Journal of Universal Computer Science, vol. 31, no. 13 (2025), 1463-1490
submitted: 12/3/2025, accepted: 5/6/2025, appeared: 28/11/2025 CC BY 4.0

Low-Footprint NLP for Reducing Teachers' Orchestration Load in Computer-Supported Case-Based Learning Environments

Claudio Alvarez
(Facultad de Ingeniería y Ciencias Aplicadas, Universidad de los Andes, Santiago, Chile
https://orcid.org/0000-0002-5442-2140, calvarez@uandes.cl)

Andres Carvallo
(National Center for Artificial Intelligence, Santiago, Chile
https://orcid.org/0000-0002-5042-8772, afcarvallo@uc.cl)

Gustavo Zurita
(Management Control & Information Systems, Universidad de Chile, Santiago, Chile
https://orcid.org/0000-0003-0757-1247, gzurita@fen.uchile.cl)

Abstract: As student cohorts grow, real-time case-based learning discussions generate increasing volumes of textual data, intensifying the orchestration load teachers must manage. Reviewing and providing feedback on student responses promptly becomes increasingly challenging, demanding efficient methods to assist educators in selecting relevant contributions to steer classroom discussions. This study proposes a low-footprint natural language processing (NLP) approach that leverages small-scale models running on commodity hardware, avoiding the computational overhead and cost associated with large language models. Our system, integrated into EthicApp, a collaborative learning platform, employs pre-trained language models such as the Universal Sentence Encoder (USE) and Bidirectional Encoder Representations from Transformers for Spanish (BETO), along with traditional text-mining techniques like Term Frequency-Inverse Document Frequency (TF-IDF). Through expert evaluations, we found that BETO exhibited superior performance in identifying relevant student responses but required GPU acceleration. At the same time, USE provided an efficient alternative that outperformed TF-IDF and remained feasible for CPU-based execution. Additionally, the methods showed a tendency—most notably BETO—to select longer responses, which, rather than introducing selection bias, was interpreted as an indicator of deeper student engagement. No significant semantic bias was found, ensuring a fair representation of students' perspectives. Our findings suggest that low-footprint NLP can effectively reduce teacher orchestration load, enabling more targeted feedback without requiring extensive computational resources.

Keywords: Natural Language Processing, Collaborative and social computing systems and tools, Laboratory experiments
Categories: H.4, J.5, L.0, L.3, L.7, M.9
DOI: 10.3897/jucs.152864

1 Introduction

Case-based learning is a widely used educational method across various disciplines, including health sciences, business, and engineering [Evans, 16; McLean, 16]. This approach is particularly prevalent in ethics education [Hess, 18; Pierrakos, 19], where it plays a crucial role in developing ethical decision-making competencies [Kobbe, 07; Stegmann, 07], a skill set increasingly demanded in the workplace. Accreditation

agencies such as ABET, AACSB, and AAA strongly promote their inclusion in academic curricula [Franks, 13; Ocone, 13]. In addition, considering high-profile scandals stemming from unethical conduct in various countries, especially in lower-income nations where corruption is more pervasive, there is an urgent need to strengthen ethics education across university programs [Heimann, 17; Instituto-de-Ingenieros, 21].

With the support of technology, students can develop and digitally share their arguments regarding the questions and decisions that arise during case-based learning activities in ethics education [Donkin, 23]. These activities can be conducted asynchronously through forums and social platforms, enabling distributed and remote participation, or synchronously in classroom environments, enhancing interaction and collaboration among learners [Caceres, 18].

When large numbers of students generate content in real time, the workload on teachers becomes substantial, making it increasingly difficult to monitor students and provide quality feedback effectively. Managing this growing volume of student contributions is a complex task that demands continuous attention and rapid decision-making. This challenge is encapsulated in classroom orchestration [Dillenbourg, 11], which refers to how teachers manage classroom activities across multiple levels—individual, group, and whole-class—while operating under real-time constraints. As orchestration involves overseeing activities across different social planes and assessing student work within strict time and curriculum constraints, the strain it places on teachers is commonly referred to as *orchestration load* [Amarasinghe, 21]. To ensure effective instruction, it is crucial to develop strategies and tools that help maintain this load manageable, preventing it from becoming an obstacle to teachers' pedagogical efficacy and well-being.

Real-time discussion activities can generate extensive amounts of argumentative text from students. As the size of the cohort increases, teachers face a growing orchestration load—the challenge of managing and processing diverse student outputs in real time—which complicates the manual review process and hinders their ability to provide students with detailed, timely feedback in classroom discussions, or at the individual level, while ensuring a broad range of approaches to the case.

Currently, large language models (LLMs) offer powerful capabilities for content analysis [Wang, 24], which can be used to automate the tasks involved in filtering and selecting the most relevant student-generated content from real-time activities. By leveraging these models, teachers can be presented with key insights and diverse perspectives without manually sifting through extensive student input. However, these large-scale models typically require expensive hardware for self-hosted deployments or necessitate subscriptions to commercial APIs, which can represent significant barriers for educational institutions [Patil, 24]. This is particularly problematic in lower-income countries, where educational opportunities are more constrained, and adopting cost-effective technology could have a transformative impact.

In this research, we explore the use of small-scale pre-trained natural language processing models that can run efficiently on commodity hardware, without even requiring a Neural Processing Unit (NPU) or Graphical Processing Unit (GPU). Specifically, we employ the Universal Sentence Encoder (USE) [Cer, 18] and BETO, a Spanish-language variant of Bidirectional Encoder Representations from Transformers (BERT) [Cañete, 20], as our primary deep learning-based approaches. Additionally, we include Term Frequency-Inverse Document Frequency (TF-IDF)

[Mohit, 14] as a lightweight text mining method as the baseline to comparatively assess neural models in automatically selecting student-generated content based on relevance. By integrating these methods into a dedicated application developed by the present authors, called EthicApp [Alvarez, 25a], our goal is to assist teachers in efficiently identifying key student contributions, reducing the orchestration load associated with manually reviewing large volumes of real-time textual responses in case-based learning activities. Based on these considerations, the study addresses the following research questions:

- RQ_1: Which variant of the automated qualitative content analysis method proposed in this research yields the best results in selecting relevant students' responses in Spanish to questions concerning a case in ethics? Variants include M_1 based on TF-IDF, M_2 based on BETO, M_3 based on USE, and M_4, an ensemble variant combining students' top responses selected by all three previous methods.
- RQ_2: Do the response selections generated by these methods introduce semantic bias concerning students' alignment with one way or another of deciding on the ethical dilemma they are asked about? Additionally, is there a bias in these methods when selecting longer answers?
- RQ_3: How do the four variants of the automated content analysis method compare regarding memory, CPU, and GPU resource utilization?

2 Related Work

2.1 Ethics and its Teaching through the Case-Based Learning Method

Ethics, defined as the study of moral judgment and rules of conduct, is crucial for discerning right and wrong [Felton, 05; Hess, 18]. Given its relevance, ethics is increasingly incorporated into higher education curricula across various fields due to frequent reports of unethical behaviour in industries and public sectors globally [Patel, 15; Zunger, 18]. Accreditation bodies like AACSB, ABET, and AAA mandate ethics in academic programs, underscoring its importance as a core component of educational standards [Apostolou, 13; Chambers, 15; Rafinda, 19].

However, implementing ethics education is challenging due to epistemological, methodological, and pedagogical differences in perceptions between students and teachers. Traditional ethics education often limits student participation, and students in scientific and technological fields may find qualitative and subjective aspects of ethics alien [Apostolou, 13; Johnson, 12]. To address these issues and enhance practical ethics education, innovative methods like using case studies, discussions, and exposure to ethical codes are employed to foster empathy and ethical reasoning [Hess, 18].

In response to these educational needs, we introduced the EthicApp tool to promote active learning in ethics through a three-phase process that enhances students' ethical reasoning and judgment in collaborative settings [Alvarez, 25a]. This approach aims to boost student engagement, socialization, and communication skills necessary for ethical decision-making.

Case-Based Learning (CBL) is a pedagogical approach that leverages real or hypothetical cases to enhance analysis, reflection, problem-solving, and decision-making skills [Johnson, 12]. This method involves presenting students with real or

simulated scenarios where they analyse, discuss, and resolve issues using their accumulated knowledge and skills, thereby gaining practical expertise applicable to real-world challenges [Kolodner, 13]. CBL aims to foster critical thinking and decision-making by applying theoretical knowledge to practical scenarios underpinned by the principles of social constructivism [Perkins, 91].

Empirical research supports CBL's effectiveness across various fields, including business, law, medicine, and e-learning, highlighting its role in enhancing decision-making and problem-solving abilities [Angeli, 4; Choi, 09; Merseth, 92; Shulman, 92]. With advancements in web technologies and Computer-Supported Collaborative Learning (CSCL), CBL can now be conducted online, offering students opportunities to tackle professional challenges through digital platforms [Li, 19; The, 19; Wang, 21; Zeng, 10]. Despite its benefits, the application of CBL in ethics education via CSCL still needs to be explored.

2.2 Instructional Design of Case-Based Learning Activities Supported by EthicApp

According to the existing literature, the case-based learning method is the most common and widely used field of ethics education [Hess, 18; Johnson, 12]. Case-based learning involves analysing problems, developing inferences with limited information, making decisions in an ambiguous context, simulating real-world scenarios, and promoting critical thinking [Kim, 6]. Based on the analysis of literature in the field of ethics education, the design principles of EthicApp [Alvarez, 25a] are the following: a) it can be integrated into ethics education courses with methodologies based on the analysis of ethical cases, b) it makes it easier for teachers to access their students' arguments; c) makes it easier for the teacher to monitor the tasks that students perform; d) facilitates the analysis of the ethical case, through student responses to dilemmatic questions; e) maintains anonymous collaborative participation of students, so that they can express their opinions and make authentic ethical judgments while decreasing their conflict anxiety; and f) facilitates the teacher's management in monitoring the responses generated by the students, through the use of dashboards.

EthicApp's instructional design includes four phases: Individual Appraisal, Appraisal Sharing, Collaborative Discussion, and a Whole-Class Discussion. Initially, students individually read an ethical case and answer three dilemmatic questions, providing detailed arguments, a process taking 5 to 8 minutes. Next, in the appraisal-sharing phase, students are grouped into small teams based on the diversity of their initial responses. Here, they anonymously review and can modify their answers after considering their peers' responses, lasting 10 to 12 minutes.

The collaborative discussion phase keeps students in the same groups to anonymously discuss and reach a consensus on the questions via chat, taking 10 to 15 minutes. Finally, the whole-class discussion phase led by the teacher allows students to openly share their views on the questions, arguments, and ethical cases discussed. This phase helps students understand the virtues or implications of potential resolutions and develop ethical frameworks applicable in their future professional or personal contexts.

2.3 Content Analysis in Educational Applications with Large Language Models

Large Language Models (LLMs) in Natural Language Processing (NLP) are transforming a wide variety of domains, including education, by offering diverse applications such as curriculum development, content generation, advanced tutoring capabilities, and discourse analysis [Kasneci, 23; Wang, 24].

Model	Parameters	Checkpoint Size	Training Tokens	Training Time
GPT-3 [Narayanan, 21]	175B	Undisclosed	300B	34 days
BLOOM [Le Scao, 23]	176B	330 GB	350B	3.5 months
LLaMA [Touvron, 23]	7B, 13B, 33B and 65B	220 TB (complete set of checkpoints)	1T, 1.0T, 1.4T, 1.4T	Up to 21 days
Alpaca [Taori, 23]	7B	27.0 GB	1T	1.5 hours
BERT [You, 19]	110M	1.3-4GB	3.3B (word corpus)	1.27 hours
USE [Cer, 18]	257M	0.5-0.9 GB	n/a	n/a

M: Million, B: Billion, T: Trillion

Table 1: LLM parameters, training tokens, and training time

Despite their transformative potential, LLMs require substantial computational resources, leading to high energy use and environmental concerns [Schwartz, 20]. The cost of training these models also increases with their size, adding financial burdens, particularly in the context of renewable energy costs, which are high in many low- and middle-income countries [Le Scao, 23; Schwartz, 20].

Various LLMs are available, ranging from commercial models like GPT to open-source options like LLaMA, Alpaca, BERT, and USE. Table 1 shows well-known LLMs of different sizes regarding parameters, training tokens, and training time. These are crucial for educational tools in developing countries, where financial constraints limit access to commercial LLMs for research and practice. Adopting pre-trained open-source models, which can be fine-tuned for specific tasks without extensive training, offers a practical solution. Word and sentence embeddings from models like BERT and USE capture semantic meanings applicable across various fields. This approach allows educational institutions to utilize these models efficiently, even for basic tasks like word or sentence embedding.

Table 2 compiles research on automatic content analysis in educational technology from key AI-focused educational sources, including Artificial Intelligence in Education (AIED), Learning Analytics and Knowledge (LAK), the European Conference on Technology Enhanced Learning (EC-TEL), the International Journal of Artificial Intelligence in Education, and the arXiv.org e-Print archive. This research mainly utilizes BERT and USE for classification, recommendation, and automation of teachers' feedback.

Publication	Application	CL	Content Analysis Method		RT	Main Results
			LM	SM		
[Mello, 21]	Classification of student essays coherence based on Rhetorical Structure Theory (RST) categories.	E		✓ (TF-IDF, LIWC, Coh-Metrix)		XGBoost Classifier, achieved performance comparable to human annotation and outperformed previous models.
[Dood, 22]	Comment classification depending on student peer-reviews and requirements in courses.	E	✓ (BERT)			Proposed model exhibits high accuracy across different peer review prompts, assignments, and courses.
[Benedict, 22]	Recommendation system of courses depending on current student requirements and previous student experiences.	E	✓ (USE)			Participants rated 70% of recommended solutions as useful.
[Mello, 22]	Tag recommendation algorithm to support teachers in assessing class open-ended questions.	E	✓ (BERT)			Tag recommendation algorithm outperforms similarity approaches for automatic statement assignation and provides interpretable predictions.
[Jensen, 21]	Automatization of teacher's feedback to students.	E	✓ (BERT)			Finetuned BERT outperforms other approaches in providing automatic feedback to students.
[Alvarez, 21]	Real-time dashboard with students' ethics course responses on a cluster scatter plot using BETO, topic modelling and ranking of the most relevant.	S	✓ (BETO)		✓	More than 80% of comment selections were found valuable, according to experts' analysis.
[Geden, 21]	Method to predict student learning outcomes using free text from forums.	E	✓ (ELMO, GLOVE)			ELMo word embeddings outperforms other language models in terms of accuracy.
[Balyan, 20]	NLP tools to identify linguistic features predictive of text difficulty.	E		✓ (FKGL+)		Deeper features of language and the usage of hierarchical machine learning techniques outperforms other approaches for predicting text difficulty.
[Altamirano, 20]	Clustering method that summarizes topics discussed in the class.	S		✓ (LDA Topic Modelling)		Clustering techniques allow teachers to have an idea of what is being discussed in a classroom.
[Wu, 22]	Framework for automatic essay scoring.	E	✓ (BERT)			Proposed model outperforms previous state-of-the-art methods in identifying key topical sentences in argumentative essays.

Publication	Application	CL	Content Analysis Method		RT	Main Results
			LM	SM		
[Mageira, 22]	AI chatbot (AsasaraBot) for teaching cultural content and foreign languages in secondary education (English/French)	E, F	✓ (NLU pretrained/ custom models on Snatchbot platform)			The AI chatbot supports Content and Language Integrated Learning (CLIL), fostering interactive learning and personalized dialogue. Evaluation shows effectiveness in teaching cultural content and moderate success in language learning. High usability and engagement reported.
[Moore, 22]	Assessing the Quality of Student-Generated Short Answer Questions Using GPT-3	E	✓ (GPT-3)			The authors investigated students' ability to generate short answer questions in online chemistry courses and the potential of using GPT-3 for quality assessment.
[Uchiyama, 23]	Chatbot using GPT-3.5-turbo to provide immediate feedback in flipped classroom preparation learning	E,J	✓(GPT-3.5-turbo)			The system enhances flipped classroom learning by providing immediate feedback aligned with video subtitles, reducing students' anxiety and improving engagement and question frequency.
[Li, 24]	Automatic dialogue analysis to promote high-quality dialogue in the classroom.	E	✓ (BERT, ERNIE & GPT)			The AI-supported framework for classroom dialogue analysis enhances discourse quality by increasing interactive features like Questions and Feedback while reducing Statements.
[Zhang, 24]	A multi-agent architecture capable of extracting meaning, generating responses, and analysing interaction content to achieve pedagogical objectives. Categorization and quantification of the types of verbal interactions occurring between agents and users in real-time. Analysis of educational material.	E	✓ (GLM-4)		✓	The multi-agent architecture powered by GLM-4 simulates complex teacher-student interactions, fostering an immersive and collaborative educational environment. Active student engagement improved knowledge retention and conceptual understanding, while the system facilitated deep discussions, critical questions, and constructive feedback.
[Mushtaq, 25]	Multi-agent system using LLMs (e.g., GPT-4) for evaluating and	E	✓ (Multi-Agent Archi-		✓	The MAS improves alignment with faculty evaluations (89%

Publication	Application	CL	Content Analysis Method		RT	Main Results
			LM	SM		
	guiding senior design projects in engineering.		tecture with GPT-4 and custom evaluation metrics)			accuracy), fosters interdisciplinary collaboration, and enhances student learning through detailed and contextual feedback. The framework integrates real-time guidance and emphasizes ethical considerations.

Table 2: Summary of literature review of content-based methods in education applications

Only recently, the non-public nature of most LLMs limited the broader research community's ability to develop them [Le Scao, 23]. This limitation led to a focus on English-language text training, with few models trained in other languages like Chinese [Wang, 21; Zeng, 10] and Korean [Kim, 21]. This contributed to a research gap and restricted access to educational technologies for non-English speakers [Kasneci, 23]. Addressing this inequity, efforts are increasing to support multilingualism within AI, though progress toward comprehensive multilingual fairness remains early.

For Spanish, widely spoken in the Global South, the BETO [Cañete, 20], a Spanish-trained version of BERT, and the Universal Sentence Encoder (USE) [Cer, 18], suitable for multilingual applications, are notable developments. Araujo et al. [Araujo, 22] found that BETO performs well on various NLP tasks. Additionally, the more computationally efficient DistillBETO [Cañete, 22], though slightly less effective in tasks like text classification than BETO and USE, reflects advancements in developing Spanish language models. Due to their efficiency and minor computational demands, both BETO and USE models are practical for real-time applications in educational technologies, such as synchronous CSCL environments.

A recent systematic review on artificial intelligence in classroom discourse (AICD) by [Wang, 24] analysed 68 studies conducted between 2012 and 2022, revealing a predominant focus on STEM-related subjects, with language learning and literature being the most studied in the arts. The review highlights two key challenges: first, the lack of practical application of AI-based classroom discourse models, as only 18 studies explored their impact on education, and second, the limited success rate, with only 25% of studies reporting positive outcomes. This underscores the need for a comprehensive evaluation to understand better the factors contributing to AI's success or failure in educational contexts.

The review also highlights a methodological shift from traditional machine learning (TML) methods, such as random forests and logistic regression, to deep learning algorithms like BERT, RNNs, and CNNs, which have gained popularity since 2018. However, the interpretability of these models remains a significant challenge, as their "black box" nature may undermine user trust and lead to inappropriate educational adjustments. Wang et al. emphasize the importance of prioritizing accountability and transparency in AICD applications, advocating for clear explanations to accompany AI-driven analyses of classroom discourse.

The use of LLMs in educational settings involves analysing varied content, with the challenge being the scarcity of text corpora for fine-tuning these models for relevance selection. It is crucial to assess if open-source models like BETO or USE can accurately select relevant student content in Spanish for real-time teacher dashboards—a significant research point. If practical, these models could provide a cost-effective solution for educational applications in developing countries with Spanish-speaking populations, without extensive training.

Our primary focus is the automated analysis of student responses to ethical cases in Spanish, emphasizing the importance of data privacy and the need for models to operate without bias. Due to privacy concerns, using specific models like GPT in public clouds (e.g., OpenAI) may not be feasible. Moreover, ensuring these models operate without biases is critical, as educators depend on dashboards that accurately reflect the original corpus distribution.

2.4 Relevant content selection based on word and sentence embeddings

This study presents an automated method to select student responses in ethics education based on relevance, focusing on two criteria. First, the 'identification criterion' is an awareness criterion that assesses whether students recognize the ethical dilemma or conflict of interest, reflecting their understanding of ethical norms and the tensions in the case. Second, the 'decision criterion' evaluates the students' ethical decisions, which is crucial for fostering decision-making skills in professional ethics. Responses directly addressing the ethical questions and context of the case are considered more relevant, as they are likely to discuss facts, actions, and behaviours pertinent to the case [Leitão, 00]. This approach aligns with enhancing interventions with the EthicApp tool. We therefore propose a heuristic for content selection: 'Student responses that are more similar to the case text and the question posed are likely to be more relevant to the teacher than those with minimal similarity to the case and question texts.'

Cosine similarity (see Equation 1) is widely used in NLP to measure similarity. This study hypothesizes that the closer the cosine similarity of students' response embeddings (generated by language models like BETO and USE) to the embeddings of the question and case, the more relevant the content according to the specified criteria. Therefore, students' responses can be ranked based on their cosine similarity to the question and case embeddings. Compared to computing word embeddings, this vector operation is computationally efficient and can be performed on standard PC hardware. Embeddings can be cached for convenience, allowing this content selection method to operate with minimal resource requirements on commodity hardware.

$$cosine\ similarity = \frac{A \cdot B}{||A||\ ||B||}$$

Equation 1. Calculation of cosine similarity scores between two n-dimensional vectors A and B.

We suggest using the TF-IDF statistical method to calculate embeddings as a baseline for benchmarking the effectiveness of the language models BETO and USE in our research. TF-IDF requires less computational power compared to these models.

Furthermore, an ensemble method combining TF-IDF, USE, and BETO rankings might improve results by leveraging each method's strengths. TF-IDF emphasizes syntactic similarity, BETO provides context-sensitive semantic analysis due to its bidirectional nature, and USE focuses on sentence-level embeddings, capturing the overall sentence meaning.

3 Proposed Content Analysis Method

The content analysis method proposed in this paper uses NLP methods to compute syntactic and semantic similarity among texts of students' responses, questions posed, and the case text. Operationally, as shown in *Figure* 1, the question related to the case, the text of students' responses, and the full text of the case are represented as vectors (V_q, V_r, V_c) using a text mining procedure or language model. Variants of the model include M_1 based on TF-IDF, M_2 based on BETO, and M_3 based on USE.

Cosine similarity between query vectors V_q and response vectors V_r is computed, sorting the responses by similarity, and retaining the top M, with M=100 determined to yield optimal results for this study's application, given the dataset.

Subsequently, the similarity between a context vector V_c and each of the top K response vectors V_r is calculated and sorted, and the top N responses are selected based on user-defined needs, with $N \leq K$.

The ensemble method (M_4) depicted in Figure 2 integrates results from TF-IDF, BETO, and USE models to enhance the selection of top student responses. Each text (questions, responses, case texts) is vectorized using these three models, generating multiple representations. Cosine similarity is then calculated across these representations, consolidated, and sorted by score. Duplicate responses are excluded, and indices for the top K vectors are retained. V_r vectors from the top K are ranked by similarity with V_c, with response texts mapped to their vectors. The client application determines how many of the top N responses to retrieve.

This method was initially prototyped on Google Colab using Python and TensorFlow and later implemented on a local PC for testing.

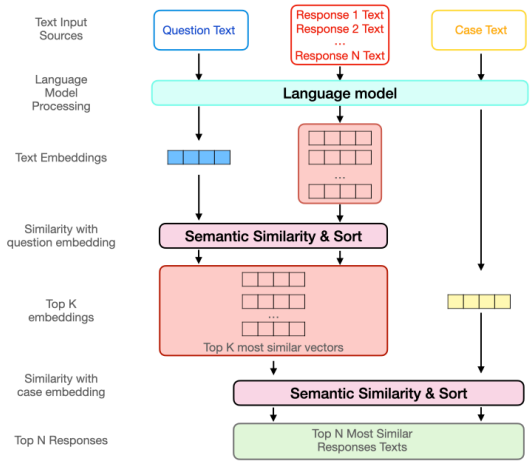

Figure 1: Retrieval of the most relevant responses using individual language models

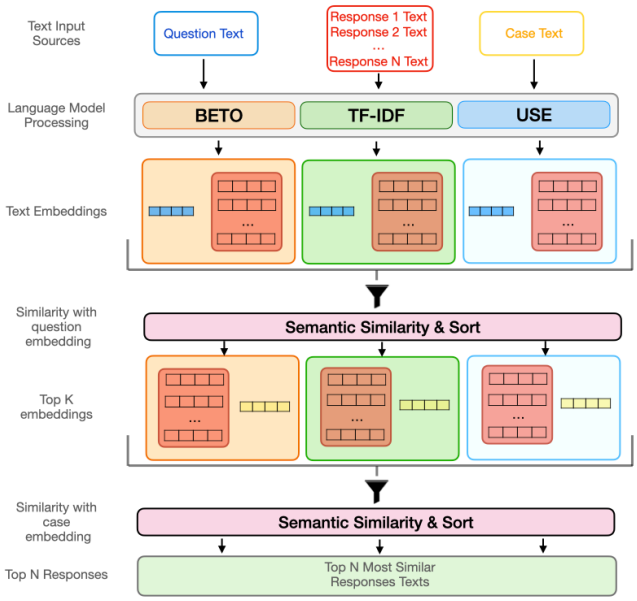

Figure 2: The ensemble variant M_4 combines the top student responses selected by the individual methods M_1 (TF-IDF), M_2 (BETO), and M_3 (USE), based on cosine similarity scores across their respective vector representations

4 Method Evaluation

4.1 Ethical Case Description and Data Set

This study employed a dataset from EthicApp activities conducted at University of Chile, involving 520 first-semester freshmen organized into 19 class groups with an average size of 30 students per group (min = 14, q_1 = 15, median = 21, q_3 = 34, max = 102). The activities centred on the ethical case of 'Laura,' an engineer navigating multiple dilemmas. Laura quit her job because of technical mistakes she made, even though she knew herself competent and could have made up for the errors. She later found a job with a higher level of responsibility. In this context, a second ethical dilemma arose about her dedication of time to family versus work, considering that she wanted to strengthen her relationship with her children and visit her aging father with health struggles.

In contrast, work offered her continual attractive challenges. Finally, based on the latter, a third dilemma arose, since in her most recent job she led the development of an offer from her company to a proposal for a very ambitious project, which meant a vital career advancement, but with a negative environmental impact that could affect the quality of citizens and ecosystems in a specific community. Her colleagues had resigned from the company because they disagreed with working on such a project.

Students responded to questions about this case quantitatively on a 7-point Semantic Differential Scale (SDS) and qualitatively in textual form, as detailed in Table 3. These responses form the basis of the database utilized in the current study, which comprises 1,102 non-empty student responses to questions 1 and 2, as shown in Table 4.

Question	Question Text	Semantic Differential Scale (SDS)*
Q_1	Is it appropriate for Laura to gradually dedicate more time to work and her professional development than to family and the other dimensions of her life?	1-7 values Left pole (1): Laura should reconsider her dedication. Right pole (7): Laura should maintain her dedication.
Q_2	Regarding the engineers who quit the project due to the impact generated, what decision seems more correct?	1-7 values Left pole (1): Quit the project. Right pole (7): Stay in the project.

*Each question required a student's written justification for the value selected on the SDS.

Table 3: Questions utilized in the EthicApp activity

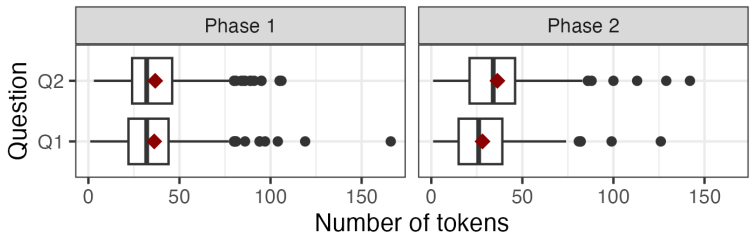

Figure 3: Distributions of the token length of students' responses per phase and question

Phase	Question	N	Token Length			
			M	***SD***	**min**	**max**
1	Q_1	257	36.1	21.0	1	166
	Q_2	371	36.7	18.8	3	106
2	Q_1	185	28.1	20.2	1	126
	Q_2	289	36.28	21.3	1	142
Total responses		1102				
Per question	Q_1	442				
	Q_2	660				

Table 4: Descriptive statistics of students' responses in the data set used in this study

Students produced the considered responses during the initial two phases of the EthicApp activity: the Individual Appraisal and the Appraisal Sharing phases previously described. The authors have noticed a trend among students to retain their original written response from the first phase when progressing to the second phase, often penning statements like, 'I maintain my judgment from the previous phase.' Consequently, the most detailed student content is typically found in the activity's early stages, particularly in the first phase. Figure 3 illustrates the distribution of student response lengths, measured in tokens, and includes stop words and punctuation. Descriptive statistics regarding response length, also in tokens, can be found in Table 4. The 'Laura' case text, written in Spanish, is 550 words long.

4.2 Expert evaluation of method efficacy

A two-round expert evaluation was conducted to compare the efficacy of the different method variants in selecting students' most relevant responses to questions concerning a case. The evaluation involved 11 experts, including the authors of this article, ethics professors, and research assistants.

The criteria for expert evaluation were twofold. Firstly, it included the 'identification' criterion, a response to question Q_j complies with this criterion if it identifies the dilemma or relevant ethical conflict. On the other hand, an answer is worse under this criterion if it partially or entirely fails to recognize the dilemma that needs to be resolved. The second criterion adopted for the expert evaluation was 'decision'. A response is better based on this criterion if it determines the best decision alternative, including explaining its arguments. Both requirements were inspired by the Ethical Reasoning Value Rubric [AAC&U, n.d.], including its categories 'Ethical Issue Recognition' and 'Evaluation of Different Ethical Perspectives.'

For each method, M_i proposed, with i in {1,2,3,4}, and case question Q_j, with j in {1,2}, the set with the five most relevant answers, i.e., B_{ij}, was selected by method M_i for question Q_j according to cosine similarity. Based on B_{ij} sets, questionnaires for expert evaluation were built using TypeForm. In the first round, the objective was to compare method variants based on LLMs, that is, M_2 (BETO) and M_3 (USE), against M_1 (TF-IDF), and verify that LLMs have superior performance than the baseline TF-IDF method. Then, in the second round, the objective was to compare the best two methods from round 1 with M_4 (ensemble method), thus determining the final winner MBF.

The questionnaires in both rounds began by presenting the evaluator with a prompt with general indications about the evaluation process, including descriptions of the two criteria. After the initial prompt, the questionnaire showed a sequence of two ranking items for each case question Q_j (see Table 3). In the first ranking item, the evaluator had to order responses from most relevant to least relevant according to the 'identification' criterion. In the second-ranking item, the ordering had to be done according to the 'decision' criterion. The responses displayed by the ranking items were the most relevant responses determined by each method (i.e., B_{ij}). When constructing the questionnaires, whenever the best responses selected by different methods were the same, the second-best responses were picked from sets B_{ij}. Thus, each ranking item in the first-round questionnaire displayed three responses, considering the best answer of each B_{ij} set, with i in {1,2,3} and j in {1,2}. In each ranking item, the answers appeared to raters displayed randomly.

The second questionnaire round was administered to evaluators a week after the first one. The structure of the second questionnaire was the same as that of the previous one. However, ranking items included the most relevant responses (B_{ij}) selected by the two best methods from round 1 and the ensemble method M_4 per question Q_j.

Per each rater R_k and question Q_j, scores of each method Mi were computed considering the order chosen for responses in the rankings. The scoring procedure for most relevant responses consisted of that per each R_k, the highest ranked response (and matching method variant M_i that selected it) was awarded 2 points, the following 1 point, and the third 0 points. Total scores per method variant M_i were calculated according to this scheme using an ad-hoc script coded in Python 3.

4.3 Evaluation of Selection Biases

In the comparative analysis of the M_1-M_4 variants of the proposed content selection method, two selection biases have been considered: bias related to the content selection of answers based on the student's inclination towards a specific decision regarding the ethical dilemma and bias associated with the length of the selected answers.

The first bias is determined by considering that questions 1 and 2 of the case are of the 7-level semantic differential type with written justification (see Table 3). The aim is to determine if the sets containing the top 30 most relevant answers selected by the M_1-M_4 method variants have a distribution of the associated semantic differential value 1-7 different from the general data distribution and whether differences among the methods exist. Depending on data normality, a parametric or non-parametric test is computed to determine if these differences exist.

The choice of 30 responses considers that it would take several minutes for a teacher to read all of this content in real-time (i.e., a considerable part of the time allotted for an EthicApp phase), given the inter-quartile range of response length distributions of 25 to 40 words, as seen in Figure 3. On the other hand, the 30 most relevant answers are expected to consist of argumentation that refers to elements of the question and the case to correctly reflect the identification of the dilemma and the corresponding decision. Therefore, the most relevant answers should be longer than the average length of the general data distribution. Thus, this analysis aims to verify whether significant differences exist between the methods for selecting answers based on length. For this purpose, like in the semantic bias analysis, parametric or non-parametric mean

comparison tests are conducted to compare the results of the selections of the method variants.

4.4. Performance Evaluation

The content selection method proposed in this research and its four variants, M_1-M_4, were evaluated based on the time taken to complete processing tasks and memory consumption. Specifically, the time taken by each variant to complete the processing of multiple sets of student responses of different sizes was measured. Sizes utilized were from 100 to 1.000 responses, with a 100 step in between. Additionally, the memory consumption of each variant was calculated by periodically sampling it at 10ms intervals (*Valgrind*, 2023).

To conduct the above evaluations, a commodity PC was equipped with a quad-core 4.2 GHz Intel Core i7 7700K CPU, 64GB RAM, and an NVIDIA GeForce GTX 1080 Ti GPU.

5 Results

5.1 Expert evaluation

Descriptive statistics of experts' scores in rounds 1 and 2 are presented in Tables 5 and 6, respectively. These include Kruskal-Wallis rank sum tests (non-parametric) for median differences among method variant total scores. The choice of the Kruskal-Wallis test was influenced by the non-normality found in the data, which was confirmed by prior Shapiro-Wilk tests, which mostly returned significant p-values, indicating deviations from normality. Figures 4 and 5 present score histograms supporting this finding and offering further insights. In round 1, as expected, LLMs BETO (M_2) and USE (M_3) outperformed TF-IDF (M_1) in selecting the most relevant student responses. It can be observed that while the mode score for TF-IDF is 0 across the board, BETO consistently has a mode score of 2, and USE's mode scores fluctuate depending on the question, never reaching a mode score of 2 in any instance. When looking at the total score, BETO emerged as the clear winner by a significant margin, with this difference being statistically significant (as evidenced in Table 5). As measured by Cliff's delta, a medium effect size is observed between M_2 and M_1 (δ=0.545) and between M_2 and M_3 (δ=0.496).

Question	**Method variant / Criteria**	**Identification**		**Decision**		**Total**	
		M	*SD*	*M*	*SD*	*M*	*SD*
Q_1	M_1 (TF-IDF)	0.73	0.79	0.73	0.79	1.45	1.44
	M_2 (BETO)	*1.36*	0.81	*1.55*	0.69	*2.91*	1.30
	M_3 (USE)	0.91	0.83	0.73	0.79	1.64	1.43
	Kruskal-Wallis $\chi^2(2) = 6.1034$, $p<0.05$						
	Method variant / Criteria	**Identification**		**Decision**		**Total**	
		M	*SD*	*M*	*SD*	*M*	*SD*
Q_2	M_1 (TF-IDF)	0.64	0.81	0.45	0.69	1.09	1.30
	M_2 (BETO)	*1.36*	0.81	*1.64*	0.67	*3.00*	1.18

	M_3 (USE)	1.00	0.77	0.91	0.70	2.91	1.22
	Kruskal-Wallis $\chi^2(2) = 10.0$, $p<0.01$						

Table 5: Scores of method variants in round 1 of expert evaluation in selecting students' most relevant responses

In the second round of expert evaluation, it is noteworthy that methods M_2 and M_4 chose the same most relevant response for Q_1. Consequently, their second-best responses were incorporated into the ranking. BETO's total score for Q_1 continues to be higher than USE's. Notably, despite the ensemble method scoring higher than USE and BETO in Q_1, its performance was inconsistent in Q_2 (see Figure 5 and Table 6). In Q_2, the Ensemble method ranked the lowest, potentially due to the inclusion of responses selected using TF-IDF. BETO's top response was evaluated more highly by experts in the identification criterion than USE's top response. However, the decision criterion obtained the opposite result, although with a higher variance in reviewer scores. BETO achieved a more significant total score than USE and the Ensemble method. However, total score differences were found to be non-significant according to the Kruskal-Wallis rank sum tests for mean differences conducted.

Question	**Method variant / Criteria**	**Identification**		**Decision**		**Total**	
		M	**SD**	**M**	**SD**	**M**	**SD**
Q_1	M_2 (BETO)	0.55	0.52	*1.27*	0.65	1.82	0.75
	M_3 (USE)	1.00	0.89	0.55	0.82	1.55	1.44
	M_4 (Ensemble)	*1.45*	0.82	1.18	0.87	*2.64*	1.50
	Kruskal-Wallis $\chi^2(2) = 4.2854$, *n.s.*						
	Method variant / Criteria	**Identification**		**Decision**		**Total**	
		M	**SD**	**M**	**SD**	**M**	**SD**
Q_2	*M_2 (BETO)*	*1.45*	0.69	1.09	0.94	*2.55*	1.29
	M_3 (USE)	0.73	0.79	*1.18*	0.87	1.91	1.51
	M_4 (Ensemble)	0.82	0.87	0.73	0.65	1.55	1.29
	Kruskal-Wallis $\chi^2(2) = 3.012$, *n.s.*						

Table 6: Scores of method variants in round 2 of expert evaluation in selecting top students' responses

The highest total score assignable by an expert using this evaluation methodology is 4.0 per question. No method achieved an average score close to this maximum; the highest average score recorded was 3.0 points, attained by the method using BETO in the first evaluation round for Q_2. However, this outcome cannot be attributed to the evaluation methods themselves but rather to the content selected by these methods, and this is due to two main reasons. Firstly, the students who participated in the activities that generated the dataset were first-year students in their first semester at the School of Engineering and Sciences, grappling with ethical dilemmas. They needed to gain prior university-level training in subjects like philosophy, ethics, and written communication with an emphasis on rhetoric, argumentation, and essay writing. Secondly, these students primarily engaged in the activity through their smartphones. The EthicApp interface provides a text area for responses on the smartphone screen that is only a few lines in size, and there is a tendency among students in this setting to write

brief and non-detailed texts (see Figure 3 and Table 4). This is further exacerbated by the fact that the time allotted for the phases of the activity is limited, typically at most 10 to 15 minutes per phase.

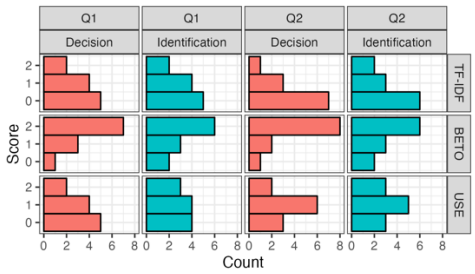

Figure 4: Expert evaluation scores in round 1 per question, criterion, and method variant

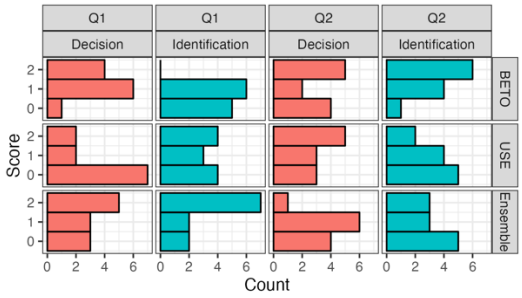

Figure 5: Expert evaluation scores in round 2 per question, criterion, and method variant

5.2 Evaluation of Selection Biases

Based on the Shapiro-Wilk test, it was determined that the distributions of student response lengths to questions 1 and 2 in the dataset were highly non-normal. Consequently, the non-parametric Kruskal-Wallis test was chosen to compare the means of the response distributions selected by the methods with the overall distribution of responses to each question. According to this test, the results in Table 7 indicate a significant difference in the length distributions of the 30 most relevant responses chosen by the methods compared to the overall data distribution in each case.

Question	Method	Min	Median	Max	Mean	SD
Q_1	M_1(TF-IDF)	15	37.0	108	39.80	19.10
	M_2 (BETO)	26	53.0	152	60.10	26.53
	M_3 (USE)	24	46.5	108	52.03	21.02
	M_4 (Ensemble)	21	49.5	152	54.30	27.35
	Full Dataset	1	30.0	152	31.92	17.57
	Kruskall-Wallis $\chi 2(2) = 89.58$, $p<0.05$					

Q_2	M_1(TF-IDF)	13	44.5	117	50.36	27.46
	M_2 (BETO)	34	75.0	117	70.23	20.50
	M_3 (USE)	20	44.5	117	51.73	26.58
	M_4 (Ensemble)	25	66.0	117	65.73	24.48
	Full Dataset	3	29.0	117	32.97	18.11
	Kruskal-Wallis $\chi 2(2) = 117.9$, $p<0.05$					

Note: The 30 most relevant responses selected by method variants are compared against the full data set (see Table 4).

Table 7: Descriptive statistics for response lengths (number of words) per question and method

Table 7 shows the method variant based on BETO (M_2) sent to select more extended responses, followed by the USE variant (M_3) and the ensemble variant. TF-IDF tends to choose shorter responses. These findings were consistently observed for responses to both questions in the dataset. More extended responses generally include more elaboration and detail; however, the responses in the EthicApp activity are typically short, mostly in the range from 37 to 75 words, while the mean of the general distribution is around 30 words. It is evident, then, that the responses selected by the methods tend to be longer than those of the general sample. If these responses are automatically presented to the teacher on an EthicApp dashboard, the teacher would have access to more content-rich responses without manually inspecting the entire set of answers, given the synchronous and time-limited context in which the activity takes place.

In Table 8, distributions of student quantitative response values in a 1-7 semantic differential scale format are displayed, linked to the written responses selected by each method variant. It should be noted that for questions 1 and 2 in the entire dataset, student responses with a value of 7 were not found. It can be observed that only in the case of question 1 is there a significant difference between the distributions. However, TF-IDF and BETO have similar results for this question, while USE tends to select responses that are marginally more inclined toward the lower end of the scale.

The ensemble method did yield a result for Q_1 that deviated from the other methods and leaned towards selecting responses biased towards the higher end of the scale. However, for Q_2, the range of differences tends to be more constrained, resulting in more minor and non-significant differences. Considering these findings, it cannot be asserted that the methods select strongly biased responses compared to the entire dataset. This is pedagogically appropriate since the methods' response selection needs to encompass a range of student stances recorded on the semantic differential scales of the questions. This assists teachers in evaluating feedback for students, ensuring they can be presented with varied perspectives.

Question	**Method**	**Min**	**Median**	**Max**	**Mean**	**SD**
Q_1	M_1(TF-IDF)	2	3	6	3.30	1.34
	M_2 (BETO)	2	3	6	3.30	1.17
	M_3 (USE)	1	3	6	3.10	1.29
	M_4 (Ensemble)	2	3	6	3.53	1.22
	Full Dataset	1	3	6	2.90	1.23
	Kruskal-Wallis $\chi 2(2) = 11.28$, $p<0.05$					

Q_2	M_1(TF-IDF)	1	2	6	2.76	1.43
	M_2 (BETO)	1	2	6	2.66	1.44
	M_3 (USE)	1	2	6	2.36	1.37
	M_4 (Ensemble)	1	2	6	2.56	1.43
	Full Dataset	1	2	6	2.59	1.45
	Kruskall-Wallis $\chi 2(2) = 1.40$, n.s.					

Note: The 30 most relevant responses selected by method variants are compared against the full data set (see Table 4).

Table 8: Descriptive statistics of semantic differential scale values linked to students' written responses per question and method

5.3 Resource evaluation

Figure 6 shows the time consumed by the different variants of the model, with two runtime configurations: execution using only CPU and execution using GPU. The M_4 variant without a GPU is the least efficient method, taking more than 250 seconds to process 1,000 student responses. This behaviour was expected, as the variant needs to run language models used in M_2 and M_3 and compute TF-IDF used in M_1.

TF-IDF in M_1 was confirmed (and expected) to require considerably fewer computational resources than language models. Its operation does not benefit from using a GPU in its implementation. Variants involving language models (i.e., M_2, M_3, and M_4) benefit from using a GPU, as processing times tend to be linear. For example, the second least efficient method is M_2 (BETO) without a GPU, which takes more than 150 seconds to process 1,000 student responses. The variant using a GPU shows an order-of-magnitude improvement. On the other hand, variant M_3 based on the USE language model shows an interesting behaviour. It is observed that the use of GPU only slightly improves performance and that it is generally more efficient than BETO.

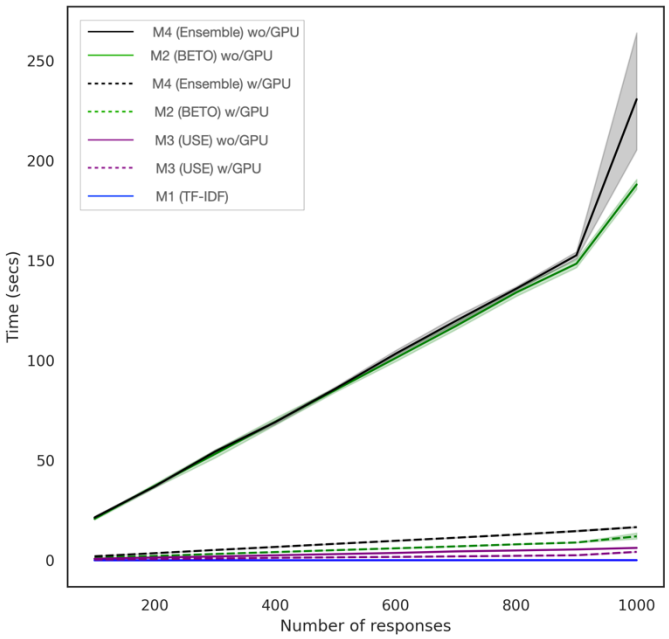

Figure 6: Processing time required by each method variant (M_1–M_4) to analyze varying numbers of student responses, comparing CPU-only and GPU-accelerated execution

Table 9 presents descriptive statistics for memory usage of each model variant when processing all student's responses for question 1 (see Table 3). M_4 was the most memory-intensive variant; memory consumption peaked at around 3 GB. This is explained by the need for the ensemble method to contain data structures and models utilized in all other variants. M_2 and M_3 had similar memory requirements (max = ~1.8 GB). This was expected, as both are transformer-based language models with similar architectures.

M_1 (18.9 MB) achieved the slightest memory usage, as it does not require an in-memory language model like the other model variants. The only relevant memory consumption comes from storing its data structures (vectors) and computing with them. Regarding variability in memory usage, M_4 had the highest, with a standard deviation of approximately 1 GB. This is explained by the ensemble's sequential use of the two memory-intensive language models.

Variant	*M*	*SD*	*min*	q1	q2	q3	q4
M_1	18.9	11.47	0	10.1	18.57	26.8	47.42
M_2	1274.1	657.2	0	784.4	1801.64	1802.4	1802.4
M_3	1030.2	493.5	0	880.9	1374.76	1374.9	1843.1
M_4	2141.4	938.3	0	1829.5	2718.61	2728.7	3098.9

Table 9: Memory usage (MB) descriptive statistics per each variant

6 Discussion

In the initial user evaluation, BETO (M_2) was superior in retrieving the top responses. As supported by [Araujo, 22], USE showed strength in handling both extremes. This demonstrates USE's versatility across diverse NLP tasks due to its multilingual training.

The first research question, RQ_1, focused on finding the most effective method for selecting relevant student responses. Initially, BETO outperformed others in retrieving top responses. In subsequent evaluations, the Ensemble method occasionally surpassed BETO, which consistently performed well overall. Despite BETO's computational demands, it emerged as a preferable option for this research setting, considering its efficiency and effectiveness without considering computational costs.

Ensemble methods suggested to improve performance by integrating outputs from multiple models showed variable success in this study. Ensemble methods have enhanced performance by amalgamating scores from various machine-learning models [Dietterich, 00]. Including the TF-IDF model may have diluted its efficacy and hampered its success. The inconsistency across data points makes a strong case for revising the ensemble approach with different language models.

The second research question, RQ_2, addressed whether the methods favoured longer or semantically different responses. BETO tended to select lengthier responses, which generally provided more detailed insights. There were no significant deviations in semantic representation across methods, suggesting the selected responses were balanced and informative for educational monitoring. Therefore, the responses chosen by the methods can be instructive for teachers monitoring the learning activities without introducing biases that might present the teacher or students with a distortion regarding the representativeness of the perspectives shared by students in their ethical judgments.

The third research question, RQ_3, evaluated computational resource use alongside performance. BETO, although requiring GPU support for optimal performance, was resource-intensive. In contrast, the USE model efficiently handled tasks without GPU enhancement, offering a cost-effective solution for CSCL platforms in resource-constrained environments like the Spanish-speaking Global South.

This study suggests that educational technology developers can use the content analysis methods proposed here for Spanish-speaking audiences in the developing world at a low cost. These methods allow for the development of dashboards in EthicApp and other CSCL domain applications requiring real-time analysis of qualitative student content. A citadel architecture incorporates a central monolithic system supplemented by specialized applications, which is beneficial for collaborative learning platforms using technologies like Ruby on Rails and ExpressJS [Thijs, 20]. This setup supports repurposing the content analysis method across various applications, enhancing the development of educational technologies using NLP.

7 Research Limitations

This study presents several limitations that must be acknowledged. First, it relies on a single dataset derived from student responses to one ethical case involving two dilemmas. While the dataset includes over one thousand responses from a large cohort,

further validation across different cases, disciplines, and student populations is necessary to ensure broader generalizability of findings.

Second, the proposed content selection heuristic—based on the cosine similarity between student responses and the question and case embeddings—prioritizes textual alignment with the instructional prompt. While this approach is both computationally efficient and pedagogically intuitive, it introduces important trade-offs. Specifically, it may overlook short but high-quality responses that demonstrate originality, critical thinking, or alternative ethical reasoning. These responses might diverge lexically from the case text or the question, yet still offer valuable and nuanced perspectives. As a result, there is a risk of underrepresenting dissenting or less conventional viewpoints, particularly those framed concisely or abstractly.

This limitation is especially relevant in ethics education, where fostering diverse interpretations and encouraging principled disagreement are pedagogical goals. Future research should therefore explore complementary methods that preserve the low computational footprint of the overall approach. For instance, lightweight clustering techniques or topic modeling using precomputed sentence embeddings could help surface responses that are semantically distinct yet relevant. Alternatively, rule-based filters or hybrid heuristics could be designed to identify indicators of originality or critical stance, without requiring full-scale supervised models. These strategies aim to broaden the scope of relevant response detection while remaining compatible with real-time processing on commodity hardware.

Third, the absence of ground truth labels across the full dataset prevents training of supervised models that could surpass heuristic-based methods. While expert evaluation provided a proxy for relevance, manual annotation of a larger corpus remains necessary to establish a gold standard.

Fourth, including the article's own authors among the expert reviewers introduces a potential source of bias, even if unavoidable in early-stage evaluations. Although steps were taken to ensure fair and independent assessment, such as anonymizing response order and distributing reviewer tasks, the lack of complete external review could influence the objectivity of the reported outcomes. Future studies should incorporate fully independent expert panels to strengthen the credibility of evaluations.

Fifth, although BETO and USE consistently tended to select longer responses, which were interpreted as more content-rich and pedagogically useful, this introduces another potential limitation. Verbosity does not always equate to higher quality. Concise responses may convey deep insight, originality, or nuanced ethical positions using minimal wording. The current approach may inadvertently undervalue these contributions, favoring length over substance. A more refined evaluation strategy is needed to avoid equating response length with quality and to ensure that brevity does not lead to the exclusion of highly relevant perspectives.

Sixth, the ensemble method explored in this study was implemented using a simple response overlap strategy to aggregate top responses across models. While this approach is easy to implement and interpret, it may not fully exploit the complementarity of the models. Future work could investigate alternative fusion strategies such as weighted aggregation or stacking-based ensemble learning to improve robustness and performance. To facilitate replication and further experimentation, we provide open access to the implementation code [Carvallo, 25] and the anonymized dataset [Alvarez, 25b] used in this study.

Lastly, although BETO proved to be the most effective method in terms of expert evaluations, it requires GPU support for optimal performance. This limits its applicability in low-resource settings, despite the study's emphasis on computational frugality. The USE-based variant offers a more scalable alternative, but with slightly lower performance.

8 Conclusions and future work

As real-time case-based learning activities generate increasing amounts of student-generated text, the orchestration load for educators becomes a critical challenge. Advanced educational technologies can help alleviate this burden by assisting in the identification of relevant student responses, allowing teachers to provide targeted and timely feedback without manually reviewing excessive amounts of text. In this study, we introduced and evaluated an automated content selection method designed to operate on commodity hardware without the computational demands of large-scale language models.

We tested four model variants: BETO (a Spanish-language BERT model), Universal Sentence Encoder (USE), the TF-IDF statistical baseline, and an ensemble method combining all three approaches. The aim was to determine the most effective approach for selecting relevant student responses in case-based learning activities within Computer-Supported Collaborative Learning (CSCL) environments.

Evaluation by 11 experts showed that BETO consistently outperformed USE and TF-IDF in selecting high-quality responses. However, BETO required GPU acceleration, making it less practical for deployment in low-resource environments. USE provided a strong alternative, performing comparably without requiring GPUs, making it suitable for real-time applications on standard CPU-based hardware. The ensemble method did not significantly improve performance over the best-performing individual models. Additionally, BETO and other methods showed a tendency to select longer responses, which was considered beneficial as these responses were often more elaborated and informative. Importantly, no significant semantic bias was observed, ensuring that the response selection process remained representative and did not skew students' perspectives on the ethical dilemmas under discussion.

Future work will explore alternative ensemble combinations and further test a USE-based EthicApp dashboard in case-based learning settings. Additionally, we aim to enhance EthicApp by incorporating small-group discussion support, allowing teachers to compare top and bottom responses to facilitate structured peer feedback. Finally, we plan to integrate topic modelling techniques to help educators quickly identify emerging themes in student discussions, further improving the dashboard's effectiveness in reducing teacher orchestration load and enhancing instructional support in case-based learning environments.

References

[AAC&U, n.d.] AAC&U: VALUE Rubrics – Ethical Reasoning, n.d. Available at: https://www.aacu.org/initiatives/value-initiative/value-rubrics/value-rubrics-ethical-reasoning (Accessed: March 11, 2025).

[Altamirano, 20] Altamirano, M., Jiménez, A. y Araya, R.: Lessons clustering using topics inferred by unsupervised modeling from textbooks, Methodologies and Intelligent Systems for Technology Enhanced Learning, 10th International Conference, 2020.

[Alvarez, 21] Alvarez, C., Zurita, G., Carvallo, A., Ramírez, P., Bravo, E. y Baloian, N.: Automatic Content Analysis of Student Moral Discourse in a Collaborative Learning Activity, in Hernández-Leo, D., Hishiyama, R., Zurita, G., Weyers, B., Nolte, A. y Ogata, H. (eds): Collaboration Technologies and Social Computing. CollabTech 2021. Lecture Notes in Computer Science, vol. 12856, Springer, Cham, 2021, https://doi.org/10.1007/978-3-030-85071-5_1.

[Alvarez, 25a] Alvarez, C., Amarasinghe, I., Zurita, G., Hernandez-Leo, D., Hakami, L. y Rojas, L.: Measurement of Teacher's Orchestration Load: A Framework and a Case Study on Tool Flexibility, IEEE Access, vol. 13, pp. 39035–39050, 2025, doi: 10.1109/ACCESS.2025.3531241.

[Alvarez, 25b] Alvarez, C., Carvallo, A., Zurita, G.: 'Spanish Dataset used in study "Low-Footprint NLP for Reducing Teacher's Orchestration Load in Computer-Supported Case-Based Learning Environments", by Alvarez, C., Carvallo, A., and Zurita, G.' (Version 1);. figshare (2025, May 28). https://doi.org/10.6084/m9.figshare.29168555.v1

[Amarasinghe, 21] Amarasinghe, I., Hernández-Leo, D., Hoppe, H. U.: Deconstructing orchestration load: Comparing teacher support through mirroring and guiding, International Journal of Computer-Supported Collaborative Learning, 16, 307–338, 2021, https://doi.org/10.1007/s11412-021-09351-9.

[Angeli, 04] Angeli, C.: The effects of case-based learning on early childhood pre-service teachers' beliefs about the pedagogical uses of ICT, Journal of Educational Media, 29(2), 139–151, 2004.

[Apostolou, 13] Apostolou, B., Dull, R. B. y Schleifer, L. L.: A framework for the pedagogy of accounting ethics, Accounting Education, 22(1), 1–17, 2013.

[Araujo, 22] Araujo, V., Carvallo, A., Kundu, S., Cañete, J., Mendoza, M., Mercer, R. E., Bravo-Marquez, F., Moens, M.-F. y Soto, A.: Evaluation Benchmarks for Spanish Sentence Representations, arXiv preprint arXiv:2204.07571, 2022.

[Balyan, 20] Balyan, R., McCarthy, K. S. y McNamara, D. S.: Applying natural language processing and hierarchical machine learning approaches to text difficulty classification, International Journal of Artificial Intelligence in Education, 30, 337–370, 2020.

[Benedict, 22] Benedict, A., Al-Hossami, E., Dorodchi, M., Benedict, A. y Wiktor, S.: Pilot Recommender System Enabling Students to Indirectly Help Each Other and Foster Belonging Through Reflections, LAK22: 12th International Learning Analytics and Knowledge Conference, 2022.

[Butts, 05] Butts, D. y Janie, B.: Nursing ethics: across the curriculum and into practice book review, Online Journal of Health Ethics, 2(2), 2, 2005.

[Cáceres, 18] Cáceres, M., Nussbaum, M., Marroquín, M., Gleisner, S., Marquínez, J. T.: Building arguments: Key to collaborative scaffolding, Interactive Learning Environments, 26(3), 355–371, 2018.

[Cañete, 20] Cañete, J., Chaperon, G., Fuentes, R., Ho, J.-H., Kang, H. y Pérez, J.: Spanish pre-trained BERT model and evaluation data, Pml4dc at ICLR, 1–10, 2020.

[Cañete, 22] Cañete, J., Donoso, S., Bravo-Marquez, F., Carvallo, A. y Araujo, V.: Albeto and distilbeto: Lightweight Spanish language models, arXiv preprint arXiv:2204.09145, 2022.

[Carvallo, 2025] Carvallo, A.: Semantic Educational Analysis – Implementation Code and Dataset, https://github.com/afcarvallo/semantic_educational_analysis.git (accessed: 27.05.2025)

[Cer, 18] Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S. y Tar, C.: Universal Sentence Encoder, arXiv preprint arXiv:1803.11175, 2018.

[Chambers, 15] Chambers, C. y Ransom, H.: Teaching ethics in higher education using the values–issues–action (VIA) model, Journal for the Study of Postsecondary and Tertiary Education, 1, 13–34, 2015.

[Dietterich, 00] Dietterich, T. G.: Ensemble methods in machine learning, Multiple Classifier Systems: First International Workshop, MCS 2000, Cagliari, Italy, 21–23 June 2000, Proceedings 1, 2000.

[Dillenbourg, 11] Dillenbourg, P., Zufferey, G., Alavi, H., Jermann, P., Do-Lenh, S., Bonnard, Q., Cuendet, S. y Kaplan, F.: Classroom orchestration: The third circle of usability, 2011.

[Donkin, 23] Donkin, R., Yule, H., Fyfe, T.: Online case-based learning in medical education: A scoping review, BMC Medical Education, 23, 564, 2023, https://doi.org/10.1186/s12909-023-04520-w.

[Dood, 22] Dood, A., Winograd, B., Finkenstaedt-Quinn, S., Gere, A. y Shultz, G.: PeerBERT: Automated Characterization of Peer Review Comments across Courses, LAK22: 12th International Learning Analytics and Knowledge Conference, 2022.

[Evans, 16] Evans, C.: Re-thinking case-based assessments in business management education, The International Journal of Management Education, 14(2), 161–166, 2016.

[Geden, 21] Geden, M., Emerson, A., Carpenter, D., Rowe, J., Azevedo, R. y Lester, J.: Predictive student modeling in game-based learning environments with word embedding representations of reflection, International Journal of Artificial Intelligence in Education, 31, 1–23, 2021.

[Heimann, 17] Heimann, F. y Pieth, M.: Confronting corruption: past concerns, present challenges, and future strategies, Oxford University Press, 2017.

[Hess, 18] Hess, J. L. y Fore, G.: A systematic literature review of US engineering ethics interventions, Science and Engineering Ethics, 24(2), 551–583, 2018.

[Instituto-de-Ingenieros, 21] Instituto-de-Ingenieros: Instituto de Ingenieros de Chile, Ética e Ingeniería, Conceptos y Orientaciones para el Ejercicio Profesional, 2021, https://www.iing.cl/document/2021-etica-e-ingenieria/.

[Jensen, 21] Jensen, E., Pugh, S. y D'Mello, S. K.: A deep transfer learning approach to modeling teacher discourse in the classroom, LAK21: 11th International Learning Analytics and Knowledge Conference, 2021.

[Johnson, 12] Johnson, J. F., Bagdasarov, Z., Connelly, S., Harkrider, L., Devenport, L. D., Mumford, M. D. y Thiel, C. E.: Case-based ethics education: The impact of cause complexity and outcome favorability on ethicality, Journal of Empirical Research on Human Research Ethics, 7(3), 63–77, 2012.

[Kasneci, 23] Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S. y Hüllermeier, E.: ChatGPT for good? On opportunities and challenges of large language models for education, Learning and Individual Differences, 103, 102274, 2023.

[Kim, 21] Kim, B., Kim, H., Lee, S.-W., Lee, G., Kwak, D., Jeon, D. H., Park, S., Kim, S., Kim, S. y Seo, D.: What changes can large-scale language models bring? intensive study on Hyperclova: Billions-scale Korean generative pretrained transformers, arXiv preprint arXiv:2109.04650, 2021.

[Kobbe, 07] Kobbe, L., Weinberger, A., Dillenbourg, P., Harrer, A., Hämäläinen, R., Häkkinen, P. y Fischer, F.: Specifying computer-supported collaboration scripts, International Journal of Computer-Supported Collaborative Learning, 2(2), 211–224, 2007.

[Kolodner, 13] Kolodner, J. L., Owensby, J. N. y Guzdial, M.: Case-based learning aids, in Handbook of Research on Educational Communications and Technology, pp. 820–852, Routledge, 2013.

[Le Scao, 23] Le Scao, T., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A. S., Yvon, F. y Gallé, M.: Bloom: A 176b-parameter open-access multilingual language model, 2023.

[Leitão, 00] Leitão, S.: The potential of argument in knowledge building, Human Development, 43(6), 332–360, 2000.

[Li, 19] Li, S., Ye, X. y Chen, W.: Practice and effectiveness of "nursing case-based learning" course on nursing student's critical thinking ability: A comparative study, Nurse Education in Practice, 36, 91–96, 2019.

[Li, 24] Li, X., Han, G., Fang, B. y He, J.: Advancing the In-Class Dialogic Quality: Developing an Artificial Intelligence-Supported Framework for Classroom Dialogue Analysis, The Asia-Pacific Education Researcher, 1–15, 2024.

[Mageira, 22] Mageira, K., Pittou, D., Papasalouros, A., Kotis, K., Zangogianni, P. y Daradoumis, A.: Educational AI Chatbots for Content and Language Integrated Learning, Applied Sciences, 12(7), 3239, 2022, https://www.mdpi.com/2076-3417/12/7/3239.

[McLean, 16] McLean, S. F.: Case-based learning and its application in medical and health-care fields: A review of worldwide literature, Journal of Medical Education and Curricular Development, 3, 2016.

[Mello, 21] Mello, R. F., Fiorentino, G., Miranda, P., Oliveira, H., Raković, M. y Gašević, D.: Towards Automatic Content Analysis of Rhetorical Structure in Brazilian College Entrance Essays, Artificial Intelligence in Education: 22nd International Conference, AIED 2021, Utrecht, The Netherlands, June 14–18, 2021, Proceedings, Part II, 2021.

[Mello, 22] Mello, R. F., Neto, R., Fiorentino, G., Alves, G., Arêdes, V., Silva, J. V. G. F., Falcão, T. P. y Gašević, D.: Enhancing instructors' capability to assess open-response using natural language processing and learning analytics, Educating for a New Future: Making Sense of Technology-Enhanced Learning Adoption: 17th European Conference on Technology Enhanced Learning, EC-TEL 2022, Toulouse, France, September 12–16, 2022, Proceedings, 2022.

[Merseth, 92] Merseth, K. K.: Cases for decision making in teacher education, Case Methods in Teacher Education, 50–63, 1992.

[Mohit, 14] Mohit, B.: Named entity recognition, Natural Language Processing of Semitic Languages, 221–245, 2014.

[Moore, 22] Moore, S., Nguyen, H. A., Bier, N., Domadia, T. y Stamper, J.: Assessing the quality of student-generated short answer questions using GPT-3, European Conference on Technology Enhanced Learning, 2022.

[Mushtaq, 25] Mushtaq, A., Naeem, M. R., Ghaznavi, I., Taj, M. I., Hashmi, I. y Qadir, J.: Harnessing Multi-Agent LLMs for Complex Engineering Problem-Solving: A Framework for Senior Design Projects, arXiv preprint arXiv:2501.01205, 2025.

[Narayanan, 21] Narayanan, D., Shoeybi, M., Casper, J., LeGresley, P., Patwary, M., Korthikanti, V., Vainbrand, D., Kashinkunti, P., Bernauer, J. y Catanzaro, B.: Efficient large-scale language model training on GPU clusters using Megatron-LM, Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, 2021.

[Ocone, 13] Ocone, R.: Engineering ethics and accreditation, Education for Chemical Engineers, 8(3), e113–e118, 2013.

[Patel, 15] Patel, P.: Engineers, ethics, and the VW scandal, IEEE Spectrum, 25, 2015.

[Patil, 24] Patil, R., Gudivada, V.: A review of current trends, techniques, and challenges in large language models (LLMs), Applied Sciences, 14(5), 2074, 2024.

[Perkins, 91] Perkins, D. V.: A case-study assignment to teach theoretical perspectives in abnormal psychology, Teaching of Psychology, 18(2), 97–99, 1991.

[Pierrakos, 19] Pierrakos, O., Prentice, M., Silverglate, C., Lamb, M., Demaske, A., Smout, R.: Reimagining engineering ethics: From ethics education to character education, in Proc. 2019 IEEE Frontiers in Education Conference (FIE), 1–9, IEEE, October 2019.

[Rafinda, 19] Rafinda, A., Gal, T. y Purwaningtyas, P.: Business Ethics Course On Student Moral Reasoning, Oradea Journal of Business and Economics, 4(Special), 60–68, 2019.

[Schwartz, 20] Schwartz, R., Dodge, J., Smith, N. A. y Etzioni, O.: Green AI, Communications of the ACM, 63(12), 54–63, 2020.

[Shulman, 92] Shulman, J.: Case methods in teacher education, Teachers College Press, Teachers College, Columbia University, 1992.

[Stegmann, 07] Stegmann, K., Weinberger, A. y Fischer, F.: Facilitating argumentative knowledge construction with computer-supported collaboration scripts, International Journal of Computer-Supported Collaborative Learning, 2(4), 421–447, 2007.

[Taori, 23] Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P. y Hashimoto, T. B.: Alpaca: A strong, replicable instruction-following model, Stanford Center for Research on Foundation Models, 3(6), 7, 2023, https://crfm.stanford.edu/2023/03/13/alpaca.html.

[The, 19] The, B., Yang, L. y Wang, Q.: What's on Your Mind? Promoting Cognitive Engagement Using Utterance Annotations in Online Collaborative Learning, ICIS, 2019.

[Thijs, 20] Thijs, C.: The Citadel Architecture at AppSignal, 2020, https://blog.appsignal.com/2020/04/08/the-citadel-architecture-at-appsignal.html.

[Touvron, 23] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E. y Azhar, F.: Llama: Open and efficient foundation language models, arXiv preprint arXiv:2302.13971, 2023.

[Uchiyama, 23] Uchiyama, S., Umemura, K. y Morita, Y.: Large Language Model-based System to Provide Immediate Feedback to Students in Flipped Classroom Preparation Learning, arXiv preprint arXiv:2307.11388, 2023.

[Valgrind, 23] Valgrind: Valgrind, 2023, https://valgrind.org/.

[Wang, 21] Wang, H., Tlili, A., Lehman, J. D., Lu, H. y Huang, R.: Investigating feedback implemented by instructors to support online competency-based learning (CBL): a multiple case study, International Journal of Educational Technology in Higher Education, 18(1), 1–21, 2021.

[Wang, 24] Wang, D., Tao, Y., Chen, G.: Artificial intelligence in classroom discourse: A systematic review of the past decade, International Journal of Educational Research, 123, 102275, 2024.

[Wu, 22] Wu, Y., Henriksson, A., Nouri, J., Duneld, M. y Li, X.: Retrieving Key Topical Sentences with Topic-Aware BERT When Conducting Automated Essay Scoring, Methodologies and Intelligent Systems for Technology Enhanced Learning, 12th International Conference, 2022.

[You, 19] You, Y., Li, J., Reddi, S., Hseu, J., Kumar, S., Bhojanapalli, S., Song, X., Demmel, J., Keutzer, K. y Hsieh, C.-J.: Large batch optimization for deep learning: Training BERT in 76 minutes, arXiv preprint arXiv:1904.00962, 2019.

[Zeng, 10] Zeng, R. y Blasi, L.: Learning through web-based multistoryline case studies: A design-based research, Quarterly Review of Distance Education, 11(3), 175, 2010.

[Zhang, 24] Zhang, Z., Zhang-Li, D., Yu, J., Gong, L., Zhou, J., Liu, Z., Hou, L. y Li, J.: Simulating classroom education with LLM-empowered agents, arXiv preprint arXiv:2406.19226, 2024.

[Zunger, 18] Zunger, J.: Computer science faces an ethics crisis. The Cambridge Analytica scandal proves it, Boston Globe, 22, 2018.