisticated adaptive retrieval strategies—such as Self-RAG [25], which retrieves on demand via learned reflection tokens, or Adaptive-RAG [27], which routes queries by complexity—could further refine evidence selection, our keyphrase-driven top- k approach was chosen for its interpretability and computational efficiency in the educational domain context.

4.5. Evidence-Grounded Generation

The generation stage operates exclusively on retrieved feedback instances and does not access the full corpus directly. For a given query q , a fixed-size set of top- K feedback instances is retrieved as

$$\mathcal{R}_q = \{f_{i_1}, \dots, f_{i_K}\} \subset \mathcal{F}_{y_q^{\text{ens}}}, \quad (5)$$

where each instance is represented as $f_i = \langle \text{text}_i, y_i^{\text{ens}}, \hat{y}_i^{\text{hdb}}, \mathbf{v}_i \rangle$ (see Section 4.4).

The large language model (LLM) receives $\mathcal{R}_q$ as explicit evidence together with structured supervisory signals derived from upstream analysis. The RAG summarization module uses GPT-4o-mini as the language model. Specifically, the input to the generator consists of (i) the raw feedback texts $\{\text{text}_i\}_{f_i \in \mathcal{R}_q}$, (ii) the ensemble-predicted category y_q^{ens} , and (iii) the set of HDBSCAN cluster identifiers $\{\hat{y}_i^{\text{hdb}}\}_{f_i \in \mathcal{R}_q}$.

The supervised category prediction y_q^{ens} conditions the high-level structure of the generated summary, ensuring alignment with institutional taxonomy. HDBSCAN cluster identifiers are used to aggregate, contrast, and balance latent semantic sub-themes within the retrieved evidence, enabling the generator to reflect internal diversity and avoid dominance by frequent or redundant patterns.

Formally, the generation process is expressed as

$$\hat{S}_q = \text{LLM}(\mathcal{R}_q \mid y_q^{\text{ens}}, \{\hat{y}_i^{\text{hdb}}\}), \quad (6)$$

where $\hat{S}_q$ denotes the generated summary.

All generated statements are explicitly grounded in $\mathcal{R}_q$, and provenance is preserved through joint reference to supervised category labels and unsupervised cluster identifiers. Unlike retrieval-only RAG architectures that condition generation solely on semantic similarity, the proposed approach jointly incorporates supervised classification and unsupervised clustering signals, enabling controllable, transparent, and institutionally aligned summarization.

5. Results

5.1. Supervised Learning

Figure 2 compares the performance of seven supervised models and the proposed ensemble across three dataset scales. As the dataset size increases from the manually annotated corpus ($n = 959$) to the 100k augmented dataset, transformer-based architectures exhibit consistent performance gains, whereas classical models show early saturation with marginal improvement. Across all scales, the TF-IDF and Hybrid models remained stable around 75%, suggesting early convergence with limited benefit from additional data. DTLF and MHAF improved moderately (from 61–69% to about 70–74%), showing some sensitivity to data diversity. DeBERTa-LoRA exhibited the steepest increase (from 19.1% to 79.2%), followed by MPNet fine-tuning (from 20.1% to 68.7%), while SetFit maintained consistently high results (77–78%) even on small datasets. These outcomes confirm that larger data volume primarily benefits transformer-based fine-tuning approaches. Based on

these observations, we selected the 50k dataset configuration as a representative large-scale setting and conducted additional classification experiments below.

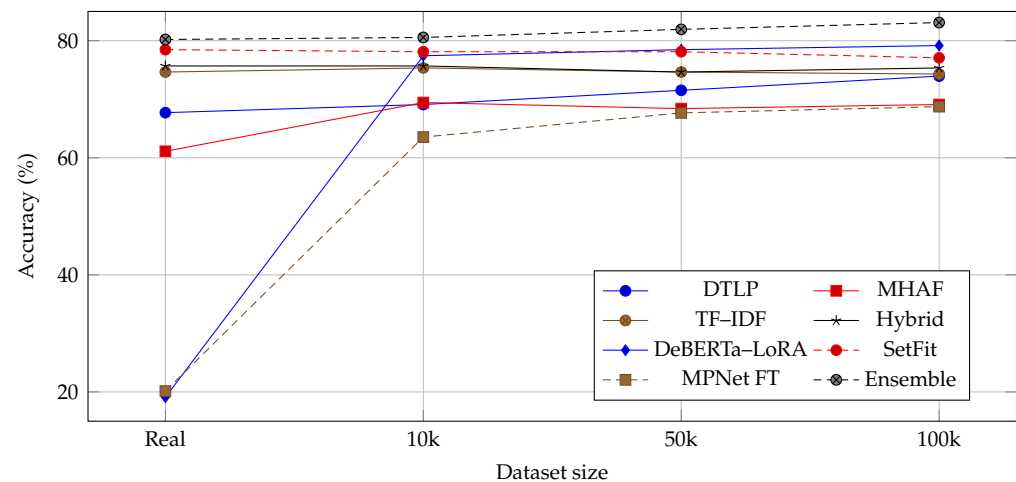


Figure 2. Accuracy versus dataset size for the seven supervised models (test set $n = 288$).

Table 3 presents supervised learning performance across three datasets evaluated independently using 5-fold cross-validation. On the primary dataset, TF-IDF achieves 80.0%/0.798 and SetFit reaches 84.5%/0.846, while DeBERTa-LoRA attains 83.3% (± 1.2)/0.820. The proposed ensemble achieves 83.0% (± 2.0) accuracy and 0.829 Macro-F1. A paired t -test between DeBERTa-LoRA and the ensemble on the primary dataset confirms that the difference in mean accuracy is not statistically significant ($p > 0.05$); however, the ensemble demonstrates a competitive F1 and substantially superior cross-dataset generalization. On EduRABSA, the ensemble achieves 81.1% (± 1.4)/0.778, the best result among all models. On Coursera, the ensemble reaches 49.8% (± 0.7)/0.534, again the highest overall. Across all three datasets, the ensemble consistently outperforms individual models, confirming its effectiveness and robustness across diverse educational contexts.

Table 3. Supervised learning performance across datasets (mean $\pm$ standard deviation over 5-fold CV).

Model	Primary		EduRABSA		Coursera	
	Acc	F1	Acc	F1	Acc	F1
<i>Classical Models</i>						
TF-IDF + LR/SVM	80.0 (± 2.7)	0.798	72.9 (± 0.9)	0.700	48.3 (± 1.6)	0.492
Hybrid (TF-IDF + Emb)	82.2 (± 2.2)	0.821	74.4 (± 1.1)	0.715	49.4 (± 0.9)	0.519
<i>Neural Models</i>						
DTLP	74.5 (± 3.4)	0.746	59.4 (± 0.7)	0.546	47.5 (± 1.0)	0.402
MHAF	71.6 (± 1.9)	0.714	61.5 (± 0.7)	0.547	48.8 (± 0.8)	0.384
<i>Transformer Models</i>						
SetFit	84.5 (± 2.5)	0.846	80.2 (± 1.1)	0.774	49.3 (± 0.7)	0.469
DeBERTa-LoRA	83.3 (± 1.2)	0.820	73.2 (± 1.1)	0.699	47.1 (± 1.2)	0.420
MPNet FT	79.2 (± 2.5)	0.799	72.9 (± 0.5)	0.688	48.4 (± 1.9)	0.438
Ensemble (Ours)	83.0 (± 2.0)	0.829	81.1 (± 1.4)	0.778	49.8 (± 0.7)	0.534

Figure 3 visualizes per-category F1 scores across datasets, revealing differential transfer patterns. Categories with explicit lexical markers (*Teaching Quality* and *Career Preparation*) maintained 90–95% of primary performance, while semantically complex categories (*General Comments* and *Improvement Suggestions*) showed larger drops (15–20%), reflecting increased sensitivity to institutional discourse patterns.

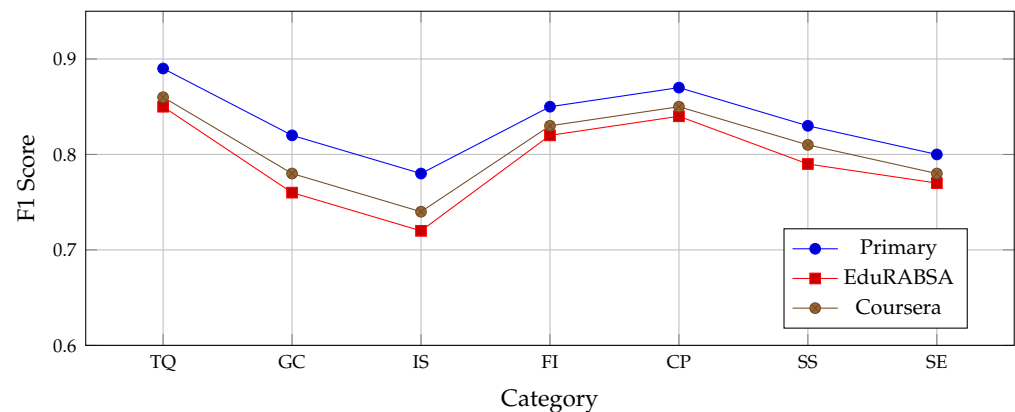


Figure 3. Per-category F1 scores across datasets (ensemble model). Abbreviations: TQ = Teaching Quality, GC = General Comments, IS = Improvement Suggestions, FI = Facilities/Inclusivity, CP = Career Preparation, SS = Support Services, and SE = Social/Extracurricular.

5.2. Unsupervised Clustering

Based on the five transformer-based sentence encoders introduced in Section 4.3.1, we implement four embedding fusion strategies: average ensemble, weighted ensemble, concatenation, and concatenation followed by PCA. The average ensemble computes the mean of ℓ_2 -normalized embedding vectors, whereas the weighted ensemble applies learned fusion weights to emphasize more reliable encoders. Concatenation directly combines embeddings from all encoders to preserve model-specific information, resulting in a high-dimensional fused representation. The Concat+PCA strategy applies principal component analysis to the concatenated embeddings, reducing the dimensionality from 2816 to 47 while preserving 66.6% of the cumulative variance, thereby retaining the underlying semantic structure.

Table 4 summarizes the HDBSCAN clustering performance under four embedding fusion strategies. Simple concatenation performs worst, yielding a low silhouette score of 0.219, a high Davies–Bouldin index of 1.521, and the largest noise ratio (721 instances, 75.1%), indicating fragmented and unstable clusters. The average ensemble improves clustering quality, achieving a silhouette score of 0.238. The weighted ensemble further enhances cluster structure, increasing the silhouette score to 0.259, lowering the Davies–Bouldin index to 1.287. Overall, the Concat + PCA strategy achieves the best performance, with the highest silhouette score of 0.271, the lowest Davies–Bouldin index of 1.241, and the highest Calinski–Harabasz index of 26.94, while also yielding the lowest noise proportion at 412 instances (42.9%).

Table 4. HDBSCAN-based clustering performance using four different embedding fusion strategies. Higher silhouette and Calinski–Harabasz scores indicate better clustering quality, while a lower Davies–Bouldin index indicates better separation.

Fusion Strategy	Silhouette $\uparrow$	Davies–Bouldin $\downarrow$	Calinski–Harabasz $\uparrow$	Clusters	Noise
Average Ensemble	0.238	1.369	24.12	8	612 (63.7%)
Weighted Ensemble	0.259	1.287	25.67	9	564 (58.7%)
Concatenation	0.219	1.521	22.34	11	721 (75.1%)
Concat + PCA	0.271	1.241	26.94	10	412 (42.9%)

To qualitatively examine the clustering behavior induced by different embedding representations, Figure 4 illustrates t-SNE projections of feedback embeddings obtained under each fusion strategy, with points colored according to HDBSCAN cluster assignments. Samples marked as noise by HDBSCAN are indicated separately.

Compared to naive concatenation and ensemble averaging, the proposed Concat + PCA fusion yields more compact and clearly separated cluster structures in the low-dimensional projection. Under this representation, HDBSCAN identifies ten stable clusters, which correspond to semantically coherent feedback groupings rather than pre-defined institutional categories. These clusters reflect distinct latent themes present in the feedback corpus and form the basis for subsequent pseudo-label interpretation and retrieval conditioning. Following the cluster interpretation procedure described in Section 4.3.3, each HDBSCAN-derived cluster was assigned a descriptive semantic label based on its dominant feedback patterns. The resulting cluster labels include AS (Academic Satisfaction), CP (Career Preparation), EA (Extracurricular Activities), FR (Facilities and Resources), IS (Improvement Suggestions), ID (Inclusivity and Diversity), OC (Overall Comments), SE (Social Experience), SS (Support Services), and TQ (Teaching Quality).

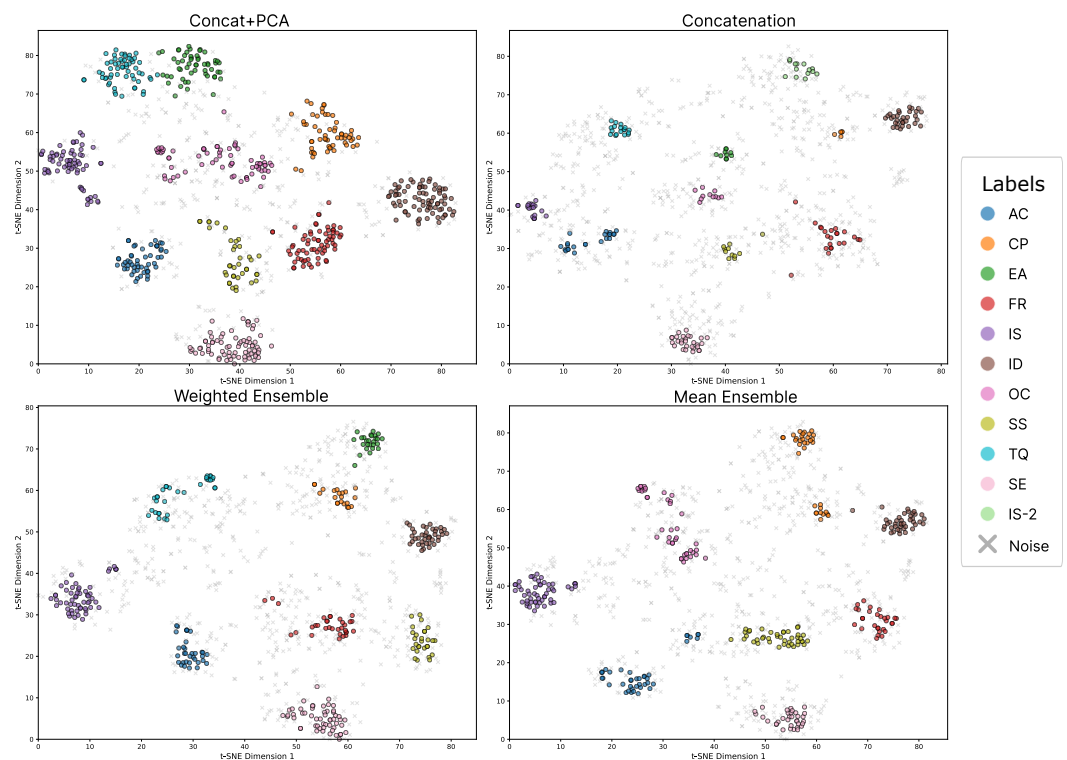


Figure 4. t-SNE visualization of fused embedding representations obtained using different fusion strategies. Points are color-coded by HDBSCAN cluster assignment, while gray crosses represent samples labeled as noise. The visualization highlights qualitative differences in cluster compactness and noise distribution across fusion methods.

5.3. RAG-Based Summarization

This section evaluates the proposed RAG-based summarization module for generating evidence-grounded, category-aware summaries of student feedback. The evaluation covers 959 feedback instances grouped into seven institutional categories. The analysis focuses on whether the generated summaries remain traceable to source feedback, contain explicit quantitative evidence, and reflect structural signals derived from supervised classification and unsupervised clustering.

To quantitatively evaluate the RAG module, we assess whether generated summary statements are grounded in the retrieved evidence, following established faithfulness evaluation protocols for RAG systems [29,39]. Summaries are decomposed into sentence-level statements, each judged by an automated LLM-based evaluator (GPT-4o-mini) with the following prompt: “Is the following statement supported by the evidence? Answer Supported/Not Supported.” Faithfulness and cluster coverage are defined as

$$\text{Faithfulness} = \frac{N_{\text{supported}}}{N_{\text{total}}}, \quad \text{Cluster Coverage} = \frac{C_{\text{represented}}}{C_{\text{retrieved}}} \quad (7)$$

Table 5 reports results across three configurations: (1) Baseline RAG with no filtering, (2) Supervised RAG using category labels only, and (3) the proposed Supervised + Unsupervised RAG integrating both category labels and HDBSCAN cluster identifiers. The results confirm that each component contributes measurably, with the proposed method achieving the highest faithfulness (0.93) and cluster coverage (0.64).

Table 5. Evaluation of RAG summarization strategies (ablation study). Higher values indicate better performance.

Model	Faithfulness ↑	Cluster Coverage ↑
Baseline RAG	0.71	0.59
Supervised RAG	0.89	0.60
Supervised + Unsupervised RAG (Ours)	0.93	0.64

Content-based analysis of the RAG outputs indicates that the generated summaries consistently include explicit quantitative evidence and structured recommendations derived from the retrieved feedback. The summaries preserve category and cluster traceability, enabling transparent verification of synthesized statements against source instances.

On average, each summary reports multiple statistical indicators and quantified recommendations, such as proportions of students mentioning an issue, prevalence of dominant themes, and recommended intervention magnitudes. All reported quantities are computed directly from the retrieved evidence subset used during generation, ensuring that percentages and action items remain anchored in the analyzed feedback rather than inferred heuristically.

Table 6 provides qualitative result examples illustrating RAG behavior under different levels of supervised–unsupervised alignment. In the high-agreement case (Teaching Quality), the majority of feedback maps to a single coherent cluster (84% in C0) with high purity (96.9%). This alignment enables stable aggregation and precise statistical reporting, resulting in summaries that highlight dominant pedagogical concerns and produce clearly quantified recommendations.

Table 6. Comparison of RAG summaries under high and moderate supervised–unsupervised agreement.

Component	Example 1: High Agreement	Example 2: Split Clusters
Category	Teaching Quality (supervised ensemble)	General Comments & Satisfaction (supervised ensemble)
Samples analyzed	105 total → 10 via RAG (top- <i>k</i>)	210 total → 10 via RAG (top- <i>k</i>)
Cluster mapping	84% → C0 (96.9% purity)	41% → C1 (90.8% purity); 32% → C3; 27% noise
Agreement level	High (C0 predominantly Teaching Quality)	Moderate (category spans multiple clusters)
Top keywords	lectures, clarity, examples, pacing, difficult, fast	overall, satisfied, workload, balance, experience
Generated summary	Lecture clarity and pacing issues dominate, with 68% negative sentiment. Difficulty following explanations (84%), rapid pacing (76%), and insufficient worked examples (62%) were frequent. Actions include 3–4 worked examples per lecture, pause intervals, scaffolding materials, and reduced content density.	Feedback is heterogeneous: 58% overall satisfaction, with helpfulness (45%) and challenging-but-rewarding perceptions (38%). Concerns include workload/resources (32%) and work–life balance (27%). Outputs include both actionable items and diagnostic flags for manual review.
Statistics provided	4 indicators (68%, 84%, 76%, 62%)	6 indicators (58%, 45%, 38%, 32%, 42%, 35%)
Actionability	4 quantified recommendations	Mixed: actionable + diagnostic
Traceability	Category: Teaching Quality; Cluster: C0; IDs: FB_0023, FB_0087, FB_0134, FB_0156, FB_0189, FB_0203, FB_0267, FB_0289, FB_0312, FB_0378	Category: General Comments; Clusters: C1, C3, noise; IDs: FB_0156, FB_0189, FB_0234, FB_0278, FB_0301, FB_0345, FB_0389, FB_0412, FB_0456, FB_0489

The same table also illustrates a split-cluster scenario for the General Comments category, where feedback is distributed across multiple clusters (41% in C1, 32% overlapping with C3, and 27% noise). Instead of forcing a single aggregated narrative, the RAG module explicitly reports this distribution and generates a structured synthesis reflecting multiple sub-themes. This output exposes internal heterogeneity within broad institutional categories and reduces the risk of overgeneralized conclusions.

Overall, the results demonstrate that the RAG module operates as an interpretive synthesis layer rather than a purely generative component. High supervised–unsupervised agreement yields focused, statistically precise summaries, while split-cluster cases produce diagnostic outputs that surface internal diversity and potential category ambiguity. These behaviors motivate the subsequent Discussion, where the implications of alignment and heterogeneity are analyzed in the context of institutional interpretation and category refinement.

6. Discussion and Limitations

This section discusses the implications of the proposed integrated framework; clarifies the benefits of combining supervised classification, unsupervised clustering, and retrieval-augmented generation; and critically examines the system’s limitations and generalizability in real educational settings.

6.1. Value of Integrated Analysis

The primary contribution of this work lies not in introducing new individual learning algorithms but in demonstrating that effective student feedback analysis requires the joint interpretation of thematic structure and evidence-grounded synthesis [3,19]. Experimental results show that each analytical pipeline captures a complementary aspect of feedback: supervised thematic categorization assigns feedback to institutionally defined categories [1,2,6], unsupervised clustering reveals latent themes [17–19], and retrieval-augmented generation translates dispersed feedback into concise summaries and recommendations [24,25,28]. When applied independently, these components provide only partial insight [3]. Their integration enables analyses that are not achievable with isolated methods, such as identifying which thematic areas are associated with predominantly negative student responses and generating summaries that reflect both thematic coverage and urgency.

6.2. From Analysis to Actionability

A key limitation of many existing feedback analysis systems is the gap between analytical outputs and pedagogical decision-making [2,3]. Aggregate scores or lists of topics often fail to indicate concrete actions [19]. The proposed framework addresses this gap by using supervised thematic category labels and unsupervised topic assignments as constraints for retrieval in the RAG-based summarization pipeline [24,25]. As a result, generated summaries are not generic condensations of feedback but action-oriented syntheses grounded in representative student comments [28]. Formal human evaluation with in-service educators remains an important direction for future work.

6.3. Robustness and Interpretability

The parallel architecture adopted in this work enhances both robustness and interpretability [3,19]. Because thematic categorization and topic discovery operate independently, inconsistencies between category assignments and thematic structure can be identified and inspected [2,19]. For example, topics associated with predominantly negative student responses can be prioritized for intervention, while mixed-response topics can be examined more closely. This separation of concerns aligns with decision-support

principles in educational technology, where transparency and traceability are essential for institutional trust and adoption [19].

6.4. Generalizability and Scope

The proposed framework is not intended as a universal sentiment or summarization model but as a decision-support architecture for educational feedback analysis [3,19]. Its modular design allows for individual components to be retrained or replaced as institutional contexts evolve [13]. While results demonstrate effectiveness across university-based and online learning platforms [1,31,32], generalizability should be interpreted in terms of framework applicability rather than direct model portability [40]. This distinction is particularly important in educational settings, where contextual and cultural differences strongly influence feedback expression [2].

6.5. Implications for Educational Practice

By enabling scalable, interpretable, and actionable analysis of student feedback, the framework supports evidence-based teaching improvement and institutional quality assurance [2,19]. Automated integration of thematic structure and grounded summaries allows educators to engage with feedback at a level of abstraction that is both manageable and meaningful [1,3], reducing reliance on anecdotal impressions or selective manual review. This positions the proposed system as a practical tool for continuous improvement rather than a purely analytical artifact.

6.6. Limitations

While the proposed framework demonstrates its strengths, several limitations remain. First, although the RAG module is evaluated using evidence-grounded automatic metrics (Table 5), formal human evaluation was not included in the current study. In particular, usability validation with in-service educators in real educational settings would further strengthen the practical relevance of the framework. Second, because the student feedback datasets do not provide ground-truth reference summaries, reference-based summarization metrics such as ROUGE and BERTScore cannot be applied in our setting; the faithfulness metric in Table 5 serves as a practical proxy. Third, although the supervised classification module was evaluated across multiple datasets, the clustering and RAG components were primarily validated within the current task setting; cross-dataset evaluation of these components would require either dataset-specific ground-truth summaries or harmonized cluster schemas, which are not currently available, and this is identified as an important direction for future work. Fourth, a more fine-grained component-level ablation, including alternative fusion strategies and clustering configurations, would further clarify the contribution of each design choice. Fifth, our automated faithfulness evaluation uses GPT-4o-mini as both the summary generator and the judge, introducing a potential self-referential bias: the evaluator may systematically favor outputs whose style aligns with its own generation patterns, independent of actual summary quality. This circularity is a recognized limitation of LLM-as-judge evaluation protocols [29,39]. The faithfulness scores in Table 5 should therefore be interpreted as indicative rather than definitive; future work should validate these results using diverse judge models or human annotation.

6.7. Future Research Directions

One promising direction for future work is to extend the framework to incorporate explicit sentiment analysis alongside the current seven-category thematic structure [1,2,6]. This would require redefining the label space (e.g., positive, neutral, and negative) and establishing reliable annotation and validation protocols to ensure consistency and reliability [21]. Such an extension would allow for the system to jointly capture both the thematic

category and sentiment polarity of student feedback, enabling richer and more nuanced institutional reporting. In addition, further research could explore improving the faithfulness and grounding of RAG-based summarization through more rigorous evaluation strategies, including human-centered assessment [28]. Finally, investigating the robustness of the framework under heterogeneous datasets and varying annotation schemes would further strengthen its applicability in real-world educational settings [20,30]. A particularly important extension is the cross-dataset evaluation of the unsupervised clustering and RAG summarization components. The current study evaluates these modules exclusively on the primary institutional corpus; extending them to external benchmarks such as EduRABSA [31] and Coursera [32] would require either dataset-specific ground-truth summaries or harmonized cluster label schemas. Developing such annotation resources and standardized evaluation protocols for educational RAG and clustering systems represents a meaningful open challenge for the community [29,39].

7. Conclusions

This paper presented an integrated framework for student feedback analysis that unifies supervised classification, unsupervised theme discovery, and evidence-grounded summarization. By combining these complementary paradigms, the framework addresses key limitations of existing educational analytics approaches, particularly with respect to interpretability, structural validation, and the actionability of insights derived from unstructured feedback.

The results demonstrate that jointly leveraging supervised and unsupervised signals enables more nuanced interpretation of student feedback than either approach in isolation. High alignment between predicted categories and discovered clusters supports the validity of institutional taxonomies, while split-cluster behavior exposes internal heterogeneity within broad categories. The RAG-based summarization component translates these analytical signals into structured, traceable outputs that emphasize quantitative grounding and practical relevance rather than purely linguistic abstraction.

Overall, the proposed framework advances student feedback analysis beyond descriptive reporting toward interpretive, evidence-based synthesis. By exposing both dominant patterns and structural ambiguities within feedback data, it supports informed institutional reflection on teaching quality, feedback instrument design, and quality assurance practices. The framework thus contributes a scalable and transparent approach for transforming unstructured student feedback into actionable knowledge within higher education contexts.

Author Contributions: Conceptualization, M.L.; methodology, Z.B. and Y.S.; software, Z.B. and Y.S.; validation, Z.B. and Y.S.; formal analysis, Z.B. and Y.S.; investigation, Z.B. and Y.S.; data curation, Z.B. and Y.S.; writing—original draft preparation, Z.B. and Y.S.; writing—review and editing, H.H. and M.L.; visualization, Z.B. and Y.S.; supervision, H.H. and M.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. AP27510939).

Institutional Review Board Statement: This study received approval from the Institutional Research Ethics Committee of Nazarbayev University (NU-IREC 738/31052025).

Informed Consent Statement: Informed consent was obtained from all participants involved in the data collection process.

Data Availability Statement: The data presented in this study are available by request. The code is publicly available at <https://github.com/nubcico/EduFeedback> (accessed on 2 April 2026).

Acknowledgments: During the preparation of this manuscript, the authors used ChatGPT (GPT-5, OpenAI) for English language editing and minor proofreading. The authors reviewed and edited all outputs and take full responsibility for the content of this publication.

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

References

1. Kastrati, Z.; Imran, A.; Kurti, A. Weakly Supervised Framework for Aspect-Based Sentiment Analysis on Students' Reviews of MOOCs. *IEEE Access* **2020**, *8*, 106799–106810. [\[CrossRef\]](#)
2. Ren, P.; Yang, L.; Luo, F. Automatic Scoring of Student Feedback for Teaching Evaluation Based on Aspect-Level Sentiment Analysis. *Educ. Inf. Technol.* **2023**, *28*, 797–814. [\[CrossRef\]](#)
3. Kastrati, Z.; Dalipi, F.; Imran, A.; Pireva Nuci, K.; Wani, M. Sentiment Analysis of Students' Feedback with NLP and Deep Learning: A Systematic Mapping Study. *Appl. Sci.* **2021**, *11*, 3986. [\[CrossRef\]](#)
4. Sindhu, I.; Daudpota, S.; Badar, K.; Bakhtyar, M.; Baber, J.; Nurunnabi, M. Aspect-Based Opinion Mining on Student's Feedback for Faculty Teaching Performance Evaluation. *IEEE Access* **2019**, *7*, 108729–108741. [\[CrossRef\]](#)
5. Joachims, T. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In *Proceedings of the European Conference on Machine Learning (ECML), Chemnitz, Germany, 21–23 April 1998*; Springer: Berlin/Heidelberg, Germany, 1998; pp. 137–142. [\[CrossRef\]](#)
6. Sebastiani, F. Machine Learning in Automated Text Categorization. *ACM Comput. Surv.* **2002**, *34*, 1–47. [\[CrossRef\]](#)
7. Dalipi, F.; Imran, A.; Kastrati, Z. MOOC Dropout Prediction Using Machine Learning Techniques: Review and Research Challenges. In *Proceedings of the IEEE Global Engineering Education Conference (EDUCON), Tenerife, Spain, 17–20 April 2018*; IEEE: New York, NY, USA, 2018; pp. 1007–1014. [\[CrossRef\]](#)
8. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT), Minneapolis, MN, USA, 2–7 June 2019*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 4171–4186. [\[CrossRef\]](#)
9. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv* **2019**, arXiv:1907.11692.
10. Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R.; Le, Q. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In *Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019*; Curran Associates Inc.: Red Hook, NY, USA, 2019; pp. 5753–5763.
11. Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; Soricut, R. ALBERT: A Lite BERT for Self-Supervised Learning of Language Representations. In *Proceedings of the International Conference on Learning Representations (ICLR), Addis Ababa, Ethiopia, 26–30 April 2020*.
12. Hu, E.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. LoRA: Low-Rank Adaptation of Large Language Models. In *Proceedings of the International Conference on Learning Representations (ICLR), Virtual, 25–29 April 2022*.
13. Ding, N.; Qin, Y.; Yang, G.; Wei, F.; Yang, Z.; Su, Y.; Hu, S.; Shi, Y.; Liu, Z.; Sun, M.; et al. Parameter-Efficient Fine-Tuning of Large-Scale Pre-Trained Language Models. *Nat. Mach. Intell.* **2023**, *5*, 220–235. [\[CrossRef\]](#)
14. Tunstall, L.; Reimers, N.; Jo, U.; Bates, L.; Korat, D.; Wasserblat, M.; Pereg, O. Efficient Few-Shot Learning Without Prompts. *arXiv* **2022**, arXiv:2209.11055 [\[CrossRef\]](#)
15. Ganaie, M.; Hu, M.; Malik, A.; Tanveer, M.; Suganthan, P. Ensemble Deep Learning: A Review. *Eng. Appl. Artif. Intell.* **2022**, *115*, 105151. [\[CrossRef\]](#)
16. Wolpert, D. Stacked Generalization. *Neural Netw.* **1992**, *5*, 241–259. [\[CrossRef\]](#)
17. Blei, D.; Ng, A.; Jordan, M. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
18. Grootendorst, M. BERTopic: Neural Topic Modeling with a Class-Based TF-IDF Procedure. *arXiv* **2022**, arXiv:2203.05794.
19. Staubitz, T.; Petrick, D.; Bauer, M.; Renz, J.; Meinel, C. Improving the Peer Assessment Experience on MOOC Platforms. In *Proceedings of the Third ACM Conference on Learning at Scale (L@S), Edinburgh, UK, 25–26 April 2016*; Association for Computing Machinery: New York, NY, USA, 2016; pp. 389–398. [\[CrossRef\]](#)
20. Reimers, N.; Gurevych, I. Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Hong Kong, China, 3–7 November 2019*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 3982–3992. [\[CrossRef\]](#)
21. Reimers, N.; Gurevych, I. Making Monolingual Sentence Embeddings Multilingual Using Knowledge Distillation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, 16–20 November 2020*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 4512–4525. [\[CrossRef\]](#)

22. Jolliffe, I.; Cadima, J. Principal Component Analysis: A Review and Recent Developments. *Philos. Trans. R. Soc. A* **2016**, *374*, 20150202. [CrossRef]
23. McInnes, L.; Healy, J.; Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv* **2018**, arXiv:1802.03426.
24. Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.t.; Rocktäschel, T.; et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Proceedings of the Advances in Neural Information Processing Systems, Virtual, 6–12 December 2020*; Curran Associates Inc.: Red Hook, NY, USA, 2020; pp. 9459–9474.
25. Asai, A.; Wu, Z.; Wang, Y.; Sil, A.; Hajishirzi, H. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. In *Proceedings of the 12th International Conference on Learning Representations (ICLR), Vienna, Austria, 7–11 May 2024*.
26. Guu, K.; Lee, K.; Tung, Z.; Pasupat, P.; Chang, M.W. REALM: Retrieval-Augmented Language Model Pre-Training. In *Proceedings of the International Conference on Machine Learning (ICML), Virtual, 13–18 July 2020*; JMLR.org: Norfolk, MA, USA, 2020; pp. 3929–3938.
27. Jeong, S.; Baek, J.; Cho, S.; Hwang, S.; Park, J. Adaptive-RAG: Learning to Adapt Retrieval-Augmented Large Language Models through Question Complexity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Mexico City, Mexico, 16–21 June 2024*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 7036–7050. [CrossRef]
28. Mialon, G.; Dessi, R.; Lomeli, M.; Nalmpantis, C.; Pasunuru, R.; Raj, K.; Riedel, S.; Khot, T.; Roller, S.; Weston, J.; et al. Augmented Language Models: A Survey. *Trans. Mach. Learn. Res.* **2023**. [CrossRef]
29. Yu, H.; Gan, A.; Zhang, K.; Tong, S.; Liu, Q.; Liu, Z. Evaluation of Retrieval-Augmented Generation: A Survey. *arXiv* **2024**, arXiv:2405.07437. [CrossRef]
30. Zhang, Y.; Jin, R.; Zhou, Z.H. Understanding Bag-of-Words Model: A Statistical Framework. *Int. J. Mach. Learn. Cybern.* **2010**, *1*, 43–52. [CrossRef]
31. Hua, Y.; Denny, P.; Wicker, J.S.; Taskova, K. The EduRABSA Dataset and the ASQE-DPT Annotation Tool. Zenodo. 2025. Available online: <https://zenodo.org/records/16935018> (accessed on 20 March 2026).
32. Mungunshagai, M. Coursera Course Reviews Dataset. Hugging Face Datasets. 2024. Available online: <https://huggingface.co/datasets/MungunshagaiT/coursera-reviews> (accessed on 20 March 2026).
33. McInnes, L.; Healy, J.; Astels, S. HDBSCAN: Hierarchical Density Based Clustering. *J. Open Source Softw.* **2017**, *2*, 205. [CrossRef]
34. Song, K.; Tan, X.; Qin, T.; Lu, J.; Liu, T.Y. MPNet: Masked and Permuted Pre-Training for Language Understanding. In *Proceedings of the Advances in Neural Information Processing Systems, Virtual, 6–12 December 2020*; Curran Associates Inc.: Red Hook, NY, USA, 2020; pp. 16857–16867.
35. Abdi, A.; Sedrakyan, G.; Veldkamp, B.; van Hillegersberg, J.; van den Berg, S. Students Feedback Analysis Model Using Deep Learning-Based Method and Linguistic Knowledge for Intelligent Educational Systems. *Soft Comput.* **2023**, *27*, 14073–14094. [CrossRef]
36. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. In *Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017*; Curran Associates Inc.: Red Hook, NY, USA, 2017; pp. 5998–6008.
37. Li, Z.; Zhang, X.; Li, Y.; Long, D.; Xie, P.; Zhang, M. Towards General Text Embeddings with Multi-Stage Contrastive Learning. *arXiv* **2023**, arXiv:2308.03281. [CrossRef]
38. Su, H.; Shi, W.; Kasai, J.; Wang, Y.; Hu, Y.; Ostendorf, M.; Yih, W.t.; Smith, N.; Zettlemoyer, L.; Yu, T. One Embedder, Any Task: Instruction-Finetuned Text Embeddings. *arXiv* **2022**, arXiv:2212.09741.
39. Es, S.; James, J.; Espinosa-Anke, L.; Schockaert, S. RAGAS: Automated Evaluation of Retrieval Augmented Generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, St. Julians, Malta, 17–22 March 2024*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 150–158. [CrossRef]
40. Conneau, A.; Kruszewski, G.; Lample, G.; Barrault, L.; Baroni, M. What you can cram into a single vector and what you can't: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL), Melbourne, Australia, 15–20 July 2018*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2018; pp. 2126–2136. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.