Conference Proceedings

Digital Technologies in the Contemporary Educational Reality

Views, Counterviews, Challenges, and Perspectives

4-6 April 2025, Preveza, Greece

Πρακτικά Συνεδρίου

Ψηφιακά Μέσα στη Σύγχρονη Εκπαιδευτική Πραγματικότητα

Θέσεις, Αντιθέσεις, Προκλήσεις και Προοπτικές

4-6 Απριλίου 2025, Πρέβεζα

Under the Aegis - Υπό την Αιγίδα

ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Υπουργείο Παιδείας, Θρησκευμάτων και Αθλητισμού

ISBN: 978-618-88572-0-9

Coorganisers - Facilitators

Διεύθυνση Δευτεροβάθμιας Εκπαίδευσης Πρέβεζας

ΔΙΕΘΝΕΣ ΠΑΝΕΠΙΣΤΗΜΙΟ ΤΗΣ ΕΛΛΑΔΟΣ

ΔΗΜΟΣ ΠΑΡΓΑΣ

ΔΗΜΟΣΖΗΡΟΥ

Περιφερειακή Ενότητα Πρέβεζας
ΠΕΡΙΦΕΡΕΙΑ ΗΠΕΙΡΟΥ

ΔΗΜΟΣ ΠΡΕΒΕΖΑΣ

ΕΠΙΜΕΛΗΤΗΡΙΟ ΠΡΕΒΕΖΑΣ

Τμήμα Λογιστικής & Χρηματοοικονομικής Πανεπιστημίου Ιωαννίνων

Secondary Education Administration of Preveza

Address: 9 Aminta, GR48100 Preveza, Greece

Tel: +30 26820 24135

email: mail@dide.pre.sch.gr

Διεύθυνση Δευτεροβάθμιας Εκπαίδευσης Πρέβεζας

Διεύθυνση: Αμύντα 9, Τ.Κ. 48100 Πρέβεζα, Ελλάδα

Τηλ: +30 26820 24135

email: mail@dide.pre.sch.gr

Digital Technologies in the Contemporary Educational Reality. Views, Counterviews, Challenges and Perspectives. Conference Proceedings.

Executive Editor:

Dr. Fani Biskanaki

Director

Secondary Education Administration of Preveza

Greece

Editors:

Fotini Dimakopoulou

Educational Quality Supervisor

Secondary Education Administration of Preveza

Greece

Dina Baga

Computer Scientist, Department of Informatics

Secondary Education Administration of Preveza

Greece

ISBN: 978-618-88572-0-9

Date Published: 4th April 2026

ePublished by DDE Prevezas (Secondary Education Administration of Preveza)

Cite as:

Biskanaki, F., Dimakopoulou, F. & Baga, D., (Eds.). (2026). *Conference Proceedings: Digital Technologies in the Contemporary Educational Reality. Views, Counterviews, Challenges and Perspectives. April 4-6, 2025, Preveza, Greece.* DDE Prevezas. 1ο ΔΙΕΘΝΕΣ ΣΥΝΕΔΡΙΟ ΔΔΕ ΠΡΕΒΕΖΑΣ (https://dinabaga.sites.sch.gr/)

*Conference articles were peer reviewed by elected reviewers as described in https://drive.google.com/file/d/1ARE_we9N92o16VfSd7S38Tc_KfZFgcmi/view

ARTIFICIAL INTELLIGENCE TO AUTOMATICALLY ASSESS STUDENT PERFORMANCE: THE CASE OF AISE PROJECT

Fani Biskanaki
Director
Diefthinsi Defterovathmias Ekpaidefsis Nomou
Prevezas, Greece

Athanasios Sypsas
Diefthinsi Defterovathmias Ekpaidefsis Nomou
Prevezas, Greece

Venediktos Hadjisavva
3rd Technical and Vocational School of Education and Training of
Lemessos, Cyprus

Andreas Paraskeva
3rd Technical and Vocational School of Education and Training of
Lemessos, Cyprus

Maria Evripidou
3rd Technical and Vocational School of Education and Training of
Lemessos, Cyprus

Achilleas Kostoulas
Doumag LTD
Cyprus

Ognjen Pavić
University of Kragujevac,
Serbia

Lazar Dašić
University of Kragujevac,
Serbia

Tijana Geroski
University of Kragujevac,
Serbia

Nenad Filipović
University of Kragujevac,
Serbia

Ana Orlović Kovačević
CARNET
Croatia

Themis Exarchos
Ionian University
Corfu, Greece

Costas Assimakopoulos
International Hellenic University
Greece

George Stalidis
International Hellenic University
Greece

Jelena Pavić
City of Kragujevac, University of Kragujevac
Serbia

Daniel Štifanić, Zlatan Car
Faculty of Engineering, University of Rijeka, Croatia
e-mail: eponymo.onoma@organisation.gr

ABSTRACT

This paper presents the methodology, results, and ongoing work of the AISE project (Artificial Intelligent platform to support Students by assessing their performance skills through predictive models created from their writing skills), funded under the Erasmus+ program. The project aims to apply state-of-the-art AI methods, especially retrieval-augmented generation (RAG) combined with Large Language Models (LLMs), for the automatic evaluation of student essays across three dimensions: Content, Organization, and Language. Essays were collected from secondary education students in Greece, Cyprus, Serbia, and Croatia, processed, and assessed using a modular AI pipeline. The results demonstrate high alignment with human ratings, especially in the Content dimension, while highlighting challenges in assessing linguistic subtleties. The AISE project paves the way for scalable, transparent, and personalized feedback systems in educational environments.

Keywords: Artificial Intelligence, Erasmus+, Language processing, Automatic Essay Scoring, Secondary Education.

INTRODUCTION

The rapid advances in Natural Language Processing (NLP) and AI have enabled the automation of complex educational tasks, including essay scoring. The AISE project, code 2023-1-EL01-KA220-SCH-000157157, addresses this challenge by developing an AI platform capable of automatically assessing secondary students' writing performance using advanced language models and structured evaluation criteria.

The AISE project is a European Erasmus+ Cooperation Partnership in the field of school education. It has been approved and funded by the Hellenic National Agency Erasmus+ (State Scholarships Foundation - IKY). The project is coordinated by the Diefthinsi Defterovathmias Ekpaidefsis Nomou Prevezas, Greece, and its partners include:

- International Hellenic University (Greece)
- Ionian University (Greece)
- University of Kragujevac (Serbia)
- City of Kragujevac (Serbia)
- University of Rijeka (Croatia)
- Ministry of Science and Education of Croatia, replaced by CARNET, (Croatia)
- 3rd Technical and Vocational School of Education and Training of Limassol (Cyprus)
- DOUMAG LTD (Cyprus)

The deployment of artificial intelligence (AI) in educational assessment accelerates marking while enriching the depth, consistency, and timeliness of feedback, ultimately fostering student growth. Within this landscape, Automated Essay Scoring (AES) stands out as a flagship application (Shetty et al., 2022) serving a key role in meeting the objectives of the AISE project. Early AES research relied on hand-crafted linguistic rules. Burstein et al. (1998), for example, built e-rater around syntactic, rhetorical, and topical features; their work showed that combining computational-linguistic indicators with statistical modelling could match human graders closely, making the system a credible "second reader" for high-stakes tests.

Since then, scores of studies have tried to capture higher-order qualities such as cohesion, relevance, and coherence—still formidable challenges. Surveys by Ramesh & Sanampudi (2021) and González-Calatayud et al. (2021) reveal that most systems emphasize style-based cues, with fewer tackling deep content evaluation. The latter review distilled 454 papers down to 22 key studies, underscoring the momentum behind AES for formative assessment: AI systems can return immediate, multidimensional feedback that helps learners revise in real time—provided they are grounded in sound pedagogical principles, not just technical accuracy. Recent work drives beyond rule-based methods. Kumar & Boulanger (2020) trained an AES model on a large corpus and scored essays across four rubrics—ideas, organization, style, and conventions—using linguistic indices to generate pointed, standards-aligned feedback. Likewise, Grivokostopoulou et al. (2016) combined interactive visualizations with automatic marking for search-algorithm exercises, reporting high agreement with human graders and measurable learning gains.

AI is also reshaping broader learning environments. Goel & Joyner (2017) embedded "nanotutors" into an online Knowledge-Based AI course, where automated micro-feedback contributed to online students outperforming their on-campus peers. Choi & McClenen (2020) fused Computerized Adaptive Testing with Dynamic Bayesian Networks to deliver personalized, diagnostic feedback in e-learning, while Ouguengay et al. (2015) applied a neuro-fuzzy inference system to gauge literacy skills in Amazigh within a learning-management system.

Two dominant technical paradigms now define AES research:

Large Language Models (LLMs)

Studies using GPT-3 (Mizumoto & Eguchi 2023), GPT-4 (Kostić et al. 2024), PaLM 2, Claude 2, and others (Pack et al. 2024) demonstrate that powerful generative models can approximate, and sometimes surpass, human scoring—though stability over time and across contexts remains a concern. Song et al. (2024) showed that open-source LLMs, combined with zero-shot, few-shot, or parameter-efficient tuning, can offer cost-effective essay scoring and revision support.

Feature-extraction approaches

When annotated data are scarce, systems that leverage educator-defined indicators still perform competitively and offer greater interpretability. Yang et al. (2019), for instance, blended shallow and deep semantic attributes to grade Chinese essays and supply constructive feedback.

In short, while large, data-hungry LLMs excel when ample training material exists, feature-driven or hybrid models remain valuable for low-resource settings or languages. The literature also highlights the importance of integrating pedagogical design, ensuring that automated scores translate into actionable, learner-centered feedback. These insights directly inform the AISE project's choice of a Retrieval-Augmented Generation framework, which combines the adaptability of LLMs with curated, educator-vetted exemplars to produce transparent, reliable, and scalable essay assessments.

PROJECT GOALS AND CONTEXT

The aim of the project is to apply Artificial Intelligence (AI) methods to create a platform capable of supporting the assessment of students' language skills in secondary education across four EU countries, namely: Greece, Cyprus, Serbia, and Croatia. To achieve this, data mining models utilizing various artificial intelligence technologies will be designed and applied. A key objective is to create predictive models that effectively assess students' skills. The project includes a data collection phase, in which sets of annotated written texts will be collected from high-school students and will be used for training and evaluating the AI models. To facilitate the data collection process, a collaborative cloud-based software platform will be developed to systematically register and store all collected student data, ensuring efficient tracking and analysis. In the data processing aspect, various AI techniques will be applied to unstructured text, collected by educators from the participating countries, in order to identify patterns and extract valuable insights.

The main activities focus on scientific research studies involving secondary education students, conducted through essays at different time periods by educators from participating European schools. State-of-the-art AI methods will be utilized to develop, evaluate, and validate highly scalable models.

METHODOLOGY

The methodology of the AISE project is structured into a set of interlinked phases, combining educational research, computational linguistics, and system engineering. The project's core aim is the development of a scalable and explainable Automated Essay Scoring (AES) platform that leverages state-of-the-art AI techniques to assess student writing across three dimensions: Content, Organization, and Language.

Preliminary Analysis and Literature Review

The first phase involved a comprehensive review of existing AI methodologies for writing skill assessment. This included a comparative analysis between two primary technological paradigms: Feature-based approaches, which rely on predefined linguistic and stylistic indicators (e.g., syntactic variety, lexical richness, cohesion metrics). Neural network–based methods, primarily centered around Large Language Models (LLMs), which learn from data patterns without explicit rule encoding. This review informed the project's decision to adopt a Retrieval-Augmented Generation (RAG) strategy—a hybrid solution that enriches LLM prompts with relevant, high-quality examples from an external knowledge base. This approach was selected to address key challenges in AES, including generalization across multilingual inputs, explainability, and alignment with pedagogical standards.

Project Requirements and Infrastructure Development

Following the conceptual foundation, a requirement analysis was conducted to define the functional and non-functional specifications of the platform. This phase involved all project partners and led to the development of:

- A cloud-based platform for secure essay storage and processing.
- A modular system architecture, enabling the separation of the essay evaluation pipeline into distinct stages.
- Text curation and harmonization workflows using AI-powered tools to clean, normalize, and annotate the collected data.

The platform was designed to support multilingual input, data de-identification, educator dashboards, and API endpoints for real-time interaction with the essay assessment engine.

Data Collection and Pre-processing

A key component of the AISE project was the establishment of a cloud-based platform designed to support the secure, structured, and collaborative collection of student essays across four participating countries: Greece, Cyprus, Croatia, and Serbia. The platform was developed to:

- Enable educators to upload essays in standardized formats (.docx or .odt).
- Associate each submission with anonymized metadata, including school ID, student code, essay prompt, and evaluation scores.
- Facilitate real-time monitoring of submissions, version tracking, and grading completeness.
- Support multilingual data inputs, ensuring language-specific encoding and compatibility.

At the first stage, essays were collected from secondary school students across four European countries: Greece, Cyprus, Croatia, and Serbia. Each essay included:

- Prompt (topic/question).
- Body (student response).
- Grading metadata (human scores in three dimensions).

All texts underwent anonymization, language normalization, and structuring into a three-part schema to ensure consistency across countries and grading teams.

Grading Schema and Label Encoding

Teachers assigned scores to each essay in the following dimensions:

- Content (maximum 40 points).
- Organization (maximum 30 points).
- Language (maximum 30 points).

These numeric scores were converted into qualitative categories—Excellent, Very Good, Good, and Marginal—based on standardized thresholds, ensuring a harmonized evaluation framework. Each essay was tagged with a combined label (e.g., "Very Good–Good–Excellent") to facilitate stratified sampling in subsequent stages, as it can be seen in the table (Table 1) below.

Table 2
Grade categories and related score ranges for each evaluation dimension

Grade	Content (40 pts)	Organization (30 pts)	Language (30 pts)
Excellent	31 – 40	24 – 30	24 – 30
Very Good	21 – 30	18 – 23	18 – 23
Good	11 – 20	11 – 17	11 – 17
Marginal	1 – 10	1 – 10	1 – 10

Train/Test Split Strategy

To ensure statistical validity and prevent data leakage, a stratified train/test split was performed using the combined qualitative label as the stratification key. Essays with rare combinations (i.e., unique or very low-frequency grade triads) were excluded from model training. The test set was completely isolated and never used during training, embedding, or retrieval, ensuring strict generalization and fairness.

Construction of Knowledge Bases

Three dimension-specific knowledge bases (KBs) were constructed—one each for Content, Organization, and Language. For each essay in the training set, the relevant body text and its qualitative grade were embedded using the MiniLM-L6-v2 transformer model, a light-weight and performant sentence encoder. These embeddings were indexed using FAISS, an efficient vector similarity search engine.

This architecture enables fast semantic retrieval of the top-k most similar training examples during model inference. Crucially, each dimension's KB is completely isolated, avoiding cross-contamination between grading criteria.

Retrieval-Augmented Generation (RAG) Models

At the heart of the platform are three LLM-based essay assessment modules, each tailored to a specific evaluation dimension. These modules are built upon DeepSeek-V3 (671B parameters) and operate within a RAG framework: At inference time, a new essay is embedded and used to query the relevant KB. The most similar annotated examples are retrieved and passed—along with the essay—to the LLM. A dimension-specific prompt template guides the model to produce a grade and rationale based on aligned pedagogical expectations. This design enables the LLM to contextualize grading, reduce hallucination, and generate traceable reasoning backed by real student examples.

Educational Research Integration

Parallel to the system development, the project initiated on-site educational research in all participating countries. Secondary school students contributed essays across different time periods. Educators participated in grading calibration workshops, feedback studies, and system evaluations. This research component supports:

- Cross-country alignment of grading standards.
- Real-world validation of AI predictions.
- The collection of educator and student feedback for platform refinement.

Performance Evaluation and Metrics

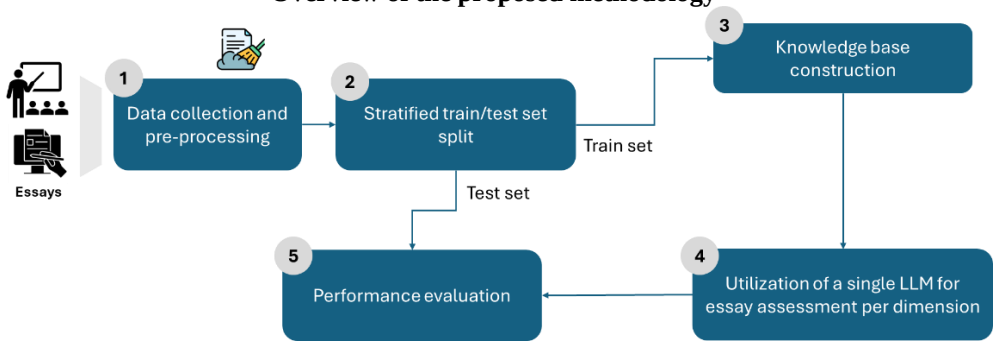
Each LLM was evaluated using the unified test set, with predicted grades compared against human-labeled ground truth. Metrics computed per dimension included:

1. Accuracy.
2. Macro Precision / Recall / F1-Score.
3. Balanced Accuracy.
4. Confusion Matrices to visualize prediction distributions.

Initial results revealed very high performance in the Content dimension (Accuracy: 0.976), strong performance in Organization, and more modest outcomes in Language—highlighting the complexity of capturing grammatical subtleties.

This multi-phase methodology ensures that the AISE system is transparent, modular, multilingual, and educationally grounded, capable of offering automated, yet interpretable, writing assessment aligned with European pedagogical standards. In the (Image 1), the overview of the proposed methodology is depicted.

Image 1
Overview of the proposed methodology



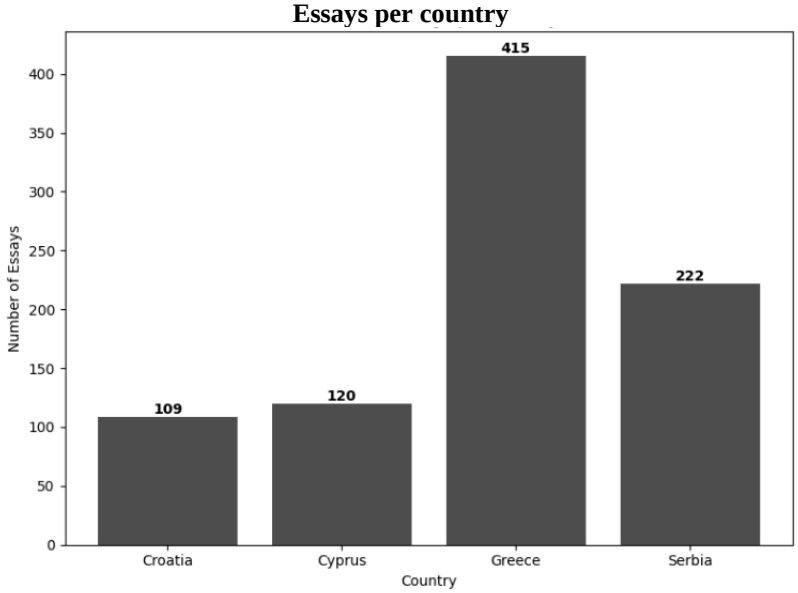
RESULTS

The first stage of essay collection resulted in a dataset of 866 student essays, gathered through coordinated efforts across the participating countries. Most submissions were received in .docx format, with a smaller number converted from .odt. Originally handwritten by students, the essays were later digitized by educators at the respective schools and uploaded to the project's cloud-based platform. Each essay was manually evaluated by teachers based on three standardized dimensions: Content (maximum 40 points), Organization (maximum 30 points), and Language (maximum 30 points). These numerical scores were subsequently mapped to four qualitative categories—Excellent, Very Good, Good, and Marginal—to facilitate interpretable, categorical classification by the AI models. All essays adhered to a consistent structural template, typically comprising an introduction, a main body, and a conclusion, which ensured

comparability across submissions and supported both human and automated analysis. This well-structured, annotated corpus serves as a representative and pedagogically grounded resource, reflecting writing diversity across countries while enabling reliable and scalable essay evaluation using retrieval-augmented large language models (LLMs).

The initial essay corpus assembled during the first stage of data collection reflects a broad geographical distribution, capturing student writing samples from four European countries participating in the AISE project. Of the 866 total essays, the majority—415 submissions—originated from Greece (specifically from DDE Prevezas), followed by 120 from Cyprus, 109 from Croatia, and the remainder from Serbia or mixed/unspecified contributors (Image 2). This cross-national diversity is essential for developing a generalizable and robust AI-based assessment framework, capable of handling variations in student writing styles, linguistic features, and educational practices. Notably, preliminary analysis revealed country-specific grading trends: for example, essays from Greece showed a higher proportion of Very Good ratings, while those from Cyprus received fewer Excellent scores across all evaluation dimensions. Such disparities underscore the importance of applying stratified sampling techniques during the train/test data split (as described in Stage 2), in order to prevent model bias and overfitting dominant patterns from a single region. Furthermore, this geographic breakdown provides a valuable foundation for future research on regional grading behaviors and opens possibilities for adapting AI-generated feedback to align with national curricula and localized writing norms.

Image 2
Essays per country



The distribution of qualitative grades across countries for the Content dimension is presented below (Image 3). The data indicate that Greece contributed the majority of essays rated as “Very Good,” with a substantial portion also falling into the “Excellent” category. Cyprus and Croatia displayed more conservative grading patterns, with fewer essays achieving the highest grade level. The relatively low number of “Marginal” grades across all countries may point to selection bias or generally high writing proficiency. These patterns highlight the need for grade-balanced sampling to ensure that the AI models do not inherit country-specific biases when learning how to assess idea relevance and development.

Each entry in the dimension-specific knowledge base (KB) consists of a student essay excerpt paired with its corresponding qualitative grade for one of the three evaluation dimensions (Content, Organization, and Language) (Image 4). This structure reflects the modular design of the proposed workflow, where a separate KB is constructed per dimension to ensure targeted retrieval during the RAG process. The example illustrates how labeled writing samples are stored in a vectorized format, allowing the system to semantically retrieve the most relevant examples during inference. These retrieved entries serve as contextual anchors, helping the LLM generate informed, transparent, and pedagogically aligned grade predictions. The AISE KB architecture supports dimension-level isolation, country-based indexing, and interpretability, making it a critical component of the pipeline's explainability and adaptability. During model inference, the test essay is embedded and compared against this vector store, retrieving the top-k most similar examples. These are then integrated into a custom prompt, specific to the evaluation dimension, and passed to the DeepSeek-V3 LLM to guide the grading process.

Image 3
Grade distribution by country and evaluation dimension

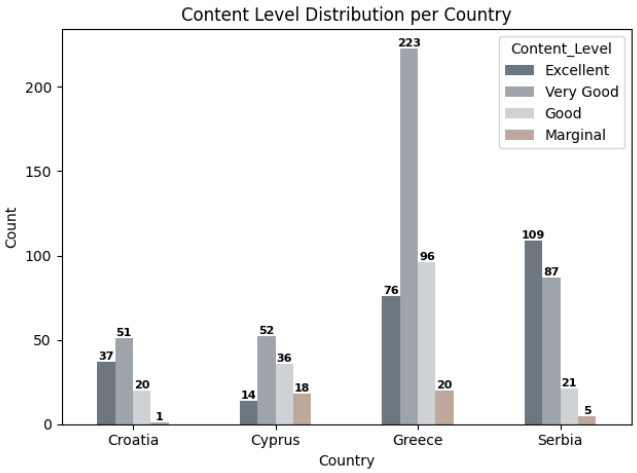
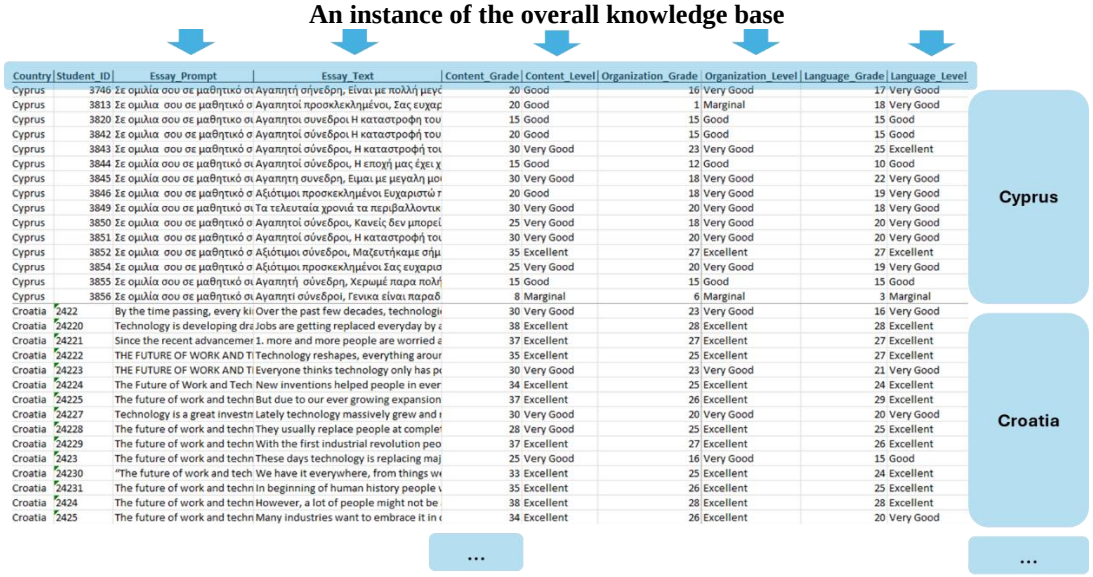


Image 4
An instance of the overall knowledge base



The evaluation of the RAG-based LLMs was conducted using a unified test set, in which each essay was independently assessed across the three evaluation dimensions—Content, Organization, and Language—by its respective dimension-specific model. Each model generated a prediction in the form of one of four qualitative grades: Excellent, Very Good, Good, or Marginal. These predictions were then compared against the grades assigned by human teachers to measure alignment.

As detailed in Table 2, the LLM assigned to the Content dimension demonstrated exceptional performance, achieving an accuracy of 0.976, macro precision of 0.988, and macro F1-score of 0.911. These results indicate a very high level of agreement with teacher evaluations. The model evaluating Organization also performed well, reaching an accuracy of 0.780 and showing strong macro recall (0.844), suggesting effective recognition of structural aspects such as coherence and logical flow.

In contrast, the model responsible for assessing Language achieved a slightly lower accuracy of 0.769, with macro recall at 0.587 and macro F1-score at 0.603. These figures reflect greater difficulty in consistently capturing nuanced linguistic features such as grammar, vocabulary precision, and stylistic variation. Such challenges may stem from increased variability across languages and countries, or from the need for more granular prompt engineering and enhanced retrieval mechanisms in this dimension.

Table 2
Performance metrics per evaluation dimension

Metric	Content	Organization	Language
Accuracy	0,976	0,780	0,769
Macro Precision	0,988	0,684	0,626
Macro Recall	0,875	0,844	0,587
Macro F1-score	0,911	0,709	0,603
Balanced Accuracy	0,875	0,844	0,587

CONLUSIONS AND FUTURE STEPS

The adopted methodology successfully integrates technical sophistication with educational applicability, ensuring that the AI-based assessment process remains both effective and pedagogically grounded. Its modular design enables targeted improvements, scalability, and future integration of more advanced or specialized language models without overhauling the entire application. While the application has demonstrated strong performance in Content evaluation, particularly in assessing idea development and relevance, the Language dimension remains more challenging. This highlights the need for ongoing enhancements in prompt design, as well as the incorporation of richer and more diverse linguistic datasets to better capture grammatical and stylistic nuances across languages. These insights will guide the next steps of development, including the refinement of the language assessment module and the continued expansion of the annotated essay corpus.

The parameter scales of several widely recognized Large Language Models (LLMs) that were evaluated during the development of the AISE pipeline, are outlined in Table 3 below. Among these, DeepSeek-V3 was selected for initial deployment due to its favorable trade-off between accuracy and computational efficiency. It offers competitive performance, fast inference times, and lower hardware requirements, making it well-suited for scalable educational applications. While more advanced models such as GPT-4 and GPT-4-Turbo have the potential to significantly enhance classification accuracy, their integration poses challenges related to increased computational costs and deployment complexity. As a result, their use is currently under consideration for future phases of the project.

Looking ahead, the project will focus on expanding the dataset, with particular emphasis on enriching the underrepresented “Marginal” category. This is expected to improve model generalization and ensure a more balanced classification across all performance levels. A critical development milestone involves the integration of the AISE essay evaluation engine into the cloud-based platform via API, enabling real-time grading and personalized feedback delivery directly to students and educators.

In parallel, further experimentation with next-generation LLMs—including GPT-4-Turbo—and parameter-efficient fine-tuning techniques is planned to improve scoring accuracy and feedback granularity. Early results already confirm the system’s capability to generate dimension-specific grade predictions and formative, pedagogically aligned suggestions, establishing a strong foundation for AI-driven personalized writing support in secondary education.

Table 3
An overview of Large Language Models (LLMs) considered for AISE essay evaluation (the selected model is shown in bold)

LLM type	Number of parameters
DeepSeek-LLM (67B)	67 billion
DeepSeek-V2	236 billion (21B activated/token)
DeepSeek-V3 (deepseek-chat)	671 billion (37B activated/token)
GPT-2	1.5 billion
GPT-3 (Davinci)	175 billion
GPT-3.5 Turbo	175 billion (optimized)
GPT-4	~1 trillion (estimated)

ACKNOWLEDGEMENTS: This research was funded by the Erasmus+ Program (State Scholarships Foundation – IKY), under the Cooperation Partnership project 2023-1-EL01-KA220-SCH-000157157.

REFERENCES

Burstein, J., Kukich, K., Wolff, S., Lu, C., & Chodorow, M. (1998). Computer analysis of essays. *In NCME symposium on automated scoring*.

Choi, Y,, & Cayce, McClenen. (2020). Development of adaptive formative assessment system using computerized adaptive testing and dynamic Bayesian networks. *Applied Sciences 10.22*: 8196.

Conijn, R., Kahr, P., & Snijders, C. C. P. (2023). The Effects of Explanations in Automated Essay Scoring Systems on Student Trust and Motivation. *Journal of Learning Analytics, 10*(1), 37-53. https://doi.org/10.18608/jla.2023.7801

Goel, Ashok K., & David A. J. (2017). Using AI to teach AI: Lessons from an online AI class. *Ai Magazine 38*.2, 48-59.

González-Calatayud, V., Prendes-Espinosa, P., & Roig-Vila, R. (2021). Artificial intelligence for student assessment: A systematic review. *Applied Sciences, 11*(12), 5467. https://doi.org/10.3390/app11125467

Grivokostopoulou, F., Perikos, I., & Hatzilygeroudis, I. (2017). An educational system for learning search algorithms and automatically assessing student performance. *International Journal of Artificial Intelligence in Education, 27*, 207–240. https://doi.org/10.1007/s40593-016-0125-2

Kumar, V., & Boulanger, D. (2020). Explainable automated essay scoring: Deep learning really has pedagogical value. *Frontiers in Education, 5*, 579125. https://doi.org/10.3389/feduc.2020.579125

Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics, 2* (2), 100050.

Ouguengay, Y. A., Nour-E. El F., & Samir B. (2015). A neuro-fuzzy inference system for the evaluation of reading/writing competencies acquisition in an e-learning environment. *Journal of Theoretical and Applied Information Technology 81*.3: 600.

Pack, A., Barrett, A., & Escalante, J. (2024). Large language models and automated essay scoring of English language learner writing: Insights into validity and reliability. *Computers and Education: Artificial Intelligence*, 100234.

Ramesh, D., & Sanampudi, S. K. (2022). An automated essay scoring systems: a systematic literature review. *Artificial Intelligence Review, 55*(3), 2495-2527.

Song, Y., Zhu, Q., Wang, H., & Zheng, Q. (2024). Automated Essay Scoring and Revising Based on Open-Source Large Language Models. *IEEE Transactions on Learning Technologies*.

Yang, Y., Xia, L., & Zhao, Q. (2019). An automated grader for Chinese essay combining shallow and deep semantic attributes. *IEEE Access, 7*, 176306-176316.

Shetty, S. V., Guruvyas, K. R., Patil, P. P., & Acharya, J. J. (2022). Essay scoring systems using AI and feature extraction: A review. In V. Bindhu, J. M. R. S. Tavares, & K. L. Du (Eds.), *Proceedings of Third International Conference on Communication, Computing and Electronics Systems* (Lecture Notes).