EduPulse: A Practical LLM-Enhanced Opinion Mining System for Vietnamese Student Feedback in Educational Platforms

Nguyen Xuan Phuc*, **Nguyen Xuan Phi***, **Nguyen Vinh Tiep**,
Dang Van Thin, **Ngan Luu-Thuy Nguyen**†
University of Information Technology-VNUHCM,
Vietnam National University, Ho Chi Minh City, Vietnam
23521213@gm.uit.edu.vn, xphi.work@gmail.com
{thindv,tiepnv,ngannlt}@uit.edu.vn

Abstract

Opinion mining from real-world student feedback presents significant practical challenges, such as handling linguistic noise (slang, teencode) and the need for scalable and maintainable systems, which are often overlooked in academic research. This paper introduces EduPulse, a practical opinion mining system designed specifically to analyze student feedback in Vietnamese. Our application[1] performs four opinion analysis tasks, including Sentiment Classification, Category-based Sentiment Classification, Suggestion Detection, and Opinion Summarization. We design the hybrid architecture that strategically balances performance, cost, and maintainability. This architecture leverages the robustness of Large Language Models (LLMs) for complex, noise-sensitive tasks as sentiment classification and suggestion detection, while employing a specialized, lightweight neural model for high-throughput, low-cost solutions. Our experiments show that applying the LLM-based approach achieves high robustness, justifying its operational cost by eliminating expensive retraining cycles. Furthermore, we demonstrate that our collaborative modular architecture significantly improves task performance (+7.6%) compared to traditional approaches, offering a practical design for industry-focused Natural Language Processing applications.

1 Introduction

Opinion Mining (OM) or Sentiment Analysis (SA) is one of the most widely used NLP applications for identifying the emotions, intentions, and attitudes based on user comments or feedbacks (Liu, 2022). Several recent studies (Wankhade et al., 2022; Mao et al., 2024; Sharma et al., 2025) comprehensively surveyed the methods, challenges, and applications of OM in mining the valuable information generated from these user comments. However, most of these works primarily focus on researching methodologies to address various technical problems within the SA field, rather than detailing their practical, real-world applications. For instance, the work by Zhang et al. (2024) presented a comprehensive evaluation investigating the capabilities of Large Language Models (LLMs) across diverse sentiment analysis tasks, ranging from simple to complex. Similarly, Zhang et al. (2025b) also explored the effectiveness of larger LLMs in addressing the challenge of labeled data scarcity in the software engineering domain. In addition, several attempts (Hellwig et al., 2025; Xu et al., 2025b) have proposed data augmentation based on the power of LLMs for various sentiment tasks.

In the education domain, mining the information embedded in student comments plays a pivotal role, particularly for comprehensive evaluation of management systems, teaching efficacy, and academic curricula (Elfeky et al., 2020). Most universities and institutions rely heavily on student feedback for quality assurance (Shaik et al., 2023a). However, the unstructured format, noise, and scalability of the student's feedback often result in manual analysis becoming an operational bottleneck. To tackle this challenge, applying artificial intelligence (AI) solutions is essential for automating the processing, extraction of insights, and comprehensive evaluation of large-scale educational feedback datasets.

For low-resource languages like Vietnamese, there has been growing interest in opinion mining (Nguyen et al., 2018a; Thin et al., 2023a). However, most studies primarily focus on evaluating proposed methods on static benchmarks, failing to address the operational fragility of these models in production. Traditional static architectures, such as fine-tuned BERT or PhoBERT

*These authors contributed equally.
†Corresponding author: ngannlt@uit.edu.vn
[1]A video demonstration is available at https://www.youtube.com/watch?v=tiWkpK-aWoI

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 5: Industry Track), pages 338–365
March 25-27, 2026

(Thin et al., 2023b; Dang et al., 2024), often struggle with out-of-distribution linguistic noisespecifically 'teencode' and evolving student slangand require expensive, time-consuming retraining cycles to maintain accuracy. This lack of adaptability creates a significant gap between high benchmark scores and actual scalability and maintainability in real-world educational environments. This gap highlights the need for a system that not only achieves high performance but also offers a robust, easy-to-update architecture for dynamic linguistic contexts.

In this paper, we introduce EduPulse, an AI-based application designed specifically to analyze the Vietnamese student feedback on educational platforms. The system is developed according to realistic requirements for ensuring training quality in universities. Specifically, EduPulse performs four main tasks: (1) **Sentiment Classification** - determining the overall sentiment polarity of student feedback across lectures, classrooms, and departments; (2) **Category-based Sentiment Classification** analyzing sentiment polarity with respect to specific aspect categories (e.g., courses, facilities, and services); (3) **Suggestion Detection** identifying and extracting suggestion-related expressions from student feedback; and (4) **Opinion Summarization** generating comprehensive analytical reports for institutional administrators. Additionally, we develop an interactive dashboard that automatically generates relevant statistics and visualizations. These features are designed to enhance the user experience of our application.

2 The EduPulse System

The EduPulse system was designed and implemented following the AI lifecycle development process (De Silva and Alahakoon, 2022), which consists of the primary stages of requirement analysis, system design, development, testing, and deployment. Among these, the requirement analysis and system design phases play the most critical roles, as they determine the alignment of system functionalities with user needs while ensuring adherence to key engineering principles such as performance, scalability, maintainability, and deployability. Based on these four principles, the AI solutions integrated into the EduPulse system are detailed in the following sections.

2.1 Requirement Analysis

This phase focuses on identifying the functional needs required to support educational quality assurance processes in Vietnamese universities. Based on interviews with institutional administrators, we determine that the system must be capable of processing large volumes of student feedback collected from multiple academic platforms. Functionally, EduPulse must classify sentiment at both the document and aspect-based levels, detect student suggestions, and generate high-quality analytical summaries that support data-driven decision-making. These requirements directly shaped the design and the selection of appropriate AI models for text understanding in Vietnamese.

Moreover, we analyze the linguistic characteristics and structural patterns present in Vietnamese student feedback, as understanding the nature of the data is essential for determining appropriate modeling solutions. We identify several key challenges associated with this type of data, as outlined below:

- **Language Evolution**: Feedback exhibits fast-changing slang and informal expressions, code-mixing or code-switching, requiring models to adapt to evolving linguistic patterns.
- **Noise and Ambiguity**: Misspellings, code-switching, and unconventional abbreviations introduce lexical noise and semantic ambiguity, complicating text pre-processing.
- **Implicit Meaning**: Many opinions are conveyed indirectly through contextual cues or rhetorical phrasing, demanding models capable of deeper semantic inference.
- **Domain-Specific Expressions**: Academic-related feedback contains specialized terminology, necessitating domain-aware representation learning.
- **Multifaceted Sentiment**: Single feedback entries often express multiple aspect-level sentiments, requiring fine-grained, aspect-based sentiment analysis.

2.2 System Design

In this section, we present our solutions for AI features in the EduPulse system. The main design principle in our solution is to balance the efficiency

of model performance, deployability, scalability and maintainability (Huyen, 2022). To develop a comprehensive solution, we conduct a thorough review of prior research for each task in the system to identify the most effective methodologies for integration (see details in Appendix A). This approach ensures that the proposed solution remains aligned with state-of-the-art advancements and reflects current progress in the field. In addition to employing AI-based methods to address the core analytical tasks, we also integrate software engineering techniques such as asynchronous and parallel programming to accelerate inference speed and enhance scalability when processing large volumes of feedback from multiple sources. The following sections describe in detail the approaches used to develop AI modules.

Sentiment Classification The purpose of this module is to classify the feedback of students to sentiment polarity in three levels: "positive", "negative", or "neutral". Previous studies (Nguyen et al., 2018a; Thin et al., 2023b; Dang et al., 2024) demonstrated the effectiveness of machine learning approaches for this task; however, the limitations are that it relies on a quality training dataset. In addition, resources to deploy these models are also a challenge in terms of system scalability. Recently, the studies of Zhang et al. (2024); Thin et al. (2024) have shown that in-context learning using large language models can achieve performance comparable to state-of-the-art methods while requiring fewer resources.

Moreover, due to their strong language understanding capabilities, LLMs can effectively address challenges present in Vietnamese student feedback, such as language evolution, noise, and ambiguity. While fine-tuned PLMs may excel on clean, static benchmarks, our hypothesis is that they fail to adapt to real-world linguistic noise (e.g., teencode) without costly retraining cycles. Therefore, we adopt an LLM-based approach as the primary solution for this task, prioritizing robustness and long-term maintainability over the raw inference speed of static models. Inspired by the work of Liu et al. (2022), we further optimize prompt engineering based on our previous dataset to enhance model performance and adaptability.

Category-based Sentiment Classification This task aims to extract aspect categories (e.g., teaching, facilities) and their corresponding sentiments from reviews (Sindhu et al., 2019; Sau et al., 2021). Addressing requirements from Section 2.1 (noisy/informal Vietnamese, aspect-level analysis, large-scale deployment), we examine three solutions:

LLM-based semantic extraction. We first use LLMs to identify predefined aspect categories and assign sentiment, experimenting with two-stage (decoupled) prompts (Jebbara and Cimiano, 2016). While interpretable, this approach shows limitations in stability and cost when processing large volumes of feedback (Polat et al., 2025).

Traditional ML pipeline on sentence embeddings. For efficiency, we design a traditional ML pipeline using fixed sentence embeddings (OpenAI text-embedding-3-large) (Wang et al., 2020), which are cached. We train conventional classifiers (e.g., XGBoost, SVM) in a two-step process: (1) multi-label aspect detection, and (2) aspect-level sentiment classification. This design is lightweight, fast, avoids online LLM calls (Ghatora et al., 2024), and is robust to informal language.

Neural two-phase model with aspect conditioning. We propose a dedicated neural architecture using the same embeddings. The *aspect detection* phase uses a multi-label Multi-Layer Perceptron (MLP) to predict aspect presence. The *aspect-conditioned sentiment* phase explicitly conditions sentiment prediction on this information: the sentence embedding is concatenated with an aspect-indicator vector (predicted at inference) and fed to a second MLP (Subbaiah and Bolla, 2024).

At inference time, the two neural phases are applied sequentially. Operating solely on fixed sentence embeddings, the system processes large streams of feedback with low latency, making it suitable for real-time monitoring. This two-phase model strikes a favourable balance: it is more efficient than LLM prompting and better aligned with the noisy nature of Vietnamese student feedback.

Suggestion Detection This module identifies and extracts constructive, actionable suggestions from student responses (Parker et al., 2024). Previous studies relied on lexical rules but failed to recognize implicit suggestions (Abdi et al., 2023; Zheng and Zhang, 2025). With in-context learning, LLMs can effectively distinguish between mere complaints and valuable suggestions (Meyer et al., 2024; SeSSler et al., 2025). Therefore, we employ LLMs to perform two tasks simultaneously: (1) **Identification**: classifying the presence

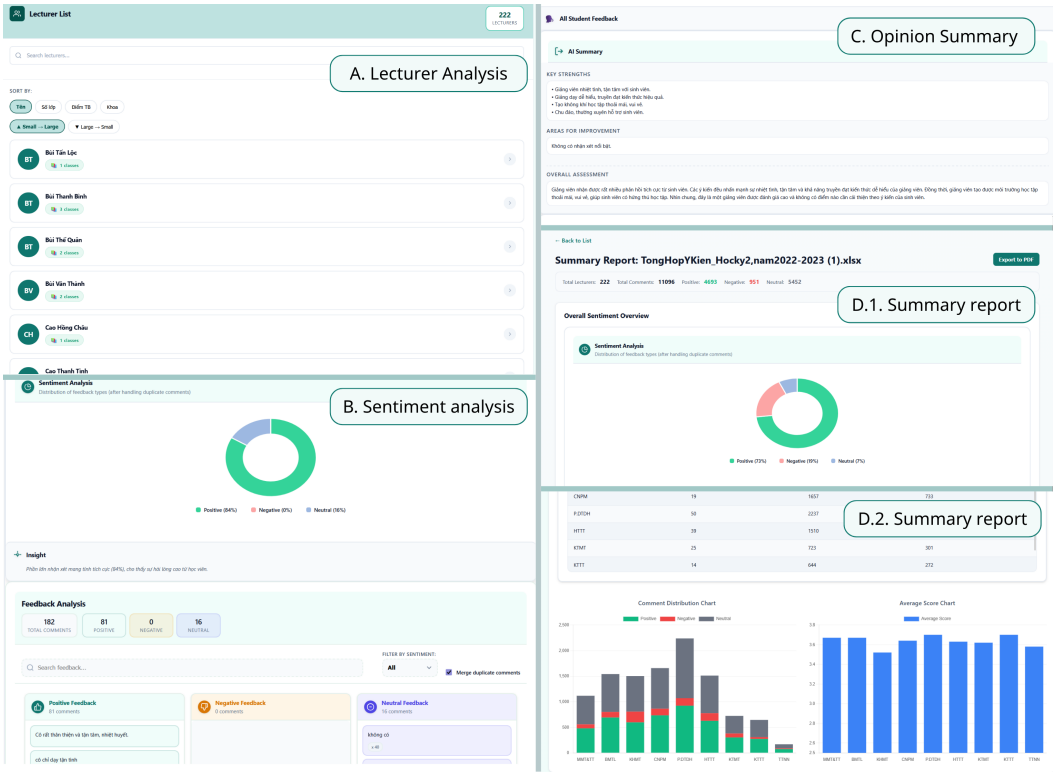

Figure 1: The user interface of the EduPulse system. A detailed breakdown of each labeled component (A-D) is provided in Appendix E.

of a suggestion, and (2) **Extraction**: retrieving its content.

Opinion Summarization This agent generates comprehensive analytical reports (Cai et al., 2025a; Zhang et al., 2025c). Since ROUGE is often unreliable for specialized topics, we prioritize informativeness and actionability (Guo et al., 2025). EduPulse employs a coordinator agent that orchestrates the operation in two steps:

Step 1: Strategic Outlining - The agent identifies recurring themes and synthesizes them into a structured outline (Zhang et al., 2025a), including main strengths and areas for improvement. Isolated opinions are disregarded to ensure objectivity.

Step 2: Report Generation - The outline is passed to an LLM to write the final summary (Sajja et al., 2024). This architecture is specifically designed to mitigate hallucinations and logical inconsistenciescommon failure cases in LLM applicationsby grounding the final report in the structured themes identified in Step 1 (Darwish et al., 2025; Kazlaris et al., 2025).

3 Experiments

In this section, we present the experimental results of our solution on existing benchmark datasets, along with the human evaluation process. The information of the evaluation benchmark dataset is described in Appendix B.

3.1 Results and Discussion

Sentiment Classification. Our evaluation for this module addresses two practical concerns: (1) performance on clean benchmarks and (2) robustness to noisy, real-world data. On the standard VSFC benchmark (Table 1), fine-tuned Vietnamese Pretrained Language Models (PLMs) like PhoBERT-base appear to be the strongest option. They achieve the highest performance with the fastest, zero-cost inference.

However, this result does not reflect the practical reality of student feedback. As hypothesized in Section 2.2, static models are 'brittle' to noise. We tested this on a challenging subset of 300 noisy comments from the Synthetic-VSFC dataset (results in Table 2, 11). The results show a stark difference: the fine-tuned PhoBERT achieved only 71.73% accuracy, struggling significantly

Table 1: Sentiment classification benchmark results on the VSFC dataset. For LLMs from this study, the result from the best-performing prompt strategy is reported.

Model	Weighted-F1	Macro-F1	Micro-F1	Time (s/1k)	Cost ($/1k)
PhoBERT-base	**0.935**	**0.828**	**0.939**	11.0	-
ViT5-base	0.922	0.821	0.924	29.0	-
Qwen 2.5 7B	0.853	0.705	0.832	96.3	$0.068
Qwen 3 32B	0.869	0.715	0.837	58.3	$0.220
Gemma 3 27B	0.906	0.765	0.892	50.2	$0.156
Llama 3.3 70B	0.916	0.787	0.911	52.5	$0.313
GPT-3.5	-	0.688	0.820	-	-
GPT-4o	-	0.689	0.811	-	-
GPT-4.1	0.881	0.723	0.866	71.9	$3.944

Table 2: Robustness Benchmark on Noisy Data (Sentiment Classification). Results averaged over 5 runs with N=300 samples per run.

Model	Accuracy	Std. Dev.
PhoBERT-base (Fine-tuned)	0.7173	ś0.0114
LLM Agent (GPT-4o)	**0.9200**	ś0.0088

with teencode and ambiguity. In contrast, the zero-shot LLM agent (GPT-4o) achieved 92.00% accuracy. While LLMs incur higher cost and latency (Table 5), this point performance gap on real-world data is decisive. The LLM's robustness validates our design choice to use it as the primary classifier. This approach justifies the inference cost by eliminating the significant, recurring operational cost and engineering effort required to constantly retrain a static PLM to keep up with language evolution.

Category-based Sentiment Classification Table 3 reveals clear performance trade-offs. Among the LLMs, GPT-4.1 demonstrates the strongest general reasoning, achieving the highest Weighted-F1 and Macro-F1 scores. However, its significant cost and latency limit its feasibility for large-scale, real-time applications. Traditional ML models like SVM offer a highly efficient, low-cost alternative but lag in overall accuracy.

Significantly, the Multi-Layer Perceptron (MLP) stands out, achieving the highest Micro-F1 score overall while operating at a minimal computational cost. This result is critical, demonstrating that for a well-defined domain task, a compact, specialized model can outperform large, general-purpose LLMs in both raw accuracy and efficiency.

This finding directly validates our choice for EduPulse. The system requires high-throughput processing of large feedback volumes, and the MLP delivers the optimal balance of accuracy, speed, and low resource usage. This design, however, remains flexible, allowing for the future integration of LLMs to handle more complex or implicit reasoning, ensuring EduPulse remains both scalable and extensible.

Suggestion Detection The suggestion detection module is evaluated on two subtasks, with methodology and dataset details in Appendix D.

Task 1: Suggestion Identification: Multiple LLMs (closed- and open-source) are tested under zero-shot and few-shot prompting, using Accuracy, Precision, Recall, and F1-score (Kojima et al., 2023). Results appear in Table 7. GPT-4o achieves the highest F1 on the VSFC dataset with examples, while Gemini 2.5 Flash Lite excels on $\mathcal{D}_{\text{UIT}}$. Closed-source models perform strongly due to recent updates and broad pre-training. Cost and latency are reported in Table 9; GPT models are costlier but offer acceptable inference speed compared to Gemini 2.5 Flash.

Task 2: Suggestion Extraction: Using the same models, token-level Precision, Recall, and F1 are computed alongside ROUGE-L for span overlap (Table 8). Exact-match metrics are low due to autoregressive prediction challenges; ROUGE-L is thus prioritized (Lin, 2004). GPT-

Table 3: ABSA benchmark on the ABSA dataset: Weighted-F1, Macro-F1, Micro-F1, inference time per 1k samples, and resource usage/cost. Bold indicates overall best (based on Micro-F1), underline indicates best within each group.

Model	Weighted-F1	Macro-F1	Micro-F1	Time (s/1k)	Cost / Resource
Large Language Models					
Gemma 3 27B	0.762	0.376	0.634	130.540	$0.106 / 1k samples
GPT-4o Mini	0.773	0.403	0.650	113.177	$0.170 / 1k samples
GPT-4o	0.791	0.405	0.673	181.084	$2.800 / 1k samples
GPT-4.1	0.827	0.437	0.728	163.268	$3.000 / 1k samples
Traditional Machine Learning Models					
Logistic Regression	0.758	0.379	0.622	0.760	CPU only
Random Forest	0.635	0.311	0.484	0.870	CPU only
LightGBM	0.760	0.379	0.625	0.528	CPU only
XGBoost	0.755	0.376	0.624	1.234	CPU only
SVM	0.808	0.405	0.693	8.212	CPU only
LSTM	-	-	0.712		CPU only
CNN	-	-	0.725	-	CPU only
BiLSTM-CNN	-	-	0.738	-	CPU only
Multi-Layer Perceptron	**0.859**	**0.705**	**0.796**	**0.022**	CPU only

4o leads with ROUGE-L lare 0.8, followed closely by Gemini 2.5 Flash in F1. However, Table 10 shows Gemini's high latency. Therefore, GPT-4o is selected as the anchor model, as it offers the best balance of high accuracy and acceptable inference speed, justifying its higher operational cost.

Opinion Summarization For the Opinion Summarization task, evaluation relies on human assessment, as automated metrics like ROUGE are notoriously unreliable for capturing the necessary qualities of an "executive-level" report, such as faithfulness and actionability. Summaries were rated on a 5-point scale (1 to 5) across four criteria: (1) Informativeness, (2) Faithfulness / Factual Consistency, (3) Conciseness, and (4) Actionability. Details on this process and the criteria are described in Appendix D.2. The mean scores for each model are reported in Table 4.

The human evaluation results in Table 4 show a clear winner: Gemini 2.5 Flash. It achieves near-perfect scores, notably receiving a perfect 5.00 for Informativeness, Faithfulness, and Actionability. This indicates that it is the most capable model for generating reports that are comprehensive, factually consistent (no hallucinations), and provide practical, decision-ready insights - all of which are critical goals for the EduPulse system.

From a deployment perspective, the trade-offs are critical for an industry application. While Gemini 2.5 Flash provides the highest quality, it also has the longest inference time. GPT-4o, while scoring well on human evaluation, is prohibitively expensive, costing many times more relative to Gemini. The comparative costlatency analysis in Table 6 further highlights these differences, showing a significant operational gap between high-end and lightweight models. This analysis suggests that for maximum-quality, non-real-time reports, Gemini 2.5 Flash is the optimal choice. However, for the system's primary, scalable, and cost-sensitive operations, Gemini 2.5 Flash Lite provides the best-in-class balance of performance, speed, and cost.

3.2 Architecture Analysis: Collaborative versus Monolithic

To validate our multi-agent design, we compare the EduPulse which uses specialized module for each subtask against a Baseline (Monolithic) architecture. This baseline uses a single, complex prompt instructing one LLM (GPT-4o) to perform all analytical tasks (Sentiment, Suggestion ID, and ABSA) simultaneously.

The detailed results, presented in the Appendix (see Tables 12, 13, and 14), reveal a clear and intentional trade-off. On one hand, the collaborative architecture (our proposed system) significantly improves task performance, achieving an average gain of +7.6% across all metrics compared to the monolithic baseline. On the other hand, this

Table 4: Human evaluation results for the Opinion Summarization task (1-5 scale). Scores reflect summaries generated by each model.

Model	Informativeness	Faithfulness	Conciseness	Actionability	Average
Qwen 2.5 7B	4.44	4.06	4.00	4.52	4.255
Llama 3.3 70B	4.70	4.48	4.46	4.60	4.560
Gemma 3 27B	4.94	4.66	4.52	4.92	4.760
GPT-4o	4.94	4.70	4.74	4.90	4.820
GPT-4o-mini	4.70	4.50	4.76	4.74	4.675
GPT-4.1	4.96	4.56	**4.88**	4.92	4.830
Gemini 2.0 Flash Lite	4.38	4.26	4.44	4.48	4.390
Gemini 2.0 Flash	4.80	4.78	4.82	4.82	4.805
Gemini 2.5 Flash Lite	4.94	4.80	4.54	5.00	4.820
Gemini 2.5 Flash	**5.00**	**5.00**	4.78	**5.00**	**4.945**

modularity comes at a computational cost.

This trade-off is a key choice for EduPulse. We choose higher accuracy and reliability over raw speed. The modular design is also much easier to maintain we can update and test each agent's prompt separately. This is essential for a real world, long-term product.

4 Limitations

Limited Benchmarking against Non-LLM Methods A notable limitation of this study is the lack of a direct performance comparison against other specialized, non-LLM-based methods for Vietnamese sentiment analysis, such as alternative fine-tuned transformer architectures (Nguyen et al., 2023). Our evaluation prioritized robustness to real-world linguistic noise over raw benchmark performance, where we demonstrated the brittleness of a standard PLM. However, a more comprehensive benchmark against other task-specific models would be necessary to fully map the trade-offs in performance, cost, and maintainability for organizations choosing between fine-tuning and LLM-based approaches (Zhou et al., 2024).

Dependency and Scalability Bottlenecks in Summarization Furthermore, the system maintains a significant dependency on LLMs for critical, high-level tasks, particularly Opinion Summarization (Xu et al., 2025a; Aly et al., 2025). While EduPulse adopts a hybrid architecture using an efficient MLP for category-based analysis, the final report generation is entirely reliant on a large, general-purpose model. This reliance creates a practical bottleneck. Our analysis shows the highest-quality models incur the highest latency, while others (like GPT-4o) are prohibitively expensive for large-scale deployment (Liu et al., 2024). This trade-off between summary quality, cost, and speed remains a key scalability challenge.

5 Future Work

While EduPulse offers a practical foundation for student feedback analysis, future work will incorporate academic performance signals, such as test scores and final grades, to contextualize individual comments. Integrating qualitative feedback with quantitative outcomes may support more robust normalization and help distinguish systematic instructional concerns from performance-driven outliers.

Second, we plan to incorporate student disposition and personality modeling by analyzing historical comment patterns. Capturing stable behavioral tendencies can enable EduPulse to support more targeted and data-driven educational interventions. In practice, this information may assist administrators in curriculum planning and scheduling decisions, for example by facilitating alignment between students and instructional styles that better match their observed learning behaviors and performance profiles.

Third, to reduce reliance on computationally expensive general-purpose LLMs for summarization, future work will explore lightweight, domain-specific models for educational report generation. This direction aims to enhance scalability while retaining the accuracy and usability observed in the current system.

Acknowledgments

This research is funded by Vietnam National University HoChiMinh City (VNU-HCM) under grant number NCM2025-26-02.

References

Asad Abdi, Gayane Sedrakyan, Bernard Veldkamp, Jos Hillegersberg, and Stéphanie van den Berg. 2023. Students feedback analysis model using deep learning-based method and linguistic knowledge for intelligent educational systems. *Soft Computing*, 27:1–22.

Farhan Aftab, Sibghat Ullah Bazai, Shah Marjan, Laila Baloch, Saad Aslam, Angela Amphawan, and Tse Kian Neo. 2023. A comprehensive survey on sentiment analysis techniques. *International Journal of Technology*, 14(6):1288–1298.

Walid Mohamed Aly, Taysir Hassan A. Soliman, and Amr Mohamed AbdelAziz. 2025. An evaluation of large language models on text summarization tasks using prompt engineering techniques. *Preprint*, arXiv:2507.05123.

L Godlin Atlas, Daniel Arockiam, Arvindhan Muthusamy, Balamurugan Balusamy, Shitharth Selvarajan, Taher Al-Shehari, and Nasser A Alsadhan. 2025. A modernized approach to sentiment analysis of product reviews using bigru and rnn based lstm deep learning models. *Scientific Reports*, 15(1):16642.

Ilya Boytsov, Vinny DeGenova, Mikhail Balyasin, Joseph Walt, Caitlin Eusden, Marie-Claire Rochat, and Margaret Pierson. 2025. End-to-end aspect-guided review summarization at scale. *Preprint*, arXiv:2509.26103.

Chang Cai, Shengxin Hong, Min Ma, Haiyue Feng, Sixuan Du, Minyang Chow, Winnie Teo, Siyuan Liu, and Xiuyi Fan. 2025a. Analyzing the teaching and learning environments through student feedback at scale: a multi-agent llms framework. *Education and Information Technologies*, 30:21815–21847.

Chang Cai, Shengxin Hong, Min Ma, Haiyue Feng, Sixuan Du, Minyang Chow, Winnie Teo, Siyuan Liu, and Xiuyi Fan. 2025b. Analyzing the teaching and learning environments through student feedback at scale: a multi-agent llms framework. *Education and Information Technologies*, 30:21815–21847.

Clayton Cohn, Nicole Hutchins, Tuan Le, and Gautam Biswas. 2024. A chain-of-thought prompting approach with llms for evaluating students formative assessment responses in science. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(21):2318223190.

Thin Van Dang, Duong Ngoc Hao, and Ngan Luu-Thuy Nguyen. 2024. A study of vietnamese sentiment classification with ensemble pre-trained language models. *Vietnam Journal of Computer Science*, 11(01):137–165.

Ahmed M. Darwish, Essam A. Rashed, and Ghada Khoriba. 2025. Mitigating llm hallucinations using a multi-agent framework. *Information*, 16(7).

Daswin De Silva and Damminda Alahakoon. 2022. An artificial intelligence life cycle: From conception to production. *Patterns*, 3(6).

Abdellah Ibrahim Mohammed Elfeky, Thouqan Saleem Yakoub Masadeh, and Marwa Yasien Helmy Elbyaly. 2020. Advance organizers in flipped classroom via e-learning management system and the promotion of integrated science process skills. *Thinking Skills and Creativity*, 35:100622.

Brittney Exline, Melanie Duffin, Brittany Harbison, Chrissa da Gomez, and David Joyner. 2025. Using sentiment analysis to investigate peer feedback by native and non-native english speakers. *Preprint*, arXiv:2507.22924.

Kathryn Fuller, Kathryn Morbitzer, Jacqueline Zeeman, Adam Persky, Amanda Savage, and Jacqueline McLaughlin. 2024. Exploring the use of chatgpt to analyze student course evaluation comments. *BMC Medical Education*, 24.

Pawanjit Singh Ghatora, Seyed Ebrahim Hosseini, Shahbaz Pervez, Muhammad Javed Iqbal, and Nabil Shaukat. 2024. Sentiment analysis of product reviews using machine learning and pre-trained llm. *Big Data and Cognitive Computing*, 8(12):199.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large language model based multi-agents: A survey of progress and challenges. *Preprint*, arXiv:2402.01680.

Yue Guo, Tal August, Gondy Leroy, Trevor Cohen, and Lucy Lu Wang. 2025. Appls: Evaluating evaluation metrics for plain language summarization. *Preprint*, arXiv:2305.14341.

Ali Hamdi, Ahmed Abdelmoneim Mazrou, and Mohamed Shaltout. 2024. Llm-sem: A sentiment-based student engagement metric using llms for e-learning platforms. In *The International Conference of Advanced Computing and Informatics*, pages 145–154. Springer.

Nils Constantin Hellwig, Jakob Fehle, and Christian Wolff. 2025. Exploring large language models for the generation of synthetic training samples for aspect-based sentiment analysis in low resource settings. *Expert Systems with Applications*, 261:125514.

Chip Huyen. 2022. *Designing machine learning systems*. " O'Reilly Media, Inc.".

Soufian Jebbara and Philipp Cimiano. 2016. Aspect-based sentiment analysis using a two-step neural network architecture. In *Semantic Web Evaluation Challenge*, pages 153–167. Springer.

Ioannis Kazlaris, Efstathios Antoniou, Konstantinos Diamantaras, and Charalampos Bratsas. 2025. From illusion to insight: A taxonomic survey of hallucination mitigation techniques in llms. *AI*, 6(10).

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large language models are zero-shot reasoners. *Preprint*, arXiv:2205.11916.

Anna Koufakou. 2024. Deep learning for opinion mining and topic classification of course reviews. *Education and Information Technologies*, 29(3):2973–2997.

Wenna Lai, Haoran Xie, Guandong Xu, and Qing Li. 2024. Rvisa: Reasoning and verification for implicit sentiment analysis. *Preprint*, arXiv:2407.02340.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Bing Liu. 2022. *Sentiment analysis and opinion mining*. Springer Nature.

Jiacheng Liu, Alisa Liu, Ximing Lu, Sean Welleck, Peter West, Ronan Le Bras, Yejin Choi, and Hannaneh Hajishirzi. 2022. Generated knowledge prompting for commonsense reasoning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3154–3169, Dublin, Ireland. Association for Computational Linguistics.

Yixin Liu, Kejian Shi, Katherine He, Longtian Ye, Alexander Fabbri, Pengfei Liu, Dragomir Radev, and Arman Cohan. 2024. On learning to summarize with large language models as references. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8647–8664, Mexico City, Mexico. Association for Computational Linguistics.

Yanying Mao, Qun Liu, and Yu Zhang. 2024. Sentiment analysis methods, applications, and challenges: A systematic literature review. *Journal of King Saud University-Computer and Information Sciences*, 36(4):102048.

Jennifer Meyer, Thorben Jansen, Ronja Schiller, Lucas W. Liebenow, Marlene Steinbach, Andrea Horbach, and Johanna Fleckenstein. 2024. Using llms to bring evidence-based feedback into the classroom: Ai-generated feedback increases secondary students text revision, motivation, and positive emotions. *Computers and Education: Artificial Intelligence*, 6:100199.

Duc Do Minh, Vinh Nguyen Van, and Thang Dam Cong. 2025. Using large language models for education managements in vietnamese with low resources. *Preprint*, arXiv:2501.15022.

Shubhangi Mohod. 2025. Ethical and societal implication of sentiment analysis using nlp in educational feedback system. *Journal of Information Systems Engineering and Management*, 10:742–749.

Dat Quoc Nguyen and Anh Tuan Nguyen. 2020. PhoBERT: Pre-trained language models for Vietnamese. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1037–1042.

Dung Ha Nguyen, Anh Thi Hoang Nguyen, and Kiet Van Nguyen. 2024. A weakly supervised data labeling framework for machine lexical normalization in vietnamese social media. *Preprint*, arXiv:2409.20467.

Kiet Van Nguyen, Duc-Vu Nguyen, Phu X. V. Nguyen, Tham T. H. Truong, and Ngan Luu-Thuy Nguyen. 2018a. Uit-vsfc: Vietnamese students feedback corpus for sentiment analysis. *2018 10th International Conference on Knowledge and Systems Engineering (KSE)*, pages 19–24.

Kiet Van Nguyen, Vu Duc Nguyen, Phu X. V. Nguyen, Tham T. H. Truong, and Ngan Luu-Thuy Nguyen. 2018b. Uit-vsfc: Vietnamese students feedback corpus for sentiment analysis. In *2018 10th International Conference on Knowledge and Systems Engineering (KSE)*, pages 19–24.

Nam Nguyen, Thang Phan, Duc-Vu Nguyen, and Kiet Nguyen. 2023. ViSoBERT: A pre-trained language model for Vietnamese social media text processing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5191–5207, Singapore. Association for Computational Linguistics.

Quy Hoang Nguyen, Minh-Van Truong Nguyen, and Kiet Van Nguyen. 2025. New benchmark dataset and fine-grained cross-modal fusion framework for vietnamese multimodal aspect-category sentiment analysis. *Multimedia Systems*, 31(1):4.

Michael J. Parker, Caitlin Anderson, Claire Stone, and YeaRim Oh. 2024. A large language model approach to educational survey feedback analysis. *International Journal of Artificial Intelligence in Education*, 35(2):444481.

Matthew E. Peters, Sebastian Ruder, and Noah A. Smith. 2019. To tune or not to tune? adapting pretrained representations to diverse tasks. *Preprint*, arXiv:1903.05987.

Fina Polat, Ilaria Tiddi, and Paul Groth. 2025. Testing prompt engineering methods for knowledge extraction from text. *Semantic Web*, 16(2):SW–243719.

Fenghua Qi, Yuxuan Gao, Meiling Wang, Tao Jiang, and Zhenhuan Li. 2024. Data mining of online teaching evaluation based on deep learning. *Mathematics*, 12(17):2692.

Ramteja Sajja, Yusuf Sermet, David Cwiertny, and Ibrahim Demir. 2024. Integrating ai and learning analytics for data-driven pedagogical decisions and personalized interventions in education. *Preprint*, arXiv:2312.09548.

Ton Nu Thi Sau, Do Phuoc Sang, and Pham Thi Thu Trang. 2021. Aspect-based sentiment analysis on students feedback in vietnamese. *TNU J. Sci. Technol.*, 226(18):48–55.

Purwo Setiawan, Arga Seta Asmara Sakti, and Dinda Safitri Ramadhani. 2025. Listening to student voices: Aspect-based sentiment analysis of academic services using bert. *Jumantara Jurnal Manajemen dan Teknologi Rekayasa*.

Kathrin SeSSler, Arne Bewersdorff, Claudia Nerdel, and Enkelejda Kasneci. 2025. Towards adaptive feedback with ai: Comparing the feedback quality of llms and teachers on experimentation protocols. *Preprint*, arXiv:2502.12842.

Thanveer Shaik, Xiaohui Tao, Christopher Dann, Haoran Xie, Yan Li, and Linda Galligan. 2023a. Sentiment analysis and opinion mining on educational data: A survey. *Natural Language Processing Journal*, 2:100003.

Thanveer Shaik, Xiaohui Tao, Christopher Dann, Haoran Xie, Yan Li, and Linda Galligan. 2023b. Sentiment analysis and opinion mining on educational data: A survey. *Natural Language Processing Journal*, 2:100003.

Neeraj Anand Sharma, ABM Shawkat Ali, and Muhammad Ashad Kabir. 2025. A review of sentiment analysis: tasks, applications, and deep learning techniques. *International journal of data science and analytics*, 19(3):351–388.

Irum Sindhu, Sher Muhammad Daudpota, Kamal Badar, Maheen Bakhtyar, Junaid Baber, and Mohammad Nurunnabi. 2019. Aspect-based opinion mining on students feedback for faculty teaching performance evaluation. *IEEE Access*, 7:108729–108741.

Palak Sood, Chengyang He, Divyanshu Gupta, Yue Ning, and Ping Wang. 2024. Understanding student sentiment on mental health support in colleges using large language models. In *2024 IEEE International Conference on Big Data (BigData)*, pages 1865–1872. IEEE.

Jiamin Su, Yibo Yan, Zhuoran Gao, Han Zhang, Xiang Liu, and Xuming Hu. 2025. Cafes: A collaborative multi-agent framework for multi-granular multimodal essay scoring. *arXiv preprint arXiv:2505.13965*.

Kalpa Subbaiah and Bharath Kumar Bolla. 2024. Aspect category learning and sentimental analysis using weakly supervised learning. *Procedia Computer Science*, 235:1246–1257.

Dang Van Thin, Duong Ngoc Hao, and Ngan Luu-Thuy Nguyen. 2024. Prompt engineering with large language models for Vietnamese sentiment classification. In *Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation*, pages 181–192, Tokyo, Japan. Tokyo University of Foreign Studies.

Van Dang Thin, Duong Ngoc Hao, and Ngan Luu-Thuy Nguyen. 2023a. A systematic literature review on vietnamese aspect-based sentiment analysis. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(8):1–28.

Van Dang Thin, Duong Ngoc Hao, and Ngan Luu-Thuy Nguyen. 2023b. Vietnamese sentiment analysis: An overview and comparative study of fine-tuning pretrained language models. *ACM transactions on Asian and low-resource language information processing*, 22(6):1–27.

Khanh Quoc Tran, Quang Phan-Minh Huynh, Oanh Thi-Hong Le, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2025. Vitasa: New benchmark and methods for vietnamese targeted aspect sentiment analysis for multiple textual domains. *Computer Speech & Language*, 93:101800.

Congcong Wang, Paul Nulty, and David Lillis. 2020. A comparative study on word embeddings in deep learning for text classification. In *Proceedings of the 4th international conference on natural language processing and information retrieval*, pages 37–46.

Peisong Wang. 2024. Student sentiment analysis and classroom feedback prediction using deep learning. *Applied Mathematics and Nonlinear Sciences*, 9(1).

Mayur Wankhade, Annavarapu Chandra Sekhara Rao, and Chaitanya Kulkarni. 2022. A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review*, 55(7):5731–5780.

Borui Xu, Yao Chen, Zeyi Wen, Weiguo Liu, and Bingsheng He. 2025a. Evaluating small language models for news summarization: Implications and factors influencing performance. *Preprint*, arXiv:2502.00641.

Hongling Xu, Yice Zhang, Qianlong Wang, and Ruifeng Xu. 2025b. DS^2-ABSA: Dual-stream data synthesis with label refinement for few-shot aspect-based sentiment analysis. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pages 15460–15478. Association for Computational Linguistics.

Mike Zhang, Amalie Pernille Dilling, Léon Gondelman, Niels Erik Ruan Lyngdorf, Euan D. Lindsay, and Johannes Bjerva. 2025a. Sefl: Enhancing educational assignment feedback with llm agents. *Preprint*, arXiv:2502.12927.

Ting Zhang, Ivana Clairine Irsan, Ferdian Thung, and David Lo. 2025b. Revisiting sentiment analysis for software engineering in the era of large language models. *ACM Transactions on Software Engineering and Methodology*, 34(3):1–30.

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Pan, and Lidong Bing. 2024. Sentiment analysis in the era of large language models: A reality check. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3881–3906. Association for Computational Linguistics.

Xueqiao Zhang, Chao Zhang, Jianwen Sun, Jun Xiao, Yi Yang, and Yawei Luo. 2025c. Eduplanner: Llm-based multiagent systems for customized and intelligent instructional design. *IEEE Trans. Learn. Technol.*, 18:416427.

Xiaofeng Zheng and Jian Zhang. 2025. The usage of a transformer based and artificial intelligence driven multidimensional feedback system in english writing instruction. *Scientific Reports*, 15.

Zixuan Zhou, Xuefei Ning, Ke Hong, Tianyu Fu, Jiaming Xu, Shiyao Li, Yuming Lou, Luning Wang, Zhihang Yuan, Xiuhong Li, Shengen Yan, Guohao Dai, Xiao-Ping Zhang, Yuhan Dong, and Yu Wang. 2024. A survey on efficient inference for large language models. *Preprint*, arXiv:2404.14294.

A Related work

A.1 Opinion Mining and Sentiment Analysis Methods

Opinion mining, or sentiment analysis, is a core NLP task for extracting opinions, emotions, and attitudes from text. Early lexicon-based methods, though interpretable, struggled with context, domain-specific language, and constructs like negation or sarcasm (Aftab et al., 2023). Machine learning models (e.g., Naive Bayes, SVM, logistic regression) improved adaptability using engineered features but required large annotated datasets (Aftab et al., 2023). More recently, deep learning models (CNNs, RNNs, LSTMs, GRUs) have captured semantic and sequential patterns automatically, enhancing performance on complex texts (Aftab et al., 2023).

Currently, transformer-based pretrained language models, such as BERT and its variants, achieve high performance in sentiment classification. Their contextualized embeddings capture subtle semantic relationships, enabling accurate predictions with minimal task-specific engineering (Atlas et al., 2025).

Aspect-based sentiment analysis (ABSA), which associates sentiment with specific entities or attributes (e.g., "teaching quality," "grading policy"), has also advanced significantly with transformer architectures. These methods jointly model aspects, context, and polarity to provide fine-grained sentiment extraction (Shaik et al., 2023b).

A.2 Vietnamese Sentiment Analysis

Aspect-Based Sentiment Analysis (ABSA) for Vietnamese has gained substantial research attention. The UIT-ABSA dataset, targeting restaurant reviews, has been extensively benchmarked with PhoBERT-based approaches. More recent work introduced ViTASA, a large-scale dataset with over 500,000 target-aspect pairs across mobile, restaurant, and hotel domains (Tran et al., 2025). Their proposed ViTASD model achieved strong macro F1-scores in these respective domains, significantly outperforming previous BERT-based methods and zero-shot LLMs including Gemma, Llama, Mistral, and Qwen (Tran et al., 2025).

In addition to text-only sentiment data, there is work on multimodal Vietnamese sentiment analysis. For example, (Nguyen et al., 2025) released the ViMACSA dataset, with nearly 4,900 textimage pairs and 14,600 fine-grained annotations in the hotel domain, and proposed a cross-modal fusion model that outperforms prior prior approaches.

These Vietnamese-specific advances highlight both the progress and the challenges in building robust, fine-grained sentiment analysis systems for Vietnamese. However, there remains relatively little work on real-world, continuous feedback systems (such as student feedback), especially in educational settings, underscoring a gap that our system aims to fill.

A.3 Large Language Models for Opinion Mining

Large Language Models (LLMs) have emerged as a promising tool for opinion mining at scale. Research investigating the use of large language models (LLMs), such as ChatGPT, for sentiment classification of student feedback has highlighted their significant potential in accurately categorizing feedback as positive, negative, or neutral (Fuller et al., 2024). A key advantage of LLMs is their capacity to interpret ambiguous comments via zero-shot and few-shot learning, with chain-of-thought prompting substantially enhancing the understanding of context-dependent feedback (Cohn et al., 2024). Comparative studies between LLMs and fine-tuned transformers (e.g., RoBERTa) for educational survey analysis have reported mixed results: while some studies show LLMs achieving substantial improvements in accuracy and F1-score, others indicate that domain-specific fine-tuned models can still outperform general-purpose LLMs in structured feedback analysis (Exline et al., 2025).

A.4 Sentiment Analysis in Educational Feedback

Educational feedback mining, analyzing student comments, course evaluations, and other qualitative data has become increasingly important for improving teaching quality, course design, and student satisfaction. Unlike general opinion mining, educational feedback often contains short, informal, and unstructured comments that include implicit sentiment, mixed languages, and domain-specific terminology, making automated analysis particularly challenging (Shaik et al., 2023b).

Early approaches applied deep learning models to large volumes of course reviews. For instance,(Koufakou, 2024) used BERT, RoBERTa, and XLNet to classify sentiment and topics, with RoBERTa achieving high accuracy. (Wang, 2024) integrated multimodal signals, combining facial expression detection with text feedback, achieving around 72% accuracy in predicting classroom sentiment.

Large Language Models (LLMs) have recently been applied to educational feedback for specialized tasks. (Sood et al., 2024) introduced the SMILECollege dataset on student feedback about mental health support and demonstrated strong LLM performance (GPT3.5, BERT) in capturing nuanced sentiment. (Hamdi et al., 2024) proposed the LLMSEM metric, which combines LLM-derived sentiment scores with platform metadata (e.g., views, likes), illustrating the potential for scalable and robust system-level monitoring of student engagement.

Aspect-level analysis has also benefited from modern pipelines. (Setiawan et al., 2025) applied a BERT-based ABSA model to academic service evaluations (e.g., staff interactions, administrative processes) and achieved 98.6% accuracy in extracting sentiment-laden aspects.(Qi et al., 2024) combined BERT and LSTM in a hybrid pipeline to classify sentiment in online teaching evaluations, also extracting association rules via Apriori, demonstrating end-to-end analysis for structured and unstructured feedback (Qi et al., 2024).

Beyond modeling, recent studies highlight deployment challenges and societal considerations. (Mohod, 2025) examined ethical and practical issues of using NLP for student feedback, including bias, privacy, and interpretability. Despite progress, many systems remain confined to small datasets or single institutions. There is limited work on scalable, cross-lingual feedback mining, particularly for Vietnamese contexts, and few studies integrate continuous, real-world deployment with actionable feedback loops for instructors and administrators.

A.5 Vietnamese NLP and Aspect-Based Sentiment Analysis

Vietnamese presents unique challenges for NLP due to its linguistic characteristics, including tone markers, syllable-based word structure, and prevalent omission of diacritics in informal text. PhoBERT (Nguyen and Nguyen, 2020), a monolingual BERT variant pre-trained on Vietnamese corpora, has emerged as the de facto standard for Vietnamese NLP tasks, consistently outperforming multilingual alternatives.

For sentiment analysis, PhoBERT-based models have achieved strong performance across multiple Vietnamese datasets. The UIT-VSFC (Vietnamese Students' Feedback Corpus), containing over 16,000 annotated sentences for sentiment and topic classification, serves as a key benchmark in educational feedback analysis (Nguyen et al., 2018a).

A.6 Multi-Agent Systems and Practical Deployment

Multi-agent architectures have recently gained traction in educational NLP for decomposing complex reasoning into coordinated specialized roles. For example, AutoFeedback (Guo et al., 2024) uses a dual-agent system where one LLM-based agent generates feedback and another evaluates and refines it to reduce hallucinations. Similarly, (Su et al., 2025) proposed a three-agent framework for multimodal essay scoring, with agents handling initial scoring, feedback pooling, and reflective refinement, demonstrating that distributed responsibilities improve evaluative consistency and robustness.

In a large-scale application, (Cai et al., 2025b) analyzed feedback from over 7,000 medical residents using agents for quantitative analysis, sentiment classification, and topic detection. By integrating multimodal data and employing specialized report-generation agents, the system produced higher-quality feedback reports with improved balance, clarity, semantic accuracy, and coherence.

These studies demonstrate a clear trend toward decomposing complex analysis into specialized,

coordinated roles. However, these approaches have primarily focused on a single task, such as feedback generation or scoring. There is a significant gap in research addressing practical, integrated pipelines that cohesively combine multiple, distinct NLP tasks (such as sentiment classification, ABSA, and suggestion detection) to provide a holistic analysis of student feedback. This gap is especially prominent for low-resource languages like Vietnamese, highlighting the need for a coordinated, modular system that can handle fine-grained interpretation at scale.

A.7 Research Gap and EduPulse Contribution

While existing research has made significant strides in automated feedback analysis, several gaps remain for real-world deployment in educational institutions, particularly for low-resource languages like Vietnamese:

- **Linguistic Dynamics**: Fine-tuned models struggle with rapid linguistic evolution (teencode, slang), requiring expensive retraining cycles that are impractical for educational institutions with limited ML expertise (Nguyen et al., 2024).
- **Implicit Reasoning**: Traditional models excel at explicit sentiment classification but often fail to detect implicit suggestions or context-dependent meanings common in student feedback (Lai et al., 2024).
- **Integration and Scalability**: Most research focuses on individual tasks (sentiment analysis or ABSA or suggestion detection) in isolation, rather than providing end-to-end pipelines that educational institutions can readily deploy (Boytsov et al., 2025).
- **Cost-Performance Trade-offs**: While LLMs show promise, limited research compares their practical deployment costs, latency, and maintenance burden against fine-tuned alternatives in educational settings (Peters et al., 2019).

EduPulse addresses these gaps by proposing an integrated, prompt-driven AI pipeline specifically designed for noisy, evolving Vietnamese student feedback (Minh et al., 2025). This system decomposes the analysis into specialized modules-Sentiment Classification, ABSA, and Suggestion Detection, which are processed in parallel before their outputs are synthesized by a final summarization agent. By leveraging LLMs as flexible, maintainable executors for these modules, the system eliminates costly retraining cycles while maintaining strong performance across all tasks. The modular architecture enables rapid adaptation to linguistic changes through simple prompt updates, offering educational institutions a practical, scalable solution that balances performance, cost, and maintainability.

B Datasets

To comprehensively evaluate the EduPulse architecture, we employ a multi-faceted evaluation strategy using both publicly available benchmark datasets and real-world student feedback data. This dual approach allows us to assess both the fundamental capabilities of our LLM-based agents on standardized tasks and their practical effectiveness in handling authentic educational feedback scenarios.

B.1 Public Benchmark Datasets

We utilize three publicly available Vietnamese sentiment analysis datasets to evaluate the foundational performance of our LLM agents on standard NLP tasks:

1. **Vietnamese Social Media Sentiment Classification (VSFC) (Nguyen et al., 2018b):** A widely-adopted benchmark dataset for Vietnamese sentiment analysis, containing **11,426** social media posts labeled with sentiment polarity. This dataset serves as the primary testbed for evaluating basic sentiment classification capabilities at the sentence level.
2. **Synthetic-VSFC[2]:** An augmented version of VSFC designed to simulate real-world linguistic challenges. This dataset, containing **8,144** samples, incorporates various forms of noise including Vietnamese teencode (slang), intentional misspellings, and colloquial expressions, enabling us to assess model robustness under challenging conditions that mirror authentic student feedback.
3. **Aspect-based Sentiment Analysis on VSFC (ABSA-VSFC):** An extension of the VSFC

[2] Synthetic-VSFC

dataset with fine-grained aspect-level annotations. Each text is labeled with identified aspects and their corresponding sentiment polarities, making it essential for evaluating the aspect-based sentiment analysis capabilities required for detailed educational feedback interpretation.

B.2 Internal Student Feedback Dataset ($\mathcal{D}_{\text{UIT}}$)

To validate the real-world applicability of **EduPulse**, we construct and utilize an internal dataset $\mathcal{D}_{\text{UIT}}$ collected from authentic student course evaluations at the University of Information Technology (UIT), VNU-HCM.

B.2.1 Data Collection and Structure

The dataset originates from official end-of-semester course evaluation surveys conducted across multiple academic terms. The raw data is stored in tabular format (Excel/CSV), where each row represents aggregated feedback for a specific course-instructor combination.

The dataset contains the following key fields:

- **Instructor and Course Metadata:** Including instructor name, faculty, course name, program type, and class code.
- **Participation Statistics:** Class size, number of participating students, number of submitted comments, and average rating.
- **Free-form Comment Fields:** Two critical unstructured text columns:
  - Positive feedback (*What you are most satisfied with about the instructors teaching activities*)
  - Negative feedback (*What you are most dissatisfied with about the instructors teaching activities*)

B.2.2 Data Preprocessing and Annotation

Given the tabular structure with multiple feedback entries per row (corresponding to different students in the same class), we perform a data transformation process to convert the raw format into a structured JSON format suitable for NLP analysis. Each individual comment is extracted and annotated with its associated metadata.

The key transformation includes:

- **Comment Extraction:** Individual comments from the positive and negative feedback columns are separated and labeled with their sentiment type (`positive` or `negative`).
- **Metadata Preservation:** Each extracted comment retains its linkage to instructor, course, class, and source information for contextual analysis.
- **Quality Filtering:** Empty responses and non-informative comments (e.g., "Không" - "None") are filtered out during preprocessing.

B.2.3 Dataset Statistics

After applying the preprocessing and quality filtering steps, we obtained the final $\mathcal{D}_{\text{UIT}}$ dataset. A total of **3,468** entries were identified as trivial or non-informative (e.g., "không có," "n/a", ".") and were subsequently discarded.

The resulting clean dataset consists of **6,407** valid, informative comments. These are categorized by their original source field as:

- **Positive Comments:** 4,854 (approx. 75.8%)
- **Negative Comments:** 1,553 (approx. 24.2%)

This distribution, heavily skewed towards positive feedback, reflects the nature of the data collection mechanism (i.e., separate fields for positive and negative remarks) rather than the overall sentiment of a single, mixed comment.

B.2.4 Dataset Characteristics and Challenges

The $\mathcal{D}_{\text{UIT}}$ dataset presents several unique challenges that distinguish it from public benchmarks:

- **Linguistic Noise:** Student feedback contains high levels of colloquial Vietnamese, teencode, misspellings, and grammatical errors that are rarely present in curated public datasets.
- **Domain Specificity:** The feedback is rich in educational terminology and context-specific references that require specialized understanding.
- **Variable Response Quality:** Comments range from single-word responses to detailed multi-sentence critiques, creating significant variance in information density.

```
{
  "source_file": "*.xlsx",
  "source_sheet": "Classes with feedback rate >= 50%",
  "lecturer": "Bui Tan Loc",
  "faculty": "Computer Science",
  "course": "Mobile Application Development",
  "class": "CSBU202.N21.KHBC",
  "type": "positive",
  "comment": "Dedicated and enthusiastic in teaching"
},
{
  "source_file": "*.xlsx",
  "source_sheet": "Classes with feedback rate >= 50%",
  "lecturer": "Bui Tan Loc",
  "faculty": "Computer Science",
  "course": "Mobile Application Development",
  "class": "CSBU202.N21.KHBC",
  "type": "negative",
  "comment": "Fails to effectively convey knowledge"
}
```


Figure 2: JSON format of the $\mathcal{D}_{\text{UIT}}$ dataset after comment extraction and metadata annotation. Each entry represents a students comment along with corresponding contextual information.

- **Response Rate Stratification:** We partition $\mathcal{D}_{\text{UIT}}$ into two subsets based on class participation rates: $\mathcal{D}_{\text{UIT}}^{\geq 50\%}$ (high response rate, more representative feedback) and $\mathcal{D}_{\text{UIT}}^{<50\%}$ (low response rate, potentially biased feedback). This stratification enables analysis of how feedback volume affects model performance and cost-efficiency.

B.2.5 Evaluation Tasks

Beyond standard sentiment classification and ABSA, the $\mathcal{D}_{\text{UIT}}$ dataset is used to evaluate additional specialized tasks:

- **Suggestion Detection:** Identifying actionable improvement recommendations embedded within student comments.
- **Multi-aspect Analysis:** Extracting multiple educational aspects (e.g., teaching methodology, course content, assessment fairness) from a single comment.
- **Cross-class Aggregation:** Synthesizing insights across multiple classes taught by the same instructor or within the same course.

C Detail on evaluation

In this appendix section, we present a comprehensive benchmark comparing our LLM-based agents against traditional fine-tuned Transformer approaches (e.g., PhoBERT) and baseline models for the key tasks in EduPulse. Our evaluation focuses on a practical trade-off: Performance (e.g., F1-score, human evaluation), Cost (API calls or compute time), and Latency (inference time per sample). We use Vietnamese-specific datasets where possible (VSFC, Synthetic Vietnamese Students' Feedback Corpus, UIT student reviews). All experiments were conducted on a standard setup. The proprietary API-based models were accessed via their respective platforms: FPT AI Cloud, Azure AI, and Google AI Studio.[3]

D Evaluation Methodology for Suggestion and Summarization Agents

D.1 Annotation Protocol for Suggestion Analysis

D.1.1 Justification for Dataset Creation

The evaluation of suggestion analysis, particularly for the Vietnamese language in an educational context, is hindered by the lack of public, gold-standard benchmark datasets. Existing NLP datasets do not cater to our specific two-part task: (1) **identifying** the presence of a suggestion in a noisy comment, and (2) **extracting** the precise, actionable suggestion spans.To rigorously evaluate our models, we developed a new, high-quality annotated corpus by applying a consistent annotation protocol to all three datasets used in our experiments: **VSFC**, **Synthetic-VSFC**, and our internal $\mathcal{D}_{UIT}$ corpus.

D.1.2 Human-in-the-Loop (HITL) Annotation Protocol

Our data was labeled using a three-step Human-in-the-Loop (HITL) process, ensuring high accuracy and consistency.

1. **Step 1: Initial Seeding via LLM** $\rightarrow$ accelerate the annotation process, we first performed a "seeding" step. A powerful Large Language Model (GPT-4) was prompted with a few-shot, chain-of-thought prompt designed to perform a first-pass analysis on the entire raw dataset. This step automatically identified potential comments containing suggestions and extracted the likely suggestion phrases.
2. **Step 2: Human Annotation (Identification Task)**
   The seeded data was then distributed to a trained panel of human annotators (comprising 3 university lecturers and 5 senior students familiar with the educational context).

[3]Provider links: FPT AI Cloud (ai.fptcloud.com), Azure AI (ai.azure.com), and Google AI Studio (aistudio.google.com).

Model	Prompt Strategy	Weighted-F1	Macro-F1	Time (s/1k)	Cost ($/1k)
Qwen 3 32B	Zero-Shot	0.827	0.675	56.5	$0.026
	Knowledge-Aided Zero-Shot	**0.869**	**0.715**	58.3	$0.220
	Zero-Shot CoT	0.830	0.682	571	$0.029
	Knowledge-Aided Zero-Shot CoT	0.867	0.713	59.2	$0.223
	Few-Shot	0.827	0.676	60.0	$0.048
	Knowledge-Aided Few-Shot	0.848	0.690	61.8	$0.242
	Few-Shot CoT	0.819	0.660	617	$0.051
	Knowledge-Aided Few-Shot CoT	0.851	0.694	63.0	$0.244
Gemma 3 27B	Zero-Shot	0.845	0.718	47.6	$0.016
	Knowledge-Aided Zero-Shot	0.904	0.762	50.5	$0.142
	Zero-Shot CoT	0.842	0.707	215.1	$0.036
	Knowledge-Aided Zero-Shot	0.900	0.756	246.1	$0.164
	Few-Shot	0.837	0.705	48.6	$0.030
	Knowledge-Aided Few-Shot	0.906	0.765	50.2	$0.156
	Few-Shot CoT	0.859	0.717	181.1	$0.047
	Knowledge-Aided Few-Shot CoT	**0.905**	**0.766**	230.4	$0.177
Llama 3.3 70B	Zero-Shot	0.839	0.714	47.3	$0.035
	Knowledge-Aided Zero-Shot	0.907	0.769	72.1	$0.292
	Zero-Shot CoT	0.837	0.701	283.3	$0.104
	Knowledge-Aided Zero-Shot CoT	0.887	0.738	520.8	$0.419
	Few-Shot	0.855	0.738	51.9	$0.063
	Knowledge-Aided Few-Shot	**0.916**	**0.787**	52.5	$0.313
	Few-Shot CoT	0.850	0.707	238.3	$0.120
	Knowledge-Aided Few-Shot CoT	0.884	0.736	531.8	$0.450
Qwen 2.5 7B	Zero-Shot	0.793	0.658	30.2	$0.007
	Knowledge-Aided Zero-Shot	0.849	0.698	32.2	$0.059
	Zero-Shot CoT	0.794	0.657	74.6	$0.013
	Knowledge-Aided Zero-Shot CoT	**0.853**	**0.705**	96.3	$0.068
	Few-Shot	0.788	0.641	36.8	$0.013
	Knowledge-Aided Few-Shot	0.841	0.688	37.6	$0.065
	Few-Shot CoT	0.792	0.647	79.7	$0.019
	Knowledge-Aided Few-Shot CoT	0.833	0.678	99.5	$0.074
GPT 4.1	Zero-Shot	0.870	0.714	70.1	$0.582
	Knowledge-Aided Zero-Shot	**0.872**	**0.719**	74.3	$3.363
	Zero-Shot CoT	0.869	0.710	114.0	$1.089
	Knowledge-Aided Zero-Shot CoT	0.870	0.718	161.7	$4.644
	Few-Shot	0.881	0.737	72.8	$0.888
	Knowledge-Aided Few-Shot	0.874	0.723	71.9	$3.942
	Few-Shot CoT	0.869	0.713	126.5	$1.526
	Knowledge-Aided Few-Shot CoT	0.870	0.720	159.1	$4.989

Table 5: Detailed ablation study of prompt strategies for sentiment classification on the VSFC dataset (N=3166). The table highlights the trade-offs between performance (Weighted-F1, Macro-F1), latency (Time), and Cost across different models and prompting techniques. Best Macro-F1 for each model is bolded. Time is measured in seconds per 1,000 samples, and cost is estimated in USD per 1,000 samples.

- **Task:** The annotators' first task was a binary classification for **Suggestion Identification**. For every single comment, they were required to assign a mandatory `"is_suggestion": true/false` label.
- **Guideline:** A comment was labeled `true` only if it contained at least one specific, actionable idea for improvement, rather than being a mere complaint (e.g., "The homework is too hard" = `false`; "The homework should include more practical examples" = `true`).
- **Outcome:** This step produced the ground truth for the classification benchmark. To ensure label quality, we measured inter-annotator agreement (IAA) on a 10% subset, achieving a Fleiss' Kappa score that indicated substantial agreement.

3. **Step 3: Human Refinement (Extraction**

Model	Avg. Gen Time (s/sample)	Avg. ROUGE-L (vs Human)	Total Gen Cost ($)
Gemini 2.0 Flash Lite	2.273	0.303	$0.0057
Gemini 2.0 Flash	3.392	0.327	$0.0087
Gemini 2.5 Flash Lite	1.960	0.299	$0.0054
Gemini 2.5 Flash	9.319	0.290	$0.0593
GPT-4o	4.268	0.329	$0.4637
GPT-4o Mini	5.206	0.335	$0.0134
GPT-4.1	3.890	0.325	$0.1819
Qwen 2.5 7B	2.030	0.311	$0.0066
Llama 3.3 70B	4.712	0.352	**$0.0054**
Gemma 3 27B	5.350	0.313	$0.0067

Table 6: Cost and Latency Analysis for Task 3: **Opinion Summarization** (50 Samples). Avg. Gen Time reflects the time taken by the model to generate one summary. Total Gen Cost reflects the cost for the generation step only.

Model	Dataset	Zero-shot				Few-shot			
		Acc.	Prec.	Recall	F1	Acc.	Prec.	Recall	F1
Qwen 2.5 7B	UIT	0.593	1.000	0.185	0.312	0.605	1.000	0.210	0.347
	VSFC	0.788	1.000	0.575	0.730	0.803	0.976	0.620	0.758
	Synthetic-VSFC	0.728	0.989	0.460	0.628	0.775	0.991	0.555	0.712
Llama 3.3 70B	UIT	0.878	0.947	0.800	0.867	0.903	0.926	0.875	0.900
	VSFC	0.848	0.836	0.865	0.850	0.818	0.775	0.895	0.831
	Synthetic-VSFC	0.878	0.842	0.930	0.884	0.860	0.803	0.955	0.872
Gemma 3 27B	UIT	0.860	0.896	0.815	0.853	0.900	0.900	0.900	0.900
	VSFC	0.830	0.777	0.925	0.845	0.815	0.740	0.970	0.840
	Synthetic-VSFC	0.918	0.907	0.930	**0.919**	0.900	0.881	0.925	0.902
Gemini 2.5 Flash Lite	UIT	0.925	0.909	0.945	**0.926**	0.928	0.901	0.960	**0.930**
	VSFC	0.800	0.736	0.935	0.824	0.778	0.719	0.910	0.804
	Synthetic-VSFC	0.883	0.828	0.965	0.891	0.870	0.819	0.950	0.880
Gemini 2.0 Flash Lite	UIT	0.855	0.955	0.745	0.837	0.865	0.951	0.770	0.851
	VSFC	0.863	0.861	0.865	0.863	0.840	0.821	0.870	0.845
	Synthetic-VSFC	0.895	0.876	0.920	0.898	0.885	0.870	0.905	0.887
Gemini 2.5 Flash	UIT	0.923	0.912	0.935	0.923	0.918	0.907	0.930	0.918
	VSFC	0.835	0.760	0.980	0.856	0.818	0.743	0.970	0.842
	Synthetic-VSFC	0.883	0.831	0.960	0.891	0.880	0.822	0.970	0.890
Gemini 2.0 Flash	UIT	0.678	0.859	0.425	0.569	0.900	0.917	0.880	0.898
	VSFC	0.915	0.888	0.950	0.918	0.845	0.772	0.980	0.863
	Synthetic-VSFC	0.900	0.994	0.805	0.890	0.910	0.932	0.885	0.908
GPT-4o Mini	UIT	0.823	0.951	0.680	0.793	0.783	0.945	0.600	0.734
	VSFC	0.878	0.917	0.830	0.871	0.860	0.914	0.795	0.850
	Synthetic-VSFC	0.895	0.895	0.895	0.895	0.898	0.925	0.865	0.894
GPT-4.1	UIT	0.848	0.943	0.740	0.829	0.853	0.967	0.730	0.832
	VSFC	0.845	0.832	0.865	0.848	0.813	0.785	0.860	0.821
	Synthetic-VSFC	0.910	0.898	0.925	0.911	0.890	0.858	0.935	0.895
GPT-4o	UIT	0.860	0.956	0.755	0.844	0.863	0.956	0.760	0.847
	VSFC	0.935	0.948	0.920	**0.934**	0.920	0.947	0.890	**0.918**
	Synthetic-VSFC	0.915	0.972	0.855	0.910	0.925	0.983	0.865	**0.920**

Table 7: Evaluation Results for Task 1: Suggestion Identification (Binary Classification)

Task)

For every comment that annotators labeled as `"is_suggestion": true`, they proceeded to the **Suggestion Extraction** task.

- **Task:** Annotators were shown the LLM-seeded extractions (from Step 1) and were tasked with meticulously refining them.
- **Guideline:** This refinement process involved: (a) **validating** correct extrac-

Model	Dataset	Zero-shot				Few-shot			
		P (Span)	R (Span)	F1 (Span)	ROUGE-L	P (Span)	R (Span)	F1 (Span)	ROUGE-L
Qwen 2.5 7B	UIT	0.011	0.009	0.010	0.592	0.041	0.045	0.043	0.693
	VSFC	0.173	0.156	0.164	0.684	0.253	0.307	0.277	0.838
	Synthetic-VSFC	0.023	0.016	0.019	0.542	0.012	0.016	0.014	0.755
Llama 3.3 70B	UIT	0.080	0.074	0.077	0.707	0.126	0.125	0.125	0.718
	VSFC	0.173	0.188	0.180	0.835	0.331	0.372	0.350	0.872
	Synthetic-VSFC	0.061	0.071	0.066	0.755	0.113	0.137	0.124	0.764
Gemma 3 27B	UIT	0.079	0.077	0.078	0.691	0.109	0.119	0.113	0.743
	VSFC	0.295	0.303	0.299	0.849	0.319	0.372	0.343	0.856
	Synthetic-VSFC	0.066	0.071	0.069	**0.795**	0.127	0.165	0.144	0.766
Gemini 2.5 Flash Lite	UIT	0.094	0.083	0.088	0.715	0.161	0.176	0.168	0.725
	VSFC	0.384	0.394	0.389	0.880	0.398	0.450	**0.422**	0.860
	Synthetic-VSFC	0.030	0.033	0.032	0.793	0.134	0.159	0.145	0.760
Gemini 2.0 Flash Lite	UIT	0.033	0.029	0.031	0.662	0.149	0.144	0.146	0.735
	VSFC	0.086	0.083	0.084	0.743	0.413	0.454	0.412	0.876
	Synthetic-VSFC	0.049	0.049	0.049	0.753	0.185	0.203	0.194	0.792
Gemini 2.5 Flash	UIT	0.190	0.189	**0.190**	**0.825**	0.255	0.263	**0.259**	**0.832**
	VSFC	0.435	0.459	**0.446**	**0.893**	0.399	0.417	0.408	0.880
	Synthetic-VSFC	0.060	0.066	0.063	0.783	0.204	0.225	**0.214**	0.802
Gemini 2.0 Flash	UIT	0.067	0.064	0.066	0.721	0.152	0.154	0.153	0.763
	VSFC	0.329	0.353	0.341	0.861	0.388	0.431	0.409	**0.884**
	Synthetic-VSFC	0.045	0.049	0.047	**0.795**	0.156	0.192	0.172	0.775
GPT-4o Mini	UIT	0.045	0.042	0.043	0.681	0.076	0.087	0.081	0.737
	VSFC	0.235	0.252	0.243	0.806	0.331	0.404	0.364	0.874
	Synthetic-VSFC	0.030	0.033	0.031	0.774	0.089	0.115	0.100	0.782
GPT-4.1	UIT	0.086	0.080	0.083	0.721	0.071	0.074	0.072	0.751
	VSFC	0.377	0.385	0.381	0.868	0.281	0.321	0.300	0.863
	Synthetic-VSFC	0.076	0.082	**0.079**	0.778	0.133	0.165	0.147	0.791
GPT-4o	UIT	0.102	0.093	0.097	0.728	0.160	0.160	0.160	0.782
	VSFC	0.399	0.381	0.390	0.820	0.398	0.450	**0.422**	**0.884**
	Synthetic-VSFC	0.059	0.060	0.060	0.791	0.175	0.198	0.186	**0.803**

Table 8: Evaluation Results for Task 2: Suggestion Extraction (Span-based)

Model	**Dataset**	**Zero-shot**			**Few-shot**		
		Avg. Time (ms)	**Total Tokens**	**Est. Cost ($)**	**Avg. Time (ms)**	**Total Tokens**	**Est. Cost ($)**
Qwen 2.5 7B	UIT	186.8	185,716	$0.021	265.2	224,494	$0.025
	VSFC	237.1	183,513	$0.021	238.6	221,370	$0.025
	Synthetic-VSFC	260.5	183,221	$0.021	256.2	221,279	$0.025
Llama 3.3 70B	UIT	191.9	171,917	$0.019	226.1	209,517	$0.023
	VSFC	201.4	169,544	$0.019	199.6	207,144	$0.023
	Synthetic-VSFC	229.1	169,750	$0.019	211.1	207,350	$0.023
Gemma 3 27B	UIT	362.3	64,141	$0.007	328.7	91,299	$0.010
	VSFC	217.9	61,433	$0.007	212.0	88,566	$0.010
	Synthetic-VSFC	324.0	62,070	$0.007	219.7	89,237	$0.010
Gemini 2.5 Flash Lite	UIT	595.3	166,542	$0.017	676.1	204,142	$0.020
	VSFC	706.3	163,833	$0.016	793.1	201,433	$0.020
	Synthetic-VSFC	655.2	164,470	$0.016	695.6	202,070	$0.020
Gemini 2.0 Flash Lite	UIT	732.1	169,637	$0.038	706.3	206,731	$0.046
	VSFC	534.1	163,172	$0.037	740.8	204,081	$0.046
	Synthetic-VSFC	685.9	167,678	$0.038	633.4	204,765	$0.046
Gemini 2.5 Flash	UIT	3722.8	166,542	$0.075	3776.8	204,677	$0.084
	VSFC	3486.1	163,833	$0.074	3666.8	201,823	$0.083
	Synthetic-VSFC	3390.7	164,470	$0.074	3529.2	202,070	$0.083
Gemini 2.0 Flash	UIT	774.2	97,237	$0.027	778.6	86,837	$0.025
	VSFC	801.9	94,086	$0.026	702.7	83,974	$0.024
	Synthetic-VSFC	826.0	95,278	$0.026	751.4	84,878	$0.024
GPT-4o Mini	UIT	570.7	177,258	$0.075	545.8	218,201	$0.091
	VSFC	576.0	174,708	$0.074	578.6	215,516	$0.091
	Synthetic-VSFC	534.7	174,881	$0.074	633.7	215,685	$0.091
GPT-4.1	UIT	616.9	177,300	$0.371	879.6	218,843	$0.450
	VSFC	591.7	174,708	$0.366	591.8	215,512	$0.448
	Synthetic-VSFC	698.7	174,881	$0.366	660.0	215,689	$0.448
GPT-4o	UIT	554.2	177,258	$0.902	578.6	218,277	$1.108
	VSFC	545.1	61,508	$0.324	563.4	89,108	$0.462
	Synthetic-VSFC	499.5	61,681	$0.324	522.8	89,281	$0.462

Table 9: Cost and Latency Analysis for Task 1: **Suggestion Identification** (400 Samples)

Model	Dataset	Zero-shot			Few-shot		
		Avg. Time (ms)	Total Tokens	Est. Cost ($)	Avg. Time (ms)	Total Tokens	Est. Cost ($)
Qwen 2.5 7B	UIT	372.8	80,728	$0.009	430.2	60,875	$0.008
	VSFC	356.6	73,918	$0.009	462.2	55,753	$0.007
	Synthetic-VSFC	270.1	73,792	$0.009	349.7	57,820	$0.007
Llama 3.3 70B	UIT	783.6	75,504	$0.009	796.2	57,295	$0.007
	VSFC	605.0	70,556	$0.008	596.1	52,595	$0.006
	Synthetic-VSFC	734.4	71,449	$0.009	683.6	53,418	$0.007
Gemma 3 27B	UIT	823.6	77,220	$0.009	703.8	58,141	$0.007
	VSFC	668.3	71,079	$0.008	580.0	53,143	$0.006
	Synthetic-VSFC	699.3	71,972	$0.008	681.5	54,273	$0.006
Gemini 2.5 Flash Lite	UIT	702.9	73,222	$0.009	678.0	55,861	$0.007
	VSFC	759.2	68,202	$0.008	666.8	50,902	$0.006
	Synthetic-VSFC	692.5	69,154	$0.008	691.9	52,127	$0.006
Gemini 2.0 Flash Lite	UIT	1238.7	74,238	$0.029	1709.4	54,127	$0.021
	VSFC	866.8	69,228	$0.027	974.9	49,953	$0.019
	Synthetic-VSFC	766.1	69,858	$0.027	776.9	50,981	$0.019
Gemini 2.5 Flash	UIT	3381.3	75,590	$0.044	3561.0	55,182	$0.030
	VSFC	2853.2	70,221	$0.041	2562.5	50,537	$0.024
	Synthetic-VSFC	2879.7	71,135	$0.041	2282.2	51,536	$0.026
Gemini 2.0 Flash	UIT	1007.0	74,375	$0.034	865.8	54,177	$0.025
	VSFC	964.6	69,396	$0.032	781.3	49,902	$0.023
	Synthetic-VSFC	999.6	70,132	$0.032	896.2	50,756	$0.023
GPT-4o Mini	UIT	885.0	76,366	$0.051	1127.0	58,081	$0.039
	VSFC	936.4	71,170	$0.048	980.8	53,127	$0.036
	Synthetic-VSFC	985.9	71,940	$0.048	917.6	53,956	$0.037
GPT-4.1	UIT	1140.4	77,690	$0.217	970.4	58,608	$0.171
	VSFC	964.9	71,482	$0.184	3745.1	53,202	$0.146
	Synthetic-VSFC	683.3	72,360	$0.190	668.6	54,126	$0.151
GPT-4o	UIT	852.5	77,545	$0.540	836.1	58,561	$0.326
	VSFC	798.2	71,571	$0.433	785.6	53,285	$0.341
	Synthetic-VSFC	808.5	72,347	$0.428	744.8	54,042	$0.324

Table 10: Cost and Latency Analysis for Task 2: **Suggestion Extraction** (200 Samples)

		PhoBERT-base (Fine-tuned)			LLM Agent (GPT-4o)		
Run	**N**	**Acc.**	**Macro-F1**	**Weighted-F1**	**Acc.**	**Macro-F1**	**Weighted-F1**
1	300	0.7267	0.48	0.77	0.9167	0.77	0.92
2	300	0.7233	0.54	0.76	0.9233	0.81	0.93
3	300	0.7267	0.50	0.76	0.9100	0.78	0.92
4	300	0.7067	0.51	0.75	0.9333	0.81	0.94
5	300	0.7033	0.51	0.76	0.9167	0.76	0.92
Average		**0.7173**	**0.508**	**0.760**	**0.9200**	**0.786**	**0.926**
Std. Dev.		0.0114	0.021	0.007	0.0088	0.024	0.009

Table 11: Detailed Performance Metrics Across 5 Experimental Runs (N=300 per run). Bold values indicate superior performance. Acc. = Accuracy.

Evaluation Aspect		Baseline (Monolithic)	EduPulse (Collaborative)	Δ	Improvement
Task Performance Metrics (N=300, gpt-4o)					
Sentiment	Macro-F1	0.823	**0.882**	+0.059	+7.2%
Suggestion ID	F1-Binary	0.667	**0.724**	+0.057	+8.5%
ABSA	Span-F1	0.339	**0.363**	+0.024	+7.1%
Computational Efficiency (Per Sample)					
Latency	Time (ms)	**1306**	2193	+887	−67.9%
Token Usage	Total Tokens	**339**	568	+229	−67.6%

Table 12: Comparative Analysis: Monolithic vs. Collaborative Agent Architecture. Bold values indicate better performance. Δ = Collaborative − Monolithic. Negative improvement percentages indicate trade-offs (higher latency/token cost).

Task	Metric	Monolithic	Collaborative	Δ
Sentiment	Macro-F1	0.823	**0.882**	+0.059
Suggestion ID	F1-Binary	0.667	**0.724**	+0.057
ABSA	Span-F1	0.339	**0.363**	+0.024
Average Gain				**+7.6%**

Table 13: Task Performance: Monolithic vs. Collaborative Agents. N=300 samples, model: gpt-4o. Δ = Collaborative − Monolithic.

Metric	Monolithic	Collaborative	Δ	Trade-off
Latency (ms)	**1306**	2193	+887	1.68× slower
Total Tokens	**339**	568	+229	1.68× more

Table 14: Computational Cost: Monolithic vs. Collaborative Agents. Per-sample average. Bold indicates better (lower) values.

tions, (b) **correcting** imprecise text spans, (c) **deleting** non-actionable complaints or vague statements that the LLM incorrectly extracted, and (d) **manually adding** any valid suggestions that the LLM had missed entirely.

- **Outcome:** This process produced the final, gold-standard `"suggestions": [,...]` array for each comment, as exemplified in our data. This array serves as the ground truth for the span-based extraction benchmark.

D.2 Human Evaluation Protocol for Opinion Summarization

D.2.1 Justification for Human Evaluation

Evaluating the quality of generated summaries is an inherently subjective task. Automated metrics (e.g., ROUGE) are notoriously unreliable as they primarily measure n-gram overlap and often fail to capture critical dimensions such as factual consistency, coverage of key topics, or the practical utility of the summary.Given that the primary goal of the EduPulse summarization agent is to produce an "executive-level report" that is both trustworthy and useful for decision-making, we employed a human evaluation protocol as the gold standard for this assessment.

D.2.2 Evaluation Setup

- **Data Sample:** We randomly selected 50 unique lecturer/course profiles from the $\mathcal{D}_{UIT}$ dataset. For each profile, we aggregated all raw positive and negative feedback and then used each candidate LLM to generate a final summary.
- **Evaluator Panel:** We recruited a panel of 5 experts (3 university lecturers and 2 quality assurance staff members) who are the target audience for such reports. Each summary was independently scored by at least 2 evaluators to ensure reliability.
- **Process:** Evaluators were presented with the complete set of raw student comments (positive and negative) alongside a single, anonymized, machine-generated summary. They were then asked to score the summary on a 5-point Likert scale (1 = Very Poor, 5 = Very Good) for each of the four criteria below.

D.2.3 Evaluation Criteria

- **Informativeness (Coverage):** How well does the summary capture all the major, recurring themes and key points from the raw feedback? A low score was given if the summary fixated on minor, isolated comments or missed significant trends (e.g., 30 students complaining about the same issue).
- **Faithfulness (Factual Consistency):** How accurately does the summary reflect the source comments? Is it free of hallucinations, exaggerations, or factual contradictions? This aligns with the agent's design goal to "not add new information". A summary stating "students loved the textbook" when the feedback was neutral would receive a very low score.
- **Conciseness (Brevity):** Is the summary brief, to the point, and free of filler language or redundancy? Does it successfully synthesize information, or does it merely list out comments? This measures its suitability as an "executive-level" report.
- **Actionability (Practicality):** This is the most critical criterion for the EduPulse system. Does the summary provide concrete, specific insights that can support pedagogical improvements and decision-making?
  - *High Actionability Example:* "Students believe the multiple-choice exam does not measure critical thinking" → (Actionable: leadership can review the exam format).

– *Low Actionability Example:* "The exam was okay" or "Students had opinions on the exam" → (Vague, not actionable).

The final scores reported in Table 4 represent the mean score averaged across all evaluators for each model and criterion.

E Detailed Interface Description

This appendix provides a detailed breakdown of the EduPulse user interface, as illustrated in Figure 1. A live demonstration of the system is also available.[4]

A. **Lecturer Analysis Dashboard:** This is the main dashboard that displays a list of all lecturers. It provides preliminary feedback metrics, the number of classes taught, and faculty affiliation for each. The interface supports filtering by various criteria and includes a search bar for locating specific lecturers. This component also features a button to generate a comprehensive overview report for all lecturers in the dataset.

B. **Sentiment Analysis:** This component displays the detailed analysis for a specific lecturer. It is the direct output of the sentiment classification module, showing the distribution of positive, negative, and neutral feedback. A key function of this module is to refine the raw data; it re-classifies feedback that may have been mislabeled in the original source files (e.g., a comment filed under "positive" that is not genuinely positive). It also groups duplicate comments to streamline the results.

C. **Opinion Summary:** This section presents the results from the suggestion detection and summarization agents. It extracts and lists actionable suggestions provided by students. It then provides a concise, AI-generated summary paragraph that synthesizes all feedback for the lecturer, structured into three categories: "Key Strengths," "Areas for Improvement," and an "Overall Assessment."

D. **Summary Report (D.1, D.2):** This component shows the aggregated report for the entire institution. When initiated, the system coordinates all AI modules (sentiment, extraction, summarization) in a parallel flow to process feedback for all lecturers. The resulting report (D.1) allows administrators to compare sentiment metrics and performance trends on a larger scale, such as between different faculties within the educational unit (D.2). The use of parallel processing is key to ensuring this large-scale analysis is completed efficiently.

[4]A video demonstration is available at: https://www.youtube.com/watch?v=tiWkpK-aWoI.

F Prompt Design

This section provides details on the prompt designs used to steer the intelligent module within the EduPulse system.

Module LLM Prompts A summary of the core prompts for each agent role is presented in Table 3. Each agent is designed to handle a distinct subtask within Vietnamese student feedback analysis, ensuring modularity and interpretability across the pipeline.

Any content enclosed in angle brackets (e.g., `<ITEMS>`, `<POSITIVE_COMMENTS>`) represents a placeholder automatically populated by the system during live annotation.

VSFC Prompts For the VSFC sentiment classification experiments in F.1, we design a set of instruction-based prompts. These include zero-shot and few-shot prompts, with or without chain-of-thought (CoT) reasoning, as well as variants augmented with a Vietnamese sentiment guideline document providing task-specific knowledge. This yields configurations such as aided-knowledge zero-shot, aided-knowledge zero-shot CoT, and others, all maintaining a consistent three-way label space (positive, negative, neutral).

ABSA Prompts For the ABSA setting, we adopt a two-stage LLM prompting pipeline in F.2 .The first stage extracts relevant aspects from a fixed ontology of Vietnamese course review categories, emphasizing semantic understanding and outputting a structured JSON list. The second stage assigns sentiment (*positive*, *negative*, or *neutral*) to each extracted aspect, producing a JSON object of `"Aspect"`, `"Sentiment"` pairs. This design decouples aspect identification from sentiment labeling while maintaining a consistent ontology and output schema across all experiments.

Agent	Prompt
Sentiment Classification Agent	You are a Vietnamese language analysis assistant. Your task is to assign sentiment labels to **each student comment** according to three categories: 1) positive, 2) negative, 3) neutral. **Rules:** Do not infer beyond the literal content. Comments such as "Không có" ("None"), "Dạ không" ("No"), "X", or empty entries ⇒ label as neutral. Return a JSON array with the fields: text, sentiment. Assign exactly one label per line. **Example Output:** [{"text": "Thầy dạy rất tốt", "sentiment": "positive"}, {"text": "Không có", "sentiment": "neutral"}, {"text": "Quá nhiều bài tập", "sentiment": "negative"}] **Input:** <ITEMS>
Aspect-Based Sentiment Analysis (ABSA) Agent	You are a senior educational analyst. Your task is to perform fine-grained Aspect-Based Sentiment Analysis on student comments, identifying specific opinions about various educational aspects. **Rules:** 1. Identify all key aspects mentioned (e.g., "lecturer", "curriculum", "facilities", "homework"). 2. Map each aspect to a general category (e.g., "teaching", "course_content", "assessment"). 3. Extract the specific opinion phrase related to that aspect. 4. Assign a sentiment ("Positive", "Negative", "Neutral", "Mixed") to that specific opinion. 5. Return a JSON array, with one object for each aspect-opinion pair found. **Example Output:** [{"aspect": "lecturer", "category": "teaching", "opinion": "teaches well but too fast", "sentiment": "Mixed"}, {"aspect": "textbook", "category": "curriculum", "opinion": "phải lên libgen tìm sách", "sentiment": "Negative"}] **Input:** <ITEMS>
Suggestion Detection Agent	You are a constructive feedback analysis assistant. Your task is to detect and extract **specific, actionable suggestions** (both explicit and implicit) from student comments. **Rules:** 1. Identify actionable improvement requests. Exclude purely descriptive or emotional comments (e.g., "bài tập khó" is a complaint, not a suggestion). 2. Classify the suggestion type as "explicit" (uses action words like "nên", "cần") or "implicit" (inferred from a strong negative comment). 3. Assign a priority ("high", "medium", "low") based on the urgency or importance implied. 4. Link the suggestion to one or more related_aspects (e.g., "homework", "lecturer", "curriculum"). 5. Return results as a JSON array with one object per suggestion. **Example Output:** [{"suggestion": "Add more practical examples to the homework", "type": "explicit", "priority": "high", "related_aspects": ["homework", "curriculum"]}, {"suggestion": "Cần giảm số lượng bài tập", "type": "explicit", "priority": "medium", "related_aspects": ["homework"]}] **Input:** <ITEMS>
Summary Generation Agent	You are an objective and insightful educational analysis expert. Your task is to read all student comments about a lecturer and produce a concise, structured summary.s **Input:** A list of positive comments and a list of negative comments. **Output:** A single JSON object with fields: "summary_positive" – bullet points (max 4) of most frequent strengths; "summary_negative" – bullet points (max 4) of most frequent improvement points; "final_summary" – a short paragraph summarizing the lecturer's overall performance. **Rules:** Focus on recurring and representative comments; ignore isolated ones. Use professional, constructive language. If no positive/negative comments exist, state "No notable comments." Do not fabricate information. **Input:** # Positive comments: <POSITIVE_COMMENTS> # Negative comments: <NEGATIVE_COMMENTS>

Figure 3: Examples of the core prompts guiding each intelligent agent in the EduPulse system.

F.1 Prompt Templates for VSFC

F.1.1 Zero-Shot Prompt

```
You are a sentiment classification system in the education domain.
Determine the sentiment of the following Vietnamese comment (positive / negative / neutral).
Only respond with one of the three labels: "positive", "negative", or "neutral".

Comment: "{text}"

Please output a JSON object of the form:
{
  "sentiment": "positive" | "negative" | "neutral"
}
```

F.1.2 Few-Shot Prompt

```
You are a sentiment classification system in the education domain.
Determine the sentiment of the following Vietnamese comment (positive / negative / neutral).
Only respond with one of the three labels: "positive", "negative", or "neutral".

Comment: "{text}"

Example Analysis:

Review: "nhiệt tình giảng dạy , gần gũi với sinh viên ."
Classification:
{
  "sentiment": "positive"
}

Review: "thời lượng học quá dài , không đảm bảo tiếp thu hiệu quả"
Classification:
{
  "sentiment": "negative"
}

Review: "không có gì đặc biệt"
Classification:
{
  "sentiment": "neutral"
}

Please output a JSON object of the form:
{
  "sentiment": "positive" | "negative" | "neutral"
}
```

F.1.3 Zero-Shot CoT Prompt

```
You are a sentiment classification system in the education domain.
Determine the sentiment of the following Vietnamese comment (positive / negative / neutral).
Only respond with one of the three labels: "positive", "negative", or "neutral".

Comment: "{text}"

Provide your reasoning, focusing on sentiment indicators, then state the classification value.

Please output a JSON object of the form:
{
  "sentiment": "positive" | "negative" | "neutral"
}
```

F.1.4 Few-Shot CoT Prompt

```
You are a sentiment classification system in the education domain.
Determine the sentiment of the following Vietnamese comment (positive / negative / neutral).
Only respond with one of the three labels: "positive", "negative", or "neutral".
Provide your reasoning, focusing on sentiment indicators, then state the classification value.

Comment: "{text}"

Example Analysis:

Review: "nhiệt tình giảng dạy , gần gũi với sinh viên ."
Classification:
{
  "sentiment": "positive"
}

Review: "thời lượng học quá dài , không đảm bảo tiếp thu hiệu quả"
Classification:
{
  "sentiment": "negative"
}

Review: "không có gì đặc biệt"
Classification:
{
  "sentiment": "neutral"
}

Please output a JSON object of the form:
{
  "sentiment": "positive" | "negative" | "neutral"
}
```

F.1.5 Rule Document for VSFC

Sentiment Detection Guidelines

This document defines linguistic and semantic rules for detecting sentiment in Vietnamese text, particularly in educational or feedback-style data.
Labels follow the numeric mapping below:
- Label 0: Negative
- Label 1: Neutral
- Label 2: Positive

Positive Sentiment

Key Patterns
- Contains complimentary or appreciative words/phrases such as:
  "hay", "tốt", "dễ hiểu", "rõ ràng", "nhiệt tình", "tận tâm", "chu đáo", "vui vẻ", "thân thiện"
- Often includes positive adverbs:
  "rất", "khá", "luôn", "hết sức"
- Mentions good teaching qualities or student experience:
  e.g., "thầy giảng bài hay", "bài giảng dễ hiểu", "nhiệt tình giảng dạy"
- Emotionally positive tone (gratitude, satisfaction, enjoyment).

Rules
1. If the text contains positive adjectives/adverbs → Label 2
2. If verbs like "giúp", "hỗ trợ", "quan tâm", "tận tình" appear → Label 2
3. If text expresses gratitude, improvement, or praise → Label 2

Negative Sentiment

Key Patterns
- Contains negation or negative adjectives/verbs such as:
  "không", "chưa", "tệ", "kém", "chán", "khó hiểu", "thiếu", "ít"
- Complaints or suggestions for improvement:
  "cần cải thiện", "nên thay đổi", "chưa tốt", "không hiệu quả"
- Often structured as:
  neutral topic + negative clause (e.g., "môn học cần cải thiện", "thầy dạy chưa ổn").
- Tone indicates frustration or deficiency.

Rules
1. Negation ("không", "chưa") + adjective/verb → Label 0
2. Suggestion or complaint terms ("nên", "cần", "mong", "hy vọng cải thiện") → Label 0
3. If focus is on problems, lack, or deficiency → Label 0

Neutral Sentiment

Key Patterns
- Descriptive, factual, or balanced tone — reports without strong emotion.
  Examples:
  - "Môn học có lý thuyết và thực hành."
  - "Sinh viên tham gia đầy đủ."
  - "Thầy giảng bài rõ nhưng hơi nhanh."
- Neutral verbs/nouns dominate:
  "có", "làm", "học", "giảng", "thực hành", "kiến thức", "môn", "bài", "lớp"
- May mix good and bad elements → balanced sentiment.
  Example: "Thầy giảng dễ hiểu nhưng bài tập hơi khó."
- Soft modifiers: "cũng", "tương đối", "bình thường", "khá ổn"
- Modal connectors: "nhưng", "tuy nhiên", "cũng được"

Rules
1. No explicit emotional markers → Label 1
2. Descriptive or factual sentences (observations, summaries) → Label 1
3. Mixed positive + negative → Label 1
4. Contains modal connectors ("tuy nhiên", "nhưng", "cũng được") → Label 1
5. Focuses on content or structure rather than feelings → Label 1

Decision Flow

IF text contains strong positive adjectives/adverbs → Label 2 (Positive)
ELSE IF contains negation or complaint terms → Label 0 (Negative)
ELSE IF factual/descriptive with no emotion → Label 1 (Neutral)
ELSE IF contains both positive and negative cues → Label 1 (Neutral)

F.2 Prompt Templates for ABSA

F.2.1 Aspect Extraction Prompt

You are an expert aspect extractor for Vietnamese course reviews.

TASK:
Given one Vietnamese review sentence (or short paragraph), identify all aspects present from this ontology:
Kỹ năng giảng dạy, Hành vi, Bài tập, Cung cấp tài liệu, Kiến thức, Kinh nghiệm, Chấm điểm, Thiết bị dạy học, Đề xuất, Chương trình học, Nói chung.

REQUIREMENTS:
- Think semantically — connect the student's opinion or evaluation with the corresponding aspect, even if the aspect word is not explicitly mentioned.
- Do NOT rely on keyword matching.
- Consider context, implied meaning, and cause-effect relations (e.g., "khó hiểu" → relates to Kỹ năng giảng dạy).
- If a sentence includes contrasts (e.g., "... nhưng ..."), extract each aspect mentioned separately.
- Each extracted aspect should represent a distinct focus of opinion.

- Only output 1 aspect type one time, do not duplicate.

Output format:
```json
{
  "Aspects": ["aspect_1", "aspect_2", ...]
}
```

Examples:
Input: "Thầy dạy dễ hiểu, nhiệt tình và cho nhiều bài tập."
Output:
```json
{"Aspects": ["Kỹ năng giảng dạy", "Hành vi", "Bài tập"]}
```

Input: "Slide còn thiếu và phòng học ồn ào."
Output:
```json
{"Aspects": ["Cung cấp tài liệu", "Thiết bị dạy học"]}
```

Input: {sentence}
Output:

F.2.2 Sentiment Classification Prompt

You are an expert in aspect-level sentiment classification for Vietnamese course reviews.

TASK:
Given a Vietnamese review sentence and a list of extracted aspects, determine the sentiment polarity for each aspect.

SENTIMENT LABELS:
- positive: expresses satisfaction, praise, or appreciation.
- negative: expresses dissatisfaction, complaint, or criticism.
- neutral: mixed, factual, or suggestion without clear emotional tone.

GUIDELINES:
- Analyze semantic meaning, not just keywords.
- Handle negations (e.g., "không tốt"), contrasts ("nhưng"), and soft tones ("hơi", "tương đối").
- For "Đề xuất": if the suggestion arises from a problem (e.g., "mong thầy chuẩn bị slide"), label negative; if it appreciates a new idea, label positive or neutral.
- Each detected aspect must have exactly one sentiment label.

Output format (valid JSON object): Must be in ```json``` block
```json
{
  "Results": [
```

```
    {"Aspect": "aspect_1", "Sentiment": "positive|negative|neutral"},
    {"Aspect": "aspect_2", "Sentiment": "positive|negative|neutral"}
  ]
}
```

Examples:
Input:
Sentence: "Thầy dạy dễ hiểu, nhiệt tình và cho nhiều bài tập."
Aspects: ["Kỹ năng giảng dạy", "Hành vi", "Bài tập"]
Output:
```json
{
  "Results": [
    {"Aspect": "Kỹ năng giảng dạy", "Sentiment": "positive"},
    {"Aspect": "Hành vi", "Sentiment": "positive"},
    {"Aspect": "Bài tập", "Sentiment": "positive"}
  ]
}
```

Input:
Sentence: {sentence}
Aspects: {aspects}
Output: (Must be in ```json``` block)