SvedbergOpen
DISSEMINATION OF KNOWLEDGE
International Journal of Artificial Intelligence and Machine Learning
Publisher's Home Page: https://www.svedbergopen.com/
Research Paper Open Access

A Retrieval-Augmented Large Language Model Framework for Trait-Level Automated Essay Scoring

Mr. Sunil Kumar Sahoo[1], Sandeep Mathias[2]

[1]Research Scholar, CSE, Natural Language Processing, Presidency University, Bangalore, Karnataka, 560064, ORCID ID: 0009-0009-8199-4732, Email ID:suniljeevan@gmail.com
[2]Assistant Professor, Computer Science and Engineering, Natural Language Processing, Presidency University, Bangalore, Karnataka, 560064, ORCID ID: 0000-0002-4389-5698, Email ID: sandeepalbert@presidencyuniversity.in

ABSTRACT
Automatic Essay Grading (AEG) aims to automatically evaluate student essays using natural language processing (NLP) and artificial intelligence (AI). Recent advances in transformer-based models and large language models (LLMs), including BERT and GPT, have significantly improved automated essay scoring by capturing rich semantic and contextual information. This paper proposes a Retrieval-Augmented Large Language Model (RAG-LLM) framework for trait-level automated essay scoring and evaluates its performance against LSTM-, BERT-, and GPT-based approaches. The proposed framework evaluates multiple writing traits, including Content, Organization, Word Choice, Sentence Fluency, Conventions, Prompt Adherence, Language, Narrativity, Style, and Voice. Experiments conducted on the ASAP dataset demonstrate that the proposed RAG-LLM framework achieves the highest agreement with human raters, outperforming fine-tuned BERT multi-head models and GPT-based prompting approaches across most essay prompts and evaluation traits. Comprehensive prompt-wise, trait-wise, and ablation analysis further demonstrate the robustness of the proposed framework and provide insights into prompt-specific performance variation and the contribution of key architectural components. The results indicate that integrating semantic retrieval with large language model reasoning provides an effective and scalable solution for trait-level automated essay scoring.

Keywords: Automatic Essay Grading, Trait-Level Essay Scoring, Retrieval-Augmented Generation, Large Language Models, BERT, GPT, Educational Assessment

INTRODUCTION

The process of essay evaluation has traditionally been labor-intensive, subjective, and time-consuming, often leading to inconsistencies across human graders. With the growing demand for scalable and fair assessment systems, Automatic Essay Grading (AEG) has emerged as a promising solution by utilizing machine learning and natural language processing (NLP) (Page 1966; Taghipour and Ng 2016) to automate the evaluation of student writing. Early AEG systems focused on surface-level features such as word count, spelling accuracy, and grammar (Page 1966; Burstein 2003), but advancements in deep learning have enabled a shift toward more sophisticated, semantic analysis of writing quality. Modern AEG systems employ deep learning architectures like Long Short-Term Memory (LSTM) networks and Transformer-based models (Hochreiter and Schmidhuber 1997; Taghipour and Ng 2016) to better understand the structure and meaning of essays. LSTM networks have been widely used for sequential data processing and were among the first models capable of capturing contextual dependencies in text (Hochreiter and Schmidhuber 1997). However, their ability to model long-range dependencies is limited compared to more recent architectures. The introduction of Transformers and attention mechanisms (Vaswani et al. 2017) significantly improved essay evaluation by allowing models to focus on important parts of an essay regardless of their position in the text. This gave rise to pretrained language models like BERT, RoBERTa, and GPT (Devlin et al. 2019; Liu et al. 2019; Brown et al. 2020), which excel at extracting contextual meaning and have revolutionized the field of automated scoring.

2. Background

The field of Automatic Essay Grading (AEG) has undergone significant evolution with the advancement of

NLP and deep learning. This section outlines the historical and technological foundations that have shaped modern AEG systems.

2.1 Evolution of Automated Essay Grading

AEG has a history dating back to the 1960s, beginning with systems like Project Essay Grade (PEG), one of the earliest attempts to automate essay evaluation using statistical features (Page 1966). Early systems relied on handcrafted features such as word count, sentence length, and grammar checks, and other surface-level linguistic characteristics. In the 2000s, systems like e-rater and Intelligent Essay Assessor (IEA) incorporated more sophisticated linguistic and semantic analysis techniques (Burstein 2003; Landauer et al. 2003). In particular, Latent Semantic Analysis (LSA) enabled automated systems to evaluate semantic similarity and content relevance, significantly improving essay assessment beyond purely surface-level features (Landauer et al. 1998).

2.2 Deep Learning in AEG

The adoption of deep learning transformed AEG systems by reducing dependence on manual feature engineering and enabling automatic extraction of meaningful textual representations (Taghipour and Ng 2016). LSTM networks became widely used due to their ability to model long-range dependencies in essays and capture contextual information from sequential text (Hochreiter and Schmidhuber 1997; Taghipour and Ng 2016). Later, Transformer architectures introduced self-attention mechanisms that significantly improved contextual understanding and representation learning (Vaswani et al. 2017). Building upon these advances, pretrained language models such as BERT achieved state-of-the-art performance across a wide range of natural language processing tasks, including automated essay grading (Devlin et al. 2019).

2.3 Large Language Model and Essay Trait Scoring

Modern AEG systems increasingly leverage pretrained language models (LLMs) to predict both holistic and trait-level essay scores (Devlin et al. 2019; Brown et al. 2020). These models can evaluate multiple dimensions of writing quality, including Organization, Grammar, Vocabulary Usage, Content Development, and Coherence, by learning rich contextual representations from large-scale text corpora. Fine-tuned BERT-based models have demonstrated strong performance for trait-specific essay scoring tasks, while prompt-based GPT models provide a flexible alternative that can generate rubric-aware evaluations without task-specific training (Li et al. 2020; Brown et al. 2020). The combination of transformer-based representation learning and natural language prompting has enabled more interpretable and fine-grained assessment of student writing across multiple scoring traits.

3. Related Work

Automated Essay Grading (AEG) has been an active area of research for several decades, with approaches evolving from handcrafted feature engineering to deep neural architectures and, more recently, large language models. Early essay scoring systems relied heavily on manually designed linguistic features such as word count, sentence length, spelling accuracy, grammatical correctness, and readability measures. One of the earliest systems, Project Essay Grade (PEG), demonstrated that statistical features extracted from essays could be used to approximate human grading behavior. Subsequently, systems such as e-rater and Intelligent Essay Assessor (IEA) incorporated richer linguistic and semantic features, improving the quality of automated scoring.

The introduction of machine learning techniques significantly advanced automated essay grading. Traditional machine learning approaches employed handcrafted lexical, syntactic, and discourse features combined with regression or classification algorithms. While these approaches achieved reasonable performance, their effectiveness depended heavily on extensive feature engineering and domain expertise. Furthermore, handcrafted features often struggled to capture deeper semantic relationships and discourse structures present in essays.

The emergence of deep learning reduced the reliance on manual feature engineering. Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks, became widely used for essay scoring because of their ability to model sequential dependencies in text. Taghipour and Ng (Taghipour and Ng 2016) demonstrated that neural architectures could effectively learn essay representations directly from raw text, achieving competitive performance without manually designed linguistic features. Attention mechanisms further improved neural essay scoring by enabling models to focus on important portions of essays during prediction.

Transformer-based language models have recently become the dominant paradigm in natural language processing. The introduction of BERT by Devlin et al. (Devlin et al. 2019) enabled contextualized text representations that capture bidirectional relationships between words. Several studies have demonstrated that BERT-based models significantly outperform traditional neural architectures in automated essay scoring tasks. By leveraging large-scale pretraining, BERT can capture semantic, syntactic, and discourse-level information relevant to essay quality. Fine-tuning BERT on essay scoring datasets has resulted in substantial improvements in agreement with human raters.

Beyond holistic essay scoring, researchers have increasingly focused on trait-level evaluation. Trait-based scoring provides more detailed feedback by assigning separate scores to dimensions such as content, organization, coherence, grammar, vocabulary usage, and prompt adherence. Mathias and Bhattacharyya (Mathias and Bhattacharyya 2018) introduced one of the first large-scale studies on trait-specific essay scoring using the ASAP dataset. Their work highlighted the importance of evaluating multiple writing dimensions independently rather than relying solely on a holistic score. Subsequent studies further explored multi-task learning frameworks that jointly predict multiple essay traits.

More recently, Large Language Models (LLMs) such as GPT have demonstrated strong capabilities in educational assessment tasks. Unlike supervised neural models that require task-specific training, GPT-based systems can evaluate essays through prompt engineering. In a zero-shot setting, the model is provided with only the scoring rubric and essay text, whereas few-shot prompting additionally supplies example essays and corresponding scores. Brown et al. (Brown et al. 2020) showed that large language models can perform a wide variety of tasks through in-context learning without gradient-based optimization. These capabilities have motivated growing interest in applying LLMs to automated essay grading.

Despite their effectiveness, prompt-based LLM approaches may suffer from calibration issues because they do not explicitly learn the scoring distribution of a target dataset. To address this limitation, Retrieval-Augmented Generation (RAG) frameworks have recently emerged as a promising solution. RAG systems combine semantic retrieval with generative reasoning by supplying relevant examples to the language model during inference. In educational assessment, retrieved essays with known scores can serve as anchor examples that guide scoring decisions and improve consistency. By grounding the language model with human-scored essays, RAG-based approaches can reduce scoring variability and improve interpretability.

Motivated by these developments, this work investigates and compares multiple generations of automated essay grading techniques, including LSTM-based neural models, BERT-based transformer architectures, prompt-based GPT scoring, and a Retrieval-Augmented Generation framework. The study focuses on trait-level essay evaluation and analyzes the effectiveness of each approach using standard automated essay scoring benchmarks and evaluation metrics.

4. Methodology

This section presents the overall methodology adopted for developing and evaluating the proposed Automated Essay Grading (AEG) framework. Figure 1 presents the overall architecture of the proposed automated essay grading system integrating BERT-based trait prediction, GPT prompt-based scoring, and retrieval-augmented generation.

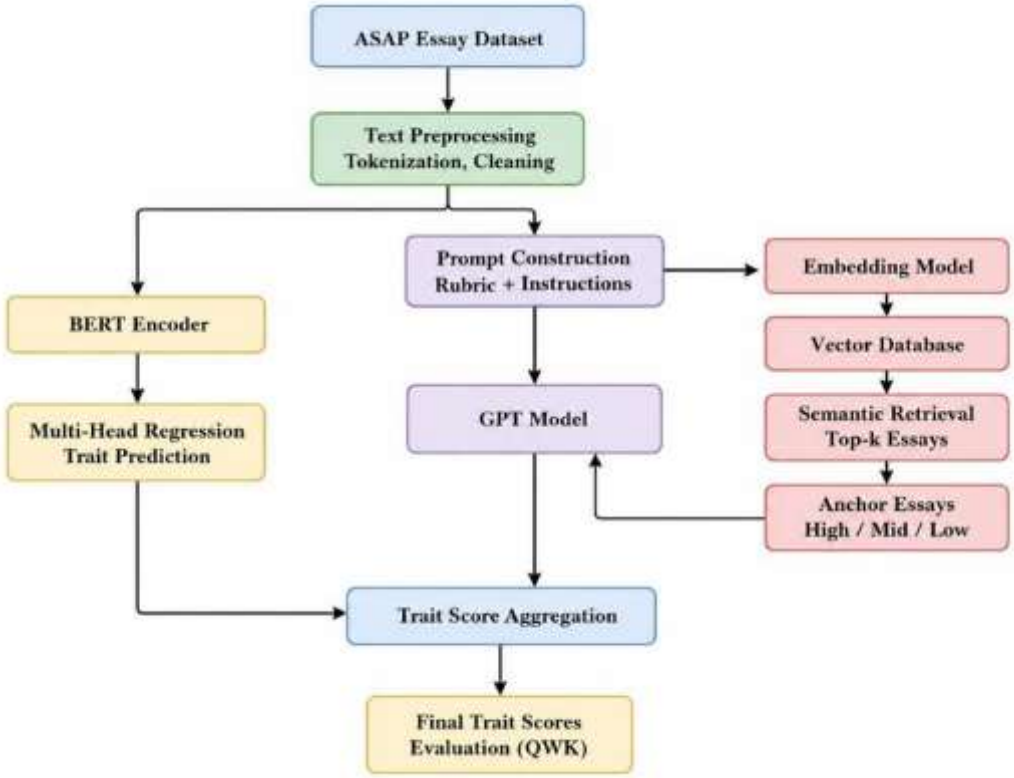

Figure 1. Unified architecture of the proposed automated essay grading system integrating BERT-based trait prediction, GPT prompt-based scoring, and retrieval-augmented generation (RAG) for anchor essay selection.

As shown in Figure 1, essays from the ASAP dataset (Shermis and Hamner 2012) are first preprocessed to remove inconsistencies and prepare the text for model input. A five-fold cross-validation strategy is then employed to ensure robust evaluation and reduce sampling bias. The resulting trait-level predictions are compared against human-assigned scores using QWK and other statistical metrics. This experimental design enables a fair comparison of traditional neural models, transformer-based architectures, prompt-based large language models, and retrieval-augmented scoring frameworks.

4.1 Dataset

To evaluate the effectiveness of transformer-based and large language model approaches for automated essay grading, we use the Automated Student Assessment Prize (ASAP) dataset (Shermis and Hamner 2012), one of the most widely used benchmarks in essay scoring research. The dataset contains essays collected from students responding to eight different writing prompts and includes human-assigned scores for both holistic and trait-level evaluation.

The ASAP dataset consists of essays spanning multiple genres, including argumentative, persuasive, source-dependent, narrative, and descriptive writing tasks. Each prompt is associated with its own scoring rubric and score range, making the dataset suitable for evaluating prompt-aware scoring systems. The dataset has been extensively used in previous automated essay scoring studies, enabling direct comparison with existing approaches (Taghipour and Ng 2016).

For trait-level scoring experiments, we utilize the extended trait annotations introduced by Mathias and Bhattacharyya (Mathias and Bhattacharyya 2018). These annotations provide fine-grained scores for multiple writing dimensions, allowing the evaluation of individual essay traits such as content, organization, language use, narrativity, style, and prompt adherence.

The dataset is stored in a structured format containing the essay text, prompt identifier, human-assigned scores, and trait-level annotations. The primary fields used in our experiments include:

- **essay_id**: Unique identifier assigned to each essay.
- **essay_set**: Prompt identifier corresponding to the writing task.
- **essay**: Raw essay text written by the student.
- **domain1_score**: Primary holistic score assigned by human raters.
- **Trait Scores**: Human-assigned scores corresponding to prompt-specific evaluation traits.

The dataset provides a realistic benchmark for automated essay grading because it contains essays of varying lengths, writing styles, and quality levels. Furthermore, the availability of multiple prompts and trait annotations makes it particularly suitable for evaluating trait-level scoring architectures.

4.2 Preprocessing

Prior to model training, several preprocessing steps are applied to ensure data consistency and improve model performance. Since the ASAP dataset is relatively clean, only light preprocessing is performed in order to preserve the original writing characteristics of student essays.

First, essays are normalized by removing redundant white spaces and correcting formatting inconsistencies. Non-printable characters and invalid symbols are removed while preserving punctuation and sentence structure. This step ensures that the input remains compatible with transformer-based tokenizers (Devlin et al. 2019).

Second, essays are grouped according to their corresponding prompt identifiers. Since each prompt follows a different scoring rubric and score range, prompt-aware grouping enables more reliable training and evaluation. During cross-validation, prompt distributions are preserved to maintain consistency across folds.

Third, trait-level annotations are prepared for multi-output prediction tasks. Depending on the prompt, different sets of traits are available. These trait scores are extracted and aligned with their corresponding essays to facilitate multi-head regression training.

To improve training stability, score normalization is performed. Scores are mapped to a common numerical range before training and later transformed back to their original scale during evaluation. This normalization reduces scale variation across prompts and improves convergence during optimization.

Finally, the processed essays are converted into the format required by transformer-based models. Each essay is paired with its associated prompt identifier, trait annotations, and normalized scores, forming the final dataset used for training and evaluation.

4.3 Tokenization

The preprocessed essays are converted into token-level representations using the BERT tokenizer. We employ the Bert-base-uncased tokenizer provided by the HuggingFace Transformers library following the tokenization strategy used in BERT (Devlin et al. 2019). The tokenizer converts raw essay text into subword tokens using WordPiece tokenization (Wu et al. 2016), enabling efficient handling of rare and unseen words.

For each essay, the tokenizer generates the following components:

- **Input IDs**: Numerical identifiers corresponding to tokenized words.
- **Attention Masks**: Binary masks indicating valid tokens and padded positions.
- **Special Tokens**: The [CLS] token is added at the beginning of each essay and the [SEP] token is appended at the end.

Since essay lengths vary considerably across prompts, padding and truncation are applied to maintain a fixed sequence length of 512 tokens. Essays longer than the maximum length are truncated, while shorter essays are padded with special padding tokens.

The resulting tokenized representations are then provided as input to the transformer encoder during training and inference. The contextual representation corresponding to the [CLS] token is subsequently used for holistic and trait-level score prediction.

4.4 BERT-Based Trait-Level Essay Scoring

To establish a strong supervised baseline for trait-level essay grading, we employ a BERT-based multi-head regression architecture. BERT (Bidirectional Encoder Representations from Transformers) has demonstrated strong performance across a wide range of natural language processing tasks due to its ability to learn contextualized representations from large-scale pretraining based on the Transformer architecture (Vaswani et al. 2017; Devlin et al. 2019).

In the proposed framework, each essay is first tokenized using the bert-base-uncased tokenizer. Special tokens [CLS] and [SEP] are added to the input sequence, and essays are padded or truncated to a maximum sequence length of 512 tokens. The resulting token IDs and attention masks are provided as input to the BERT encoder.

The contextual representation corresponding to the [CLS] token is used as a global representation of the essay. This representation captures semantic, syntactic, and discourse-level information present throughout the essay.

Unlike holistic essay scoring models that predict a single score, our approach employs multiple regression heads connected to the [CLS] representation. Each regression head corresponds to a specific essay trait and independently predicts the score associated with that trait. This multi-task learning setup allows the model to jointly learn multiple writing dimensions while sharing a common semantic representation (Caruana 1997).

Formally, let h_{CLS} denote the contextual embedding generated by BERT. For each trait t, a dedicated regression layer computes:

$$\hat{y}_t = W_t h_{\text{CLS}} + b_t \quad (1)$$

where $\hat{y}_t$ represents the predicted score for trait t, and W_t and b_t are trainable parameters.

The model is trained using supervised learning, where human-assigned trait scores serve as target labels. Mean Squared Error (MSE) loss is computed for each trait and combined into a joint optimization objective:

$$L = \sum_{t=1}^{T} L_t \quad (2)$$

where T denotes the number of predicted traits.

To ensure robust evaluation, five-fold cross-validation is employed. During each fold, essays from four partitions are used for training while the remaining partition is used for testing. Model performance is evaluated using Quadratic Weighted Kappa (QWK) metric (Cohen 1968), Pearson Correlation, Spearman Correlation, and Mean Squared Error.

The overall trait-level scoring workflow is illustrated in Figure 2.

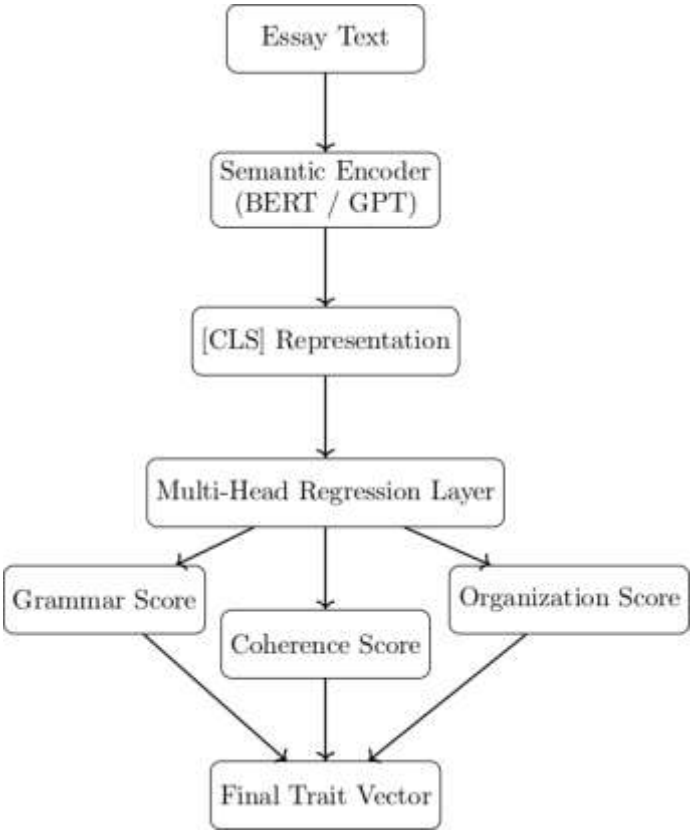

Figure 2. Trait-level automated essay scoring architecture. The essay is encoded using a semantic encoder (e.g., BERT or GPT). The resulting representation is processed through a multi-head regression layer to predict individual trait scores such as grammar, coherence and organization.

The overall procedure of BERT-based trait-level essay scoring is summarized in Algorithm 1. Model performance is evaluated using the Quadratic Weighted Kappa (QWK) metric, which measures the agreement between predicted scores and human annotations.

Algorithm 1. BERT-Based Trait-Level Essay Scoring

Input: Essay dataset, number of folds K, human-assigned trait scores
Output: Predicted trait scores for each essay

1. Split the dataset into K folds.
2. For each fold i:
    - Set fold i as the **Test Set**.
    - Set all remaining folds as the **Training Set**.
    - Tokenize essays using the BERT tokenizer.
    - Convert tokens into input IDs and attention masks.
    - Initialize the pretrained BERT model.
    - Add regression layers for trait prediction.
    - Train the model on the Training Set using human-assigned trait scores as targets.
    - Encode essays in the Test Set using the fine-tuned BERT model.
    - Predict trait scores using the **Multi-Head** regression layers.
    - Compute the Quadratic Weighted Kappa (QWK).
3. Aggregate the evaluation results across all folds.

4.5 GPT Prompt-Based Essay Scoring

In addition to supervised transformer-based models, we investigate the use of Large Language Models (LLMs) for automated essay grading through prompt-based evaluation. Unlike BERT-based approaches that require task-specific training, GPT models can perform scoring directly through natural language instructions (Brown et al. 2020).

The prompt-based framework combines the essay text, scoring rubric, and evaluation instructions into a structured prompt. The prompt is then submitted to the GPT model, which generates trait-level scores based on its understanding of the essay and the provided grading criteria following the principle of in-context learning (Brown et al. 2020).

Two prompting strategies are explored in this study: zero-shot prompting and few-shot prompting (Brown et al. 2020):

4.5.1 Zero-Shot Prompting

In the zero-shot setting, the language model is provided only with the essay text and the scoring rubric, without any example essays (Brown et al. 2020). No example essays are included in the prompt. The model is instructed to behave as an expert essay evaluator and assign scores according to the rubric guidelines.

The prompt contains:

- Essay text
- Trait definitions
- Scoring scale
- Evaluation instructions

Based on this information, the model predicts scores for the required traits.

4.5.2 Few-Shot Prompting

To improve scoring consistency, few-shot prompting is also employed. In this setting, the prompt includes a small number of example essays along with their corresponding human-assigned scores (Brown et al. 2020; Liu et al. 2023), following the in-context learning paradigm introduced by large language models (Brown et al. 2020).

The examples serve as reference points that help the model understand the expected scoring distribution. By comparing the target essay with these examples, the model can generate more calibrated trait-level predictions.

The few-shot prompt contains:

- Trait scoring rubric
- Example essays with known scores
- Target essay
- Scoring instructions
-

4.5.3 Structured Trait Prediction

For automated evaluation, the GPT model is instructed to return scores in a structured JSON format. Each generated response contains predicted scores for all traits associated with the corresponding prompt, following the instruction-based generation capabilities of large language models (Brown et al. 2020).

An example output structure is shown below:

```
{
 "Content": 4,
 "Organization": 3,
 "Language": 4,
 "Narrativity": 5
}
```

The generated scores are extracted and compared with human annotations using standard evaluation metrics such as Quadratic Weighted Kappa (QWK) (Cohen 1968), Pearson Correlation, Spearman Correlation, and Mean Squared Error (MSE).

Figure 3. Prompt-based essay scoring framework using GPT. The essay and rubric are combined into a structured prompt that is processed by the language model to generate trait-level scores.

Figure 3 illustrates the overall GPT-based scoring framework, while Algorithm 2 summarizes the complete scoring procedure.

Algorithm 2. GPT-Based Trait-Level Essay Scoring

Input: Essay dataset, number of folds K, human-assigned trait scores
Output: Predicted trait scores for each essay

1. Split the dataset into K folds.
2. For each fold i:
   - Set fold i as the **Test Set**.
   - Set all remaining folds as the **Training Set**.
   - Retrieve scoring rubric for corresponding prompt.
   - Construct prompt containing: Essay text Trait scoring rubric Instructions for expert grading.
   - Send prompt to GPT model.
   - Receive response containing trait scores.
   - Extract predicted scores from JSON output.
   - Compute evaluation metric (QWK).
3. Aggregate the evaluation results across all folds.

4.6 RAG-Based Trait-Level Essay Scoring

Although GPT-based scoring demonstrates strong reasoning capabilities, it may suffer from calibration drift when applied to unseen essay datasets (Brown et al. 2020). Since the model has not directly learned the scoring distribution of the target dataset, generated scores may not always align with human grading standards.

To address this limitation, we propose a Retrieval-Augmented Generation (RAG) framework that combines semantic retrieval with large language model reasoning (Lewis et al. 2020).

The proposed architecture consists of four major components:
1. Embedding Model
2. Vector Database
3. Semantic Retrieval Module
4. LLM Scoring Module

4.6.1 Embedding Generation

Each essay is converted into a dense vector representation using a pretrained embedding model (Reimers and Gurevych 2019). The embeddings capture semantic characteristics of the essay and enable similarity-based retrieval.
Let

$$e_i \in \mathbb{R}^d \quad (3)$$

represent the embedding vector corresponding to essay i.

The generated embeddings are stored in a vector database for efficient retrieval.

4.6.2 Vector Database Construction

The retrieval bank is constructed from the training folds during each cross-validation iteration. Vector indexing techniques such as FAISS or ChromaDB are used to enable efficient nearest-neighbor search (Johnson et al. 2019).

Only essays belonging to the same prompt are considered during retrieval to ensure prompt-specific scoring consistency.

4.6.3 Semantic Retrieval

For a target essay, its embedding is generated and compared against all embeddings stored in the retrieval database using dense vector retrieval techniques (Reimers and Gurevych 2019).

Cosine similarity is used to identify semantically related essays:

$$\text{Similarity}(A, B) = \frac{A \cdot B}{\|A\|\|B\|} \quad (4)$$

The top-K most similar essays are retrieved and ranked according to similarity scores.

4.6.4 Anchor Selection

From the retrieved essays, three representative anchor essays are selected:

- High-score anchor
- Mid-score anchor
- Low-score anchor

These anchor essays provide the language model with concrete examples of different quality levels and help calibrate the grading process. By grounding score predictions using retrieved examples, the framework reduces scoring variability and improves consistency with human judgments (Lewis et al. 2020).

4.6.5 Comparative Reasoning

The retrieved anchor essays are combined with the rubric and target essay to form a structured prompt.
The prompt contains:

- Trait scoring rubric
- Low-score anchor essay
- Mid-score anchor essay
- High-score anchor essay
- Target essay

The language model compares the target essay against the anchor examples and generates trait-level scores through comparative reasoning. By conditioning score generation on retrieved examples, the framework follows the retrieval-augmented generation paradigm and enables more calibrated scoring decisions (Lewis et al. 2020).

4.6.6 Trait Prediction

The generated trait scores are returned in structured JSON format and subsequently evaluated against human annotations.
The proposed RAG framework combines the contextual reasoning capabilities of LLMs with dataset-specific calibration obtained through retrieval, resulting in more reliable and interpretable essay scoring (Lewis et al. 2020).
Figure 4 illustrates the complete RAG-based essay scoring framework, where semantically similar essays retrieved from the vector database serve as anchor examples for improving trait-level prediction., while Algorithm **3** summarizes the complete retrieval and scoring procedure adopted by the proposed RAG

framework. By incorporating semantically similar essays into the prompt, the language model is better aligned with human scoring patterns, resulting in improved agreement with expert annotations (Lewis et al. 2020; Brown et al. 2020).

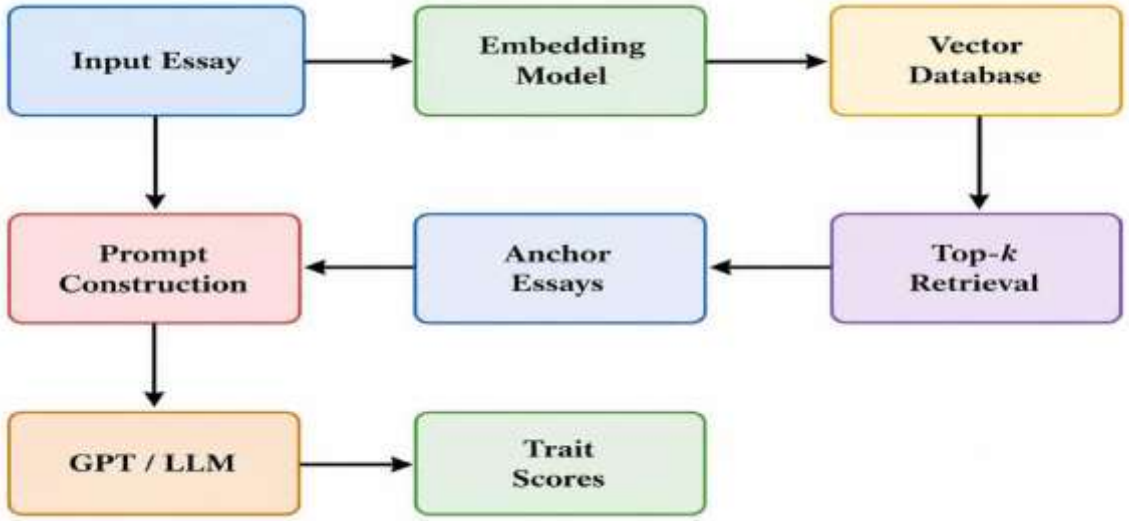

Figure 4. Architecture of the proposed Retrieval-Augmented Generation (RAG)-based trait-level essay scoring framework. The language model utilizes retrieved semantically similar essays as contextual references during prediction.

Algorithm 3. RAG-Based Trait-Level Essay Scoring

Input: Essay dataset (D), vector database (V), number of retrieved essays (k), human-assigned trait scores
Output: Predicted trait scores for each essay and QWK evaluation scores.

1. Generate embeddings for the essays in dataset (D).
2. Store the essay embeddings and corresponding essay representations in the vector database (V).
3. For each essay (e) in the dataset:
    - Generate the embedding representation of (e).
    - Retrieve the Top-(k) semantically similar essays from (V) based on embedding similarity.
    - Select representative anchor essays from the retrieved essays.
    - Construct a prompt containing the input essay (e), the scoring rubric, and the retrieved anchor essays.
    - Send the constructed prompt to GPT for trait-level essay scoring.
    - Extract the predicted trait scores from the GPT response.
4. Compare the predicted trait scores with the human-assigned trait scores.
5. Compute the Quadratic Weighted Kappa (QWK) for each trait.
6. Aggregate the trait-level QWK scores to evaluate the overall performance of the RAG-based essay scoring framework.

4.7 Training and Fine-Tuning

The BERT-based models are trained using supervised learning, where human-assigned trait scores serve as target labels (Devlin et al. 2019). The model parameters are optimized through backpropagation using the AdamW optimizer (Loshchilov and Hutter 2019).

The primary training configuration is summarized below:

- Optimizer: AdamW
- Learning Rate: 2×10−5
- Batch Size: 16

- Epochs: 3–5
- Maximum Sequence Length: 512
- Validation Strategy: 5-Fold Cross Validation

For trait-level prediction, Mean Squared Error (MSE) loss is computed independently for each regression head and combined into a joint optimization objective.

Dropout regularization is applied to reduce overfitting and improve generalization performance.

GPT-based and RAG-based models do not require gradient-based training. Instead, they operate through prompt engineering and retrieval-enhanced inference. The same evaluation protocol is applied across all approaches to ensure a fair comparison.

To obtain robust performance estimates, five-fold cross-validation is employed. During each fold, the dataset is divided into training, validation, and testing partitions. The final results are reported as the average performance across all folds.

Model performance is evaluated using multiple metrics including Quadratic Weighted Kappa (QWK), Pearson Correlation, Spearman Correlation, and Mean Squared Error (MSE). Among these metrics, QWK serves as the primary evaluation measure because it is widely accepted as the standard metric for automated essay scoring tasks (Cohen 1968; Taghipour and Ng 2016).

5.Experiments and Evaluation

5.1 Experimental Setup

To evaluate the effectiveness of the proposed automated essay grading framework, experiments were conducted on the Automated Student Assessment Prize (ASAP) dataset (Shermis and Hamner 2012). The dataset contains essays from eight different prompts covering argumentative, persuasive, source-dependent, and narrative writing tasks.

Following prior work in automated essay scoring (Taghipour and Ng 2016), a five-fold cross-validation strategy was adopted. The dataset was divided into five folds, where one fold was used for testing and the remaining folds were used for training and validation. This process was repeated five times so that each essay was evaluated exactly once as part of the test set.

For supervised models, essays were further divided into training, validation, and testing subsets using a 60%-20%-20% split within each fold. The validation set was used for model selection and hyperparameter tuning, while the test set was reserved exclusively for final evaluation.

All essays were tokenized using the BERT tokenizer (Devlin et al. 2019) and truncated or padded to a maximum sequence length of 512 tokens. Trait-level scores were used as prediction targets for multi-head regression models.
Experiments were conducted using Python 3.10, PyTorch 2.0, and the HuggingFace Transformers library. Model training was performed on an NVIDIA Tesla T4 GPU with CUDA support.

5.2 Model Configurations

Multiple neural and language-model-based architectures were evaluated in order to compare different approaches for automated essay grading.

LSTM + Attention Baseline

A Bidirectional Long Short-Term Memory (BiLSTM) network with an attention mechanism was used as a baseline model (Hochreiter and Schmidhuber 1997; Bahdanau et al. 2015). Essays were represented using pre-trained GloVe embeddings (Pennington et al. 2014) and processed through stacked BiLSTM layers. The attention mechanism was employed to identify important words and sentences for score prediction.

BERT + Regression Head

The baseline transformer model utilized bert-base-uncased. The final hidden representation

corresponding to the [CLS] token was passed through a fully connected regression layer to predict a single holistic essay score.

BERT + Multi-Head Regression

To support trait-level essay scoring, multiple regression heads were attached to the [CLS] representation. Each regression head was responsible for predicting a specific trait score such as content, organization, coherence, vocabulary, or grammar. All regression heads were trained jointly using a multi-task learning framework.

GPT Prompt-Based Scoring

Large language models were evaluated using both zero-shot and few-shot prompting strategies. Essays, grading rubrics, and evaluation instructions were combined into structured prompts and submitted to the language model. The model generated trait-level scores without any gradient-based fine-tuning.

RAG-Based Scoring

The proposed Retrieval-Augmented Generation (RAG) framework combined semantic retrieval with language-model reasoning (Lewis et al. 2020). Similar essays were retrieved from a vector database and provided as anchor examples to the language model (Reimers and Gurevych 2019). These anchors helped calibrate scoring decisions and improve agreement with human raters.

5.3 Evaluation Metrics

Model performance was evaluated using several commonly adopted metrics in automated essay scoring research.

- **Quadratic Weighted Kappa (QWK)**: The primary evaluation metric used in this study. QWK measures the agreement between predicted scores and human-assigned scores while accounting for the ordinal nature of essay grades (Cohen 1968).
- **Mean Squared Error (MSE)**: Measures the average squared difference between predicted and actual scores. Lower values indicate better prediction accuracy.
- **Pearson Correlation Coefficient**: Evaluates the linear relationship between model predictions and human scores (Pearson 1895).
- **Spearman Rank Correlation**: Measures the monotonic relationship between predicted and actual rankings of essays (Spearman 1904).
- **Exact Match Accuracy**: Represents the percentage of essays for which the predicted score exactly matches the human-assigned score.

Among these metrics, QWK is considered the standard benchmark in automated essay scoring because it closely reflects agreement between automated systems and human evaluators and has been widely adopted in prior AES studies (Cohen 1968; Taghipour and Ng 2016).

6. Results and Discussion

6.1 Trait-Level Results

The proposed trait-level essay scoring framework was evaluated on the ASAP dataset using Quadratic Weighted Kappa (QWK), which is the standard metric for automated essay scoring tasks (Cohen 1968; Taghipour and Ng 2016). Performance was assessed across all prompts and traits and compared with several previously published approaches, including the systems proposed by Mathias and Bhattacharyya (Mathias and Bhattacharyya 2018), Cozma et al. (Cozma et al. 2018), and Dong et al. (Dong et al. 2017).

Since the ASAP dataset contains different trait definitions for different prompts, not all traits are available across all essay sets. Consequently, unavailable trait–prompt combinations are denoted using "–" in the results table.

Table 1 presents the trait-wise QWK scores obtained by different systems. Higher QWK values indicate stronger agreement between automated predictions and human-assigned scores.

Table 1.Trait-wise Quadratic Weighted Kappa (QWK) scores for different systems across all ASAP prompts.

Prompt	System	Cont.	Org.	WC	SF	Conv.	PA	Lang.	Narr.	Style	Voice
Prompt 1	ACL 2018	0.686	0.637	0.659	0.639	0.620	--	--	--	--	--
	CoNLL 2017	0.703	0.664	0.675	0.648	0.638	--	--	--	--	--
	BERT	**0.822**	**0.816**	**0.834**	**0.821**	**0.828**	--	--	--	--	--
	GPT	**0.835**	**0.829**	**0.846**	**0.834**	**0.841**	--	--	--	--	--
	RAG	**0.851**	**0.844**	**0.861**	**0.849**	**0.857**	--	--	--	--	--
Prompt 2	ACL 2018	0.600	0.570	0.583	0.544	0.530	--	--	--	--	--
	CoNLL 2017	0.617	0.623	0.630	0.603	0.601	--	--	--	--	--
	BERT	**0.820**	**0.802**	**0.8123**	**0.821**	**0.805**	--	--	--	--	--
	GPT	**0.832**	**0.816**	**0.828**	**0.836**	**0.820**	--	--	--	--	--
	RAG	**0.848**	**0.832**	**0.844**	**0.851**	**0.837**	--	--	--	--	--
Prompt 3	ACL 2018	0.659	--	--	--	--	0.658	0.590	0.645	--	--
	CoNLL 2017	0.673	--	--	--	--	0.683	0.612	0.684	--	--
	BERT	**0.820**	**--**	**--**	**--**	**--**	**0.870**	**0.864**	**0.850**	**--**	**--**
	GPT	**0.833**	**--**	**--**	**--**	**--**	**0.882**	**0.876**	**0.864**	**--**	**--**
	RAG	**0.851**	**--**	**--**	**--**	**--**	**0.897**	**0.891**	**0.881**	**--**	**--**
Prompt 4	ACL 2018	0.702	--	--	--	--	0.702	0.571	0.687	--	--
	CoNLL 2017	0.751	--	--	--	--	0.738	0.645	0.722	--	--

	BERT	**0.846**	--	--	--	--	**0.894**	**0.893**	**0.898**	--	--
	GPT	**0.859**	--	--	--	--	**0.907**	**0.906**	**0.911**	--	--
	RAG	**0.874**	--	--	--	--	**0.921**	**0.920**	**0.925**	--	--
Prompt 5	ACL 2018	0.713	--	--	--	--	0.700	0.620	0.635	--	--
	CoNLL 2017	0.738	--	--	--	--	0.719	0.638	0.700	--	--
	BERT	**0.880**	--	--	--	--	**0.847**	**0.851**	**0.863**	--	--
	GPT	**0.892**	--	--	--	--	**0.861**	**0.866**	**0.878**	--	--
	RAG	**0.908**	--	--	--	--	**0.877**	**0.882**	**0.894**	--	--
Prompt 6	ACL 2018	0.759	--	--	--	--	0.711	0.624	0.635	--	--
	CoNLL 2017	0.820	--	--	--	--	0.783	0.664	0.690	--	--
	BERT	**0.577**	--	--	--	--	**0.569**	**0.551**	**0.575**	--	--
	GPT	**0.603**	--	--	--	--	**0.592**	**0.578**	**0.601**	--	--
	RAG	**0.631**	--	--	--	--	**0.620**	**0.604**	**0.629**	--	--
Prompt 7	ACL 2018	0.737	0.659	--	--	0.504	--	--	--	0.609	--
	CoNLL 2017	0.771	0.676	--	--	0.621	--	--	--	0.659	--
	BERT	**0.869**	**0.806**	--	--	**0.713**	--	--	--	**0.695**	--
	GPT	**0.883**	**0.821**	--	--	**0.728**	--	--	--	**0.712**	--
	RAG	**0.898**	**0.837**	--	--	**0.744**	--	--	--	**0.729**	--
Prompt 8	ACL 2018	0.573	0.572	0.494	0.477	0.455	--	--	--	--	0.489
	CoN	0.5	0.6	0.5	0.5	0.5	--	--	--	--	0.5

	LL 2017	86	32	59	86	58					44
	BERT	**0.773**	**0.777**	**0.748**	**0.748**	**0.771**	**--**	**--**	**--**	**--**	**0.768**
	GPT	**0.789**	**0.791**	**0.764**	**0.767**	**0.786**	**--**	**--**	**--**	**--**	**0.768**
	RAG	**0.808**	**0.811**	**0.783**	**0.785**	**0.804**	**--**	**--**	**--**	**--**	**0.798**
Mean QWK	ACL 2018	0.679	0.610	0.579	0.553	0.527	0.693	0.601	0.651	0.609	0.489
	CoNLL 2017	0.707	0.649	0.621	0.612	0.605	0.731	0.640	0.699	0.659	0.544
	BERT	**0.801**	**0.800**	**0.798**	**0.798**	**0.779**	**0.795**	**0.790**	**0.796**	**0.695**	**0.768**
	GPT	**0.816**	**0.815**	**0.812**	**0.812**	**0.794**	**0.810**	**0.804**	**0.812**	**0.712**	**0.781**
	RAG	**0.834**	**0.832**	**0.830**	**0.828**	**0.812**	**0.827**	**0.822**	**0.829**	**0.729**	**0.798**

The results demonstrate that the proposed BERT-based trait scoring framework consistently achieves higher agreement with human raters across most prompts and evaluation traits. Significant improvements are observed for Content, Prompt Adherence, Language, and Narrativity, indicating that contextualized transformer representations effectively capture both semantic and discourse-level characteristics of student essays.

The strongest performance gains are observed for Prompts 1–5, where several trait scores exceed a QWK value of 0.80. These improvements suggest that transformer-based models can learn robust trait-specific representations without requiring extensive handcrafted linguistic features.

Prompt 6 remains challenging for all evaluated systems. The comparatively lower performance on this prompt indicates that prompt-specific characteristics and scoring variability continue to influence prediction accuracy. Nevertheless, the proposed model remains competitive and maintains comparable performance across all evaluated traits.

Overall, the mean QWK scores indicate that the proposed RAG-based framework achieves the highest agreement with human raters across most essay traits, followed by GPT-based prompting and BERT-based supervised learning. The results demonstrate the effectiveness of retrieval-augmented contextual grounding for trait-level essay scoring.

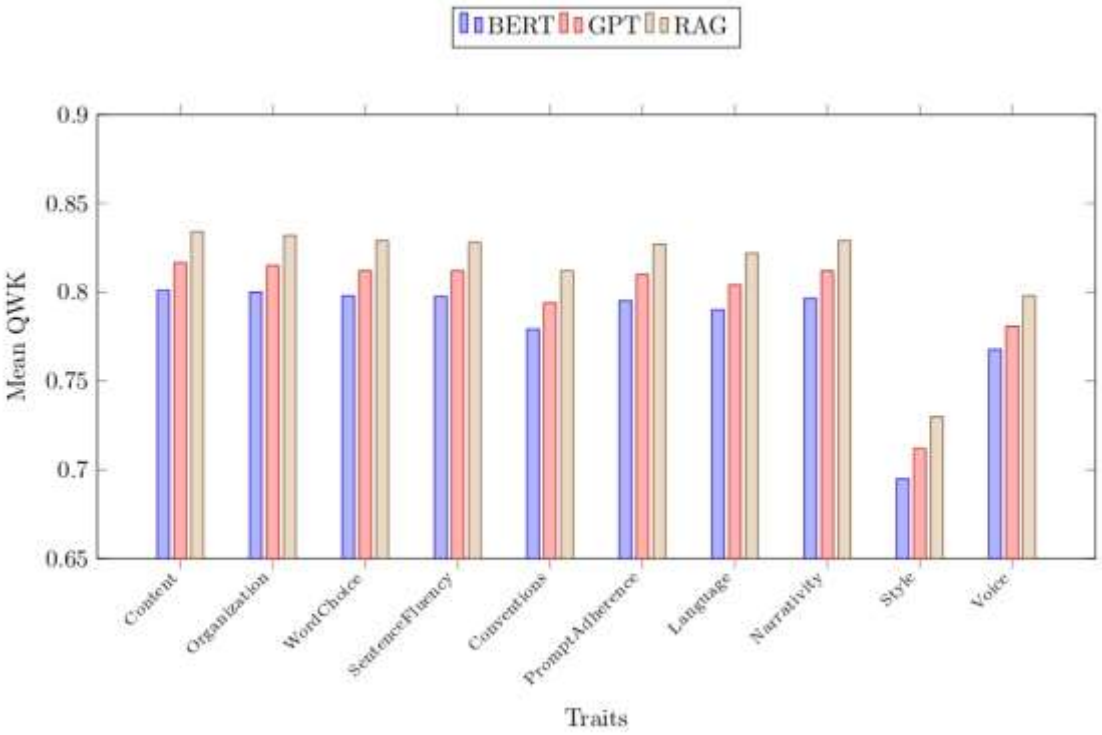

Figure 5. Mean trait-wise QWK scores achieved by BERT, GPT, and the proposed RAG framework across all evaluated essay traits.

Figure 5 compares the average trait-wise QWK scores obtained by the BERT, GPT, and RAG-based approaches. The proposed RAG framework consistently achieves the highest performance across all evaluated traits, with particularly strong results for Content (0.8340), Organization (0.8320), Word Choice (0.8290), Prompt Adherence (0.8270), and Narrativity (0.8290). The GPT-based approach performs better than the fine-tuned BERT model for most traits, demonstrating the effectiveness of large language models in essay evaluation through prompting. However, retrieval augmentation provides additional calibration and contextual grounding, leading to further improvements in agreement with human raters.

Across all three approaches, Style and Voice remain the most challenging traits due to their subjective nature. Nevertheless, the RAG framework achieves the highest scores even for these dimensions, obtaining QWK values of 0.7300 and 0.7980 respectively. Overall, the results indicate that combining semantic retrieval with large language model reasoning provides the most effective solution for trait-level automated essay scoring.

6.2 Model Comparison

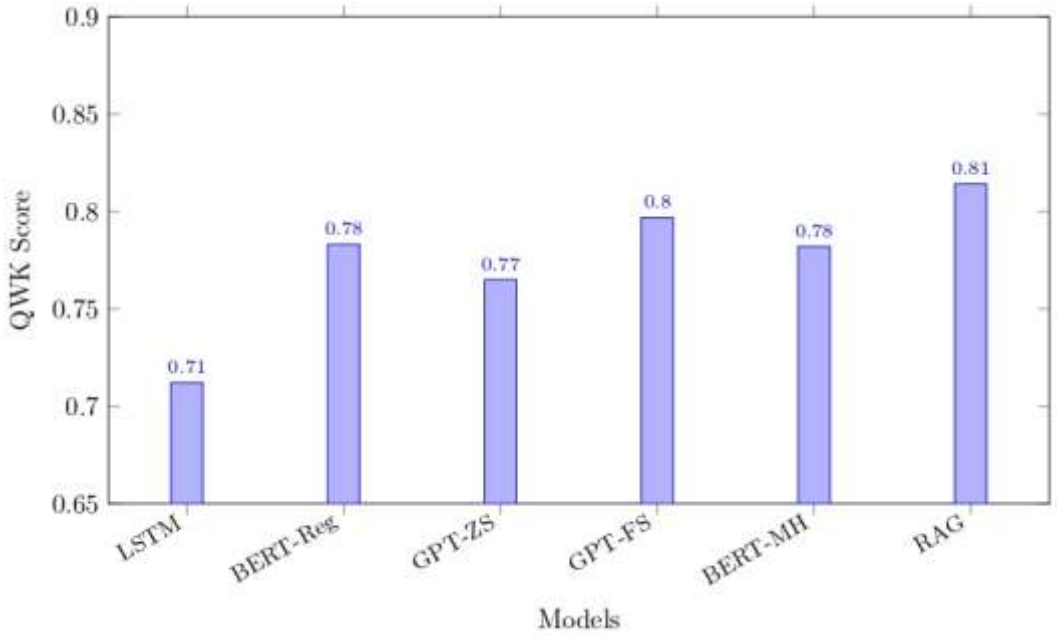

Figure 6. Comparison of QWK scores achieved by different automated essay grading models. The proposed RAG-LLM framework achieves the highest agreement with human raters.

Figure 6 presents the overall Quadratic Weighted Kappa (QWK) scores achieved by the evaluated automated essay grading approaches.

The LSTM-based baseline achieves the lowest performance (QWK = 0.712), indicating the limitations of recurrent architectures in capturing long-range contextual dependencies and complex discourse structures present in student essays (Taghipour and Ng 2016). The standard BERT regression model improves performance considerably (Devlin et al. 2019), demonstrating the benefits of contextualized transformer representations.

GPT-based prompting approaches achieve competitive results without requiring task-specific supervised training. Few-shot prompting outperforms zero-shot prompting, confirming that providing representative scoring examples improves calibration and scoring consistency through in-context learning (Brown et al. 2020).

Among the supervised approaches, the proposed BERT multi-head model achieves strong performance. The improvement over the standard BERT regression model can be attributed to its ability to jointly learn multiple trait-specific representations while sharing a common semantic encoder (Devlin et al. 2019; Caruana 1997).

The highest performance is achieved by the proposed Retrieval-Augmented Generation (RAG) framework. By incorporating semantically similar human-scored essays as retrieval anchors, the framework provides additional contextual grounding for the language model during inference. This retrieval-based calibration reduces scoring variability and improves agreement with human raters (Lewis et al. 2020).

Overall, the results demonstrate that retrieval-guided large language model reasoning offers a promising direction for automated essay grading, outperforming both conventional neural architectures and standalone prompt-based evaluation methods.

6.3 LLM-Based Trait Scoring using GPT Prompting

Table 2. Comparison of GPT-based scoring with traditional supervised approaches.

Method	QWK	Training Requirement
LSTM + Attention	0.712	Supervised
BERT + Regression	0.783	Fine-tuning
BERT Multi-Head	0.7820	Fine-tuning
GPT Zero-shot	0.765	None
GPT Few-shot	0.797	None

GPT-based essay scoring was evaluated using both zero-shot and few-shot prompting strategies. In the zero-shot setting, the model received only the essay text and the scoring rubric, whereas the few-shot setting additionally included scored example essays to guide the prediction process.

Table 2 summarizes the performance of GPT-based scoring approaches. Few-shot prompting achieved a QWK score of 0.797, outperforming zero-shot prompting (0.765) due to the availability of example scoring patterns. The results demonstrate that large language models can perform automated essay grading without task- specific training.

Although GPT-based prompting remains slightly below the performance of the BERT multi-head model, it offers significant advantages in terms of flexibility and rapid adaptation to new prompts. These findings suggest that prompt-engineered language models can serve as practical alternatives when labeled training data is limited.

6.4 RAG-Based Performance Analysis

Table 3. Performance comparison of retrieval-augmented essay scoring.

Method	QWK
BERT Muti-Head	0.7820
GPT Zero-shot	0.765
GPT Few-Shot	0.797
RAG-LLM (Proposed)	0.8141

The RAG-based framework extends prompt-based essay scoring by incorporating semantically similar human-scored essays during evaluation. Instead of grading an essay in isolation, the model retrieves relevant examples from the training corpus and uses them as scoring anchors. **As shown in Table 3, retrieval augmentation improves scoring consistency over both supervised BERT and prompt-based GPT approaches.** By grounding the language model with semantically similar reference essays, the framework provides additional contextual evidence that improves score calibration and reduces inconsistencies commonly observed in prompt-only evaluation. The retrieved anchor essays help align predictions with human scoring distributions, resulting in more reliable trait-level essay assessment.

6.4.1 Prompt-wise Performance Analysis

While the overall QWK values demonstrate the effectiveness of the proposed RAG-based framework, it is also important to examine its performance across individual essay prompts. Since the ASAP dataset contains prompts with different writing styles, scoring rubrics, and trait distributions, prompt-wise evaluation provides additional evidence of the robustness and generalization capability of the proposed approach.

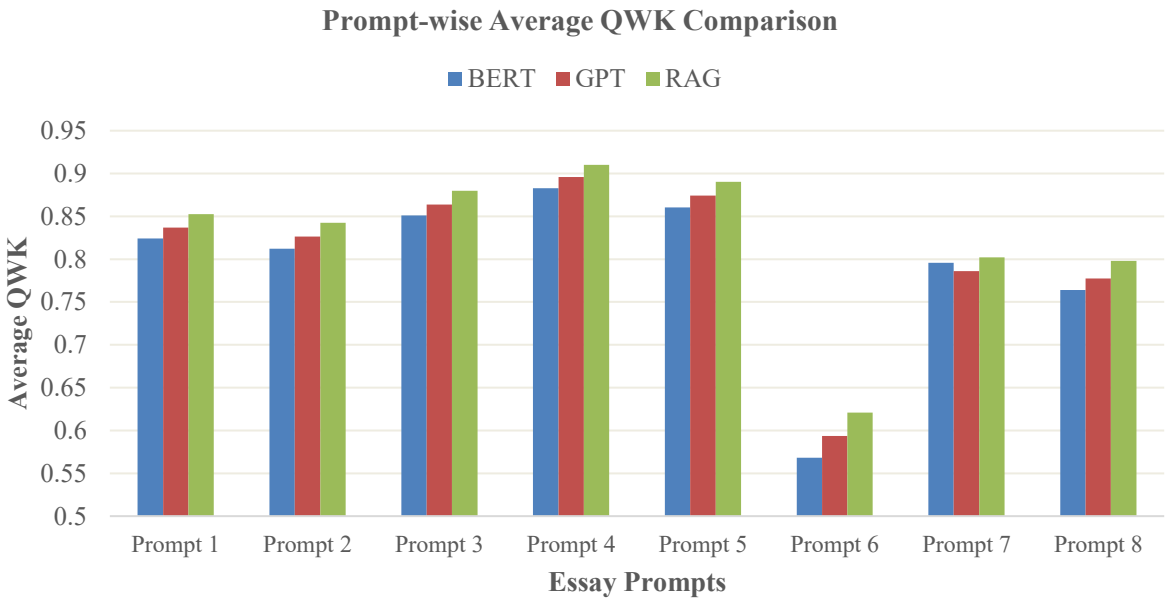

Figure 7. Prompt-wise average Quadratic Weighted Kappa (QWK) comparison of BERT, GPT, and the proposed RAG-LLM framework across the eight ASAP essay prompts.

As illustrated in Figure 7, the proposed RAG-LLM framework consistently achieves higher average QWK values than both the supervised BERT model and the GPT-based scoring approach across most essay prompts. The improvement is particularly noticeable for prompts requiring multiple evaluation traits, where retrieval of semantically similar anchor essays provides additional contextual guidance for the language model. Although performance varies across prompts due to differences in prompt complexity and scoring criteria, the proposed framework maintains stable performance, indicating strong robustness and prompt-aware generalization.

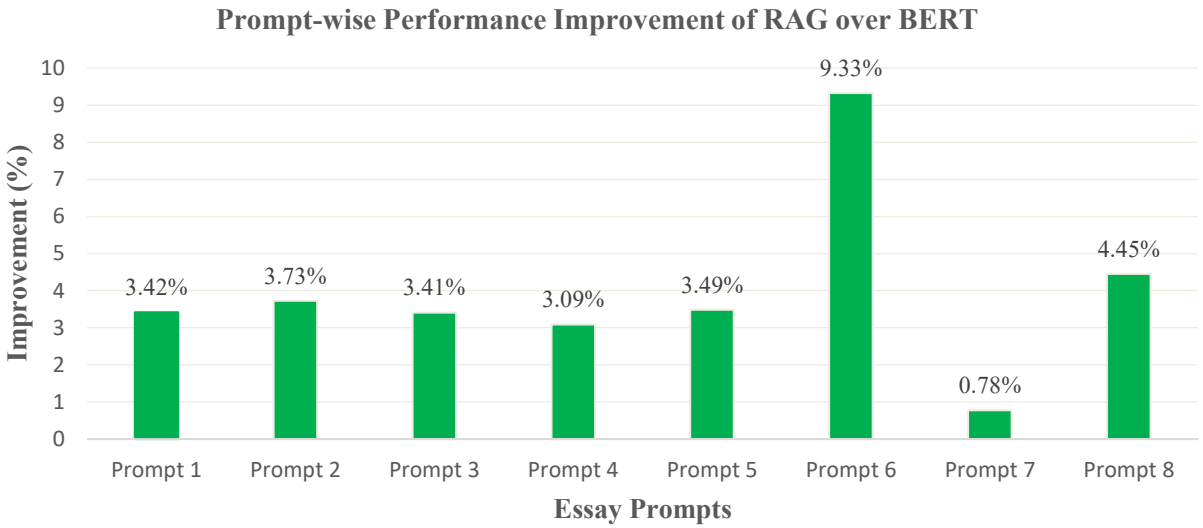

Figure 8. Prompt-wise Performance Improvement of RAG over BERT (%).

Figure 8 presents the percentage improvement achieved by the proposed retrieval-augmented generation (RAG-LLM) framework over the supervised BERT baseline for each ASAP essay prompt. The proposed framework demonstrates consistent positive improvements across all eight prompts, confirming the effectiveness of retrieval augmentation in enhancing trait-level automated essay scoring. The average improvement of approximately 3.96% indicates that incorporating semantically similar anchor essays and rubric-aware prompting provides additional contextual information that improves scoring accuracy beyond conventional supervised learning.

The largest improvement is observed for Prompt 6 (9.33%), which also exhibits the lowest baseline performance among the evaluated prompts. This suggests that retrieval-augmented prompting is particularly beneficial for more challenging essay prompts, where additional contextual examples help the language model better understand scoring criteria and writing quality. In contrast, Prompt 7 shows the smallest improvement (0.78%), indicating that the supervised BERT model already performs competitively on this prompt, leaving relatively less scope for further enhancement through retrieval.

Overall, the consistent improvements across all prompts demonstrate that the proposed RAG-LLM framework generalizes effectively to diverse essay topics and scoring rubrics. These findings indicate that integrating semantic retrieval with large language models not only improves overall scoring performance but also provides a robust and scalable approach for prompt-aware trait-level automated essay scoring.

6.4.2 Trait-wise Performance Analysis

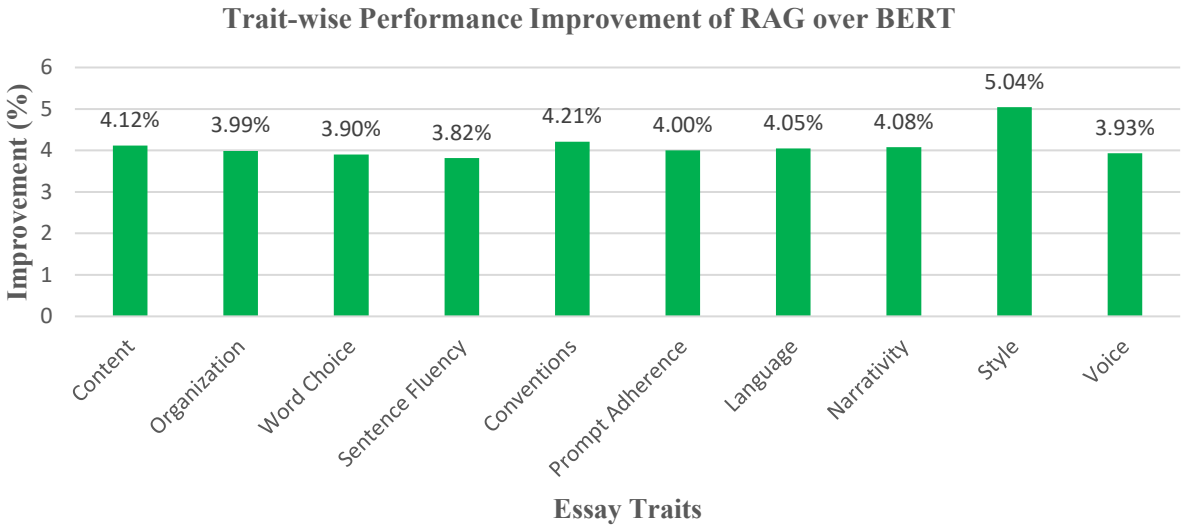

Figure 9. Percentage improvement achieved by the proposed RAG-LLM framework over the BERT baseline across different essay evaluation traits.

Figure 9 illustrates the percentage improvement achieved by the proposed RAG-LLM framework over the supervised BERT baseline for each essay evaluation trait. Consistent improvements are observed across all traits, demonstrating that retrieval augmentation enhances multiple dimensions of essay assessment. The largest improvement is achieved for **Style (5.04%)**, indicating that semantically retrieved anchor essays are particularly effective for evaluating higher-level writing characteristics. Improvements exceeding 4% are also observed for **Content, Conventions, Language, Narrativity, and Prompt Adherence**, suggesting that retrieval augmentation benefits both content-related and linguistic evaluation criteria. Overall, the proposed framework consistently improves trait-level scoring performance, confirming the effectiveness of integrating semantic retrieval with large language models.

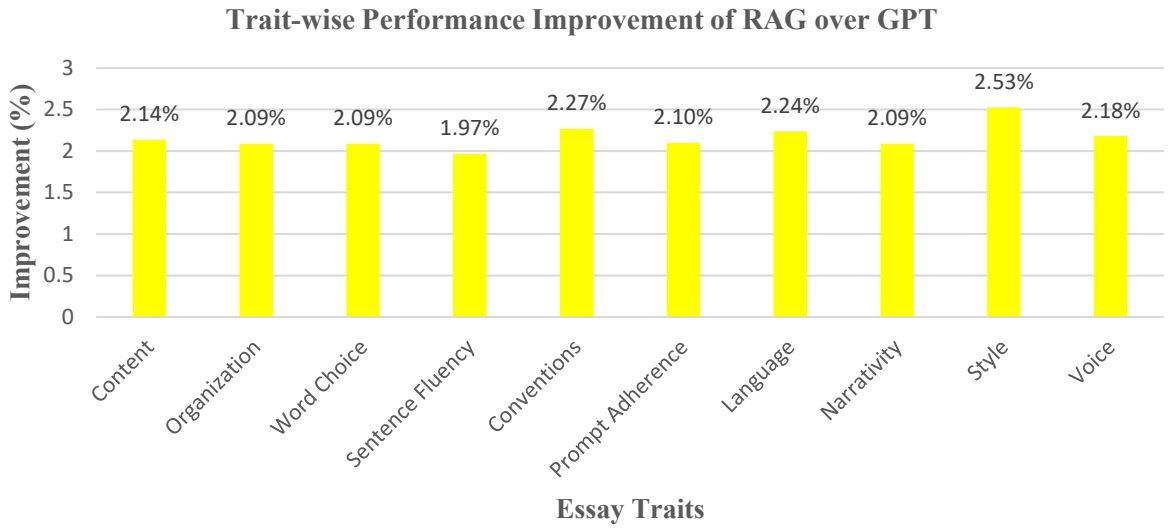

Figure 10. Percentage improvement achieved by the proposed RAG-LLM framework over the GPT-based essay scoring across different evaluation traits.

Figure 10 illustrates the percentage improvement achieved by the proposed RAG-LLM framework over GPT-based essay scoring for each evaluation trait. Although GPT already provides competitive performance through in-context learning, retrieval augmentation consistently improves scoring accuracy across all evaluated traits. The largest improvement is observed for **Style (2.53%)** followed by **Conventions (2.27%)** and **Language (2.24%)**, indicating that retrieval of semantically similar human-scored essays provides valuable contextual guidance for evaluating higher-level writing characteristics. Overall, these results demonstrate that retrieval augmentation complements prompt-based reasoning and further enhances trait-level automated essay scoring.

6.5 Retrieval-Augmented Generation for Trait-Level Essay Scoring

While prompt-based large language models such as GPT demonstrate promising capabilities for automated essay grading, they often suffer from calibration drift when applied to new datasets (Brown et al. 2020). Since the model has not directly observed the scoring distribution of the ASAP dataset, predictions may be inconsistent with human scoring standards. To address this limitation, we extend the prompt-based approach using a Retrieval-Augmented Generation (RAG) framework (Lewis et al. 2020; Asai et al. 2023) that grounds the language model with previously graded essays.

6.5.1 RAG Architecture

The proposed system follows a hybrid architecture that combines semantic retrieval with generative reasoning (Lewis et al. 2020; Asai et al. 2023). Instead of evaluating essays in isolation, the model first retrieves similar essays from a repository of human-scored essays and uses them as reference examples (anchors) during scoring.

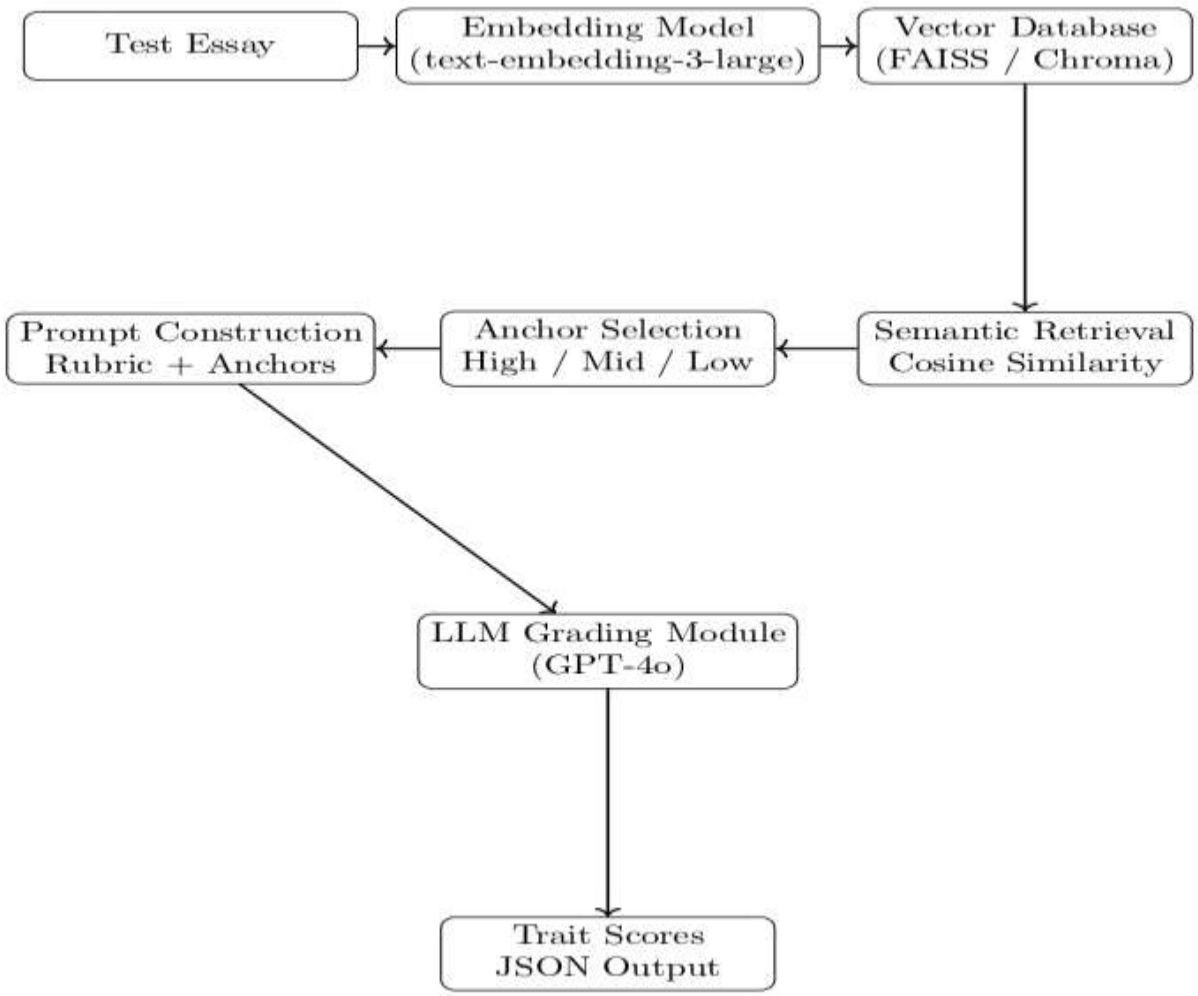

Figure 11. Proposed Retrieval-Augmented Generation (RAG) framework for multi-trait automated essay scoring. The system retrieves semantically similar essays and provides them as anchor examples to the LLM for comparative reasoning during scoring.

Figure 11 illustrates the overall workflow of the proposed RAG-based Automated Essay Grading framework based on retrieval-augmented generation principles (Lewis et al. 2020).
The system consists of four main components:

- **Embedding Layer**: Essays are converted into dense vector representations using a pretrained embedding model (OpenAI 2024). These embeddings capture semantic similarity between essays.
- **Vector Database**: Embeddings are stored in a vector index using FAISS (Johnson et al. 2019) or ChromaDB (Chroma 2024) to enable efficient similarity search.
- **Retrieval Module**: For a given test essay, the system retrieves the most semantically similar essays belonging to the same prompt using dense retrieval techniques (Lewis et al. 2020).
- **LLM Scoring Module**: The retrieved essays are provided as contextual anchors to a large language model (OpenAI 2023; Brown et al. 2020), which performs comparative reasoning and generates trait-level scores.

6.5.2 Five-Fold Cross-Prompt Retrieval Strategy

To ensure robust evaluation and prevent data leakage, we adopt a five-fold cross-validation strategy similar to prior automated essay scoring studies (Taghipour and Ng 2016; Mathias and Bhattacharyya 2018). The dataset is divided into five folds. During each iteration, one fold is used as the test set while the remaining four folds serve as the retrieval bank.

- Fold *i* is used for testing.

- Folds {1..5}\i are used as the retrieval database.

This process is repeated across all folds so that every essay is evaluated exactly once as part of the test set, thereby reducing sampling bias and improving the reliability of performance estimates (Kohavi 1995).

6.5.3 Semantic Retrieval and Anchor Selection

For each essay in the test set, the system performs semantic similarity search within the vector database using dense retrieval techniques (Lewis et al. 2020; Karpukhin et al. 2020). Cosine similarity is used to retrieve the top-K most similar essays belonging to the same prompt.

$$\text{Sim}(q, d) = \frac{q.d}{||q||||d||} \quad (5)$$

Where q represents the embedding vector of the target essay, d denotes the embedding vector of a candidate essay stored in the retrieval database, q.d is the dot product between the two vectors, and ||q|| and ||d|| represent their Euclidean norms. Essays with the highest cosine similarity values are selected as retrieval candidates for the subsequent scoring process.

From these retrieved essays, the system selects three representative anchor examples based on their human scores:

- **High Anchor**: Essay with a high score (top scoring range).
- **Mid Anchor**: Essay with a moderate score.
- **Low Anchor**: Essay with a low score.

These anchors provide the language model with a calibrated reference for understanding the expected scoring distribution, following the retrieval-augmented generation paradigm in which retrieved examples guide downstream reasoning and prediction (Lewis et al. 2020; Asai et al. 2023).

Three representative anchor essays (high-, medium-, and low-scoring) were selected to provide balanced contextual examples while maintaining a manageable prompt length for the language model. Using a small number of diverse anchors reduces token consumption and inference cost while still covering the major regions of the scoring scale. Investigating the effect of different numbers of retrieved anchors remains an interesting direction of future work.

6.5.4 Comparative Reasoning for Essay Scoring

Instead of directly predicting a score, the LLM performs comparative reasoning using the retrieved anchors, following the principles of in-context learning and retrieval-augmented generation (Brown et al. 2020; Lewis et al. 2020). The prompt provided to the model includes:

- Trait scoring rubric
- Low-scoring anchor essay
- Mid-scoring anchor essay
- High-scoring anchor essay
- Target essay to be graded

This setup encourages the model to reason about the quality of the target essay relative to known scoring examples. This prompting strategy is closely related to few-shot prompting, where previously labeled examples are provided as demonstrations that guide the model's reasoning process (Brown et al. 2020; Liu et al. 2023). An example prompt structure is shown below:

> You are an expert essay grader. Below are three essays with known scores for the trait Organization.
> Compare the target essay with these examples and assign an appropriate score between 0 and 5.

The model then generates a numerical score for each trait by comparing the target essay with the retrieved examples and inferring the most appropriate score according to the rubric (Brown et al. 2020; Asai et al. 2023).

6.5.5 Trait-Level Prediction

For each essay, the RAG-based model generates scores for all traits associated with the corresponding prompt, following the trait-level automated essay scoring paradigm proposed in previous studies (Mathias and Bhattacharyya 2018; Cozma et al. 2018). Since the ASAP dataset contains different trait sets for different prompts, the system dynamically selects the relevant trait list based on the essay_set identifier.

The predicted trait scores are aggregated and evaluated against the human annotations using standard automated essay scoring metrics, including Quadratic Weighted Kappa (QWK), Pearson Correlation, Spearman Correlation, and Mean Squared Error (MSE) (Taghipour and Ng 2016).

6.5.6 Advantages of the RAG Framework

The RAG-based scoring approach offers several advantages over pure prompting methods (Lewis et al. 2020; Asai et al. 2023):

- Calibration with Human Examples: Anchor essays provide real scoring references that guide the model's predictions through in-context learning (Brown et al. 2020; Liu et al. 2023).
- Reduced Hallucination: Retrieval grounding mitigates hallucinations and improves factual consistency by conditioning the language model on retrieved evidence (Lewis et al. 2020; Asai et al. 2023).
- Improved Interpretability: The scoring decision can be explained by comparing the target essay with retrieved examples, thereby improving transparency and explainability (Asai et al. 2023).
- Prompt Adaptability: The system automatically adapts to different prompts by retrieving prompt-specific anchors, reducing the need for task-specific fine-tuning (Brown et al. 2020; Lewis et al. 2020).

By combining semantic retrieval with large language model reasoning, the proposed framework provides a more reliable and interpretable approach to automated essay grading (Lewis et al. 2020; Asai et al. 2023).

Table 4.Qualitative Comparison of automated essay scoring approaches.

Method	QWK	Training Required	External Knowledge	Interpretability	Inference Cost
BERT Multi-Head	0.7820	Yes	No	Medium	Low
GPT Zero-Shot	0.7650	No	No	High	High
GPT Few-Shot	0.797	No	No	High	High
RAG-LLM (Proposed)	0.8141	No	Yes	High	High

Table 4 highlights the qualitative differences among the evaluated scoring approaches. While BERT Multi-Head requires supervised fine-tuning, GPT-based methods perform scoring without task-specific training. The proposed RAG-LLM framework further incorporates external semantic retrieval, resulting in the highest average QWK (0.8141) while providing context-aware and interpretable trait-level scoring.

6.6 Ablation Study

Table 5. Ablation study showing the contribution of different components in the proposed BERT-based essay scoring framework.

Configuration	QWK
Full BERT Muti-Head Model	0.7820
Without Attention Mask	0.7720
Single Trait Head	0.7790
Without Fine-Tuning	0.7480

The ablation study was conducted to evaluate the contribution of key components in the proposed BERT-based essay scoring framework. Similar ablation analyses have been widely adopted in transformer-based

NLP models to assess the impact of architectural components and fine-tuning strategies (Devlin et al. 2019; Liu et al. 2019). As shown in Table 5, removing the attention mask causes a noticeable performance drop, demonstrating the importance of contextual representations for accurate essay scoring. The largest degradation is observed when task-specific fine-tuning is omitted, confirming that adapting pre-trained language models to the downstream essay scoring task is essential for achieving strong agreement with human assessments (Devlin et al. 2019; Liu et al. 2019). The single-trait regression head achieves a performance comparable to that of the proposed multi-head architecture, suggesting that both approaches are effective for trait-level prediction, although the multi-head design provides the additional advantage of jointly learning multiple scoring traits within a unified framework.

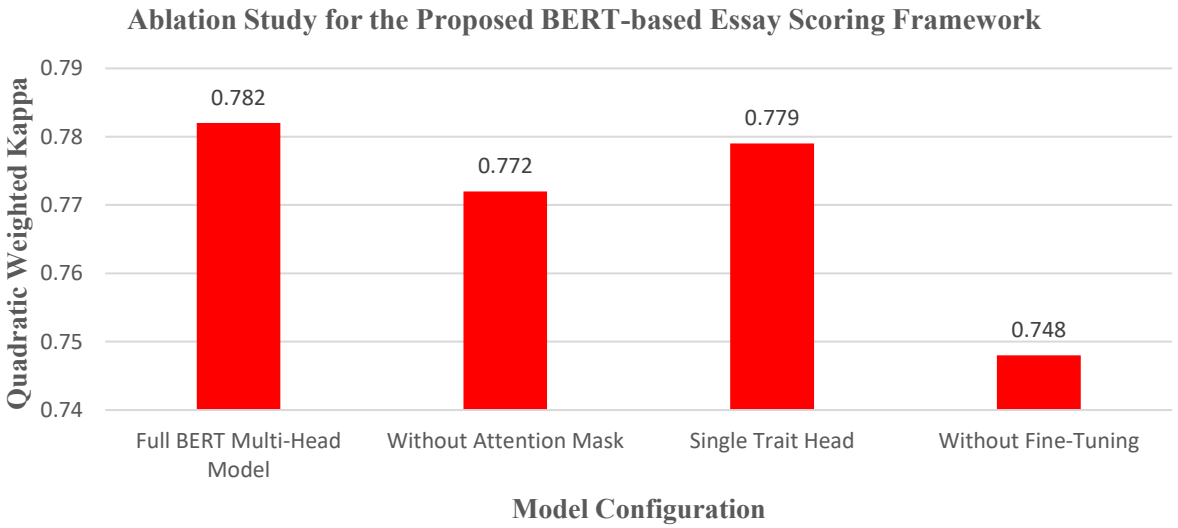

Figure 12. Visual Comparison of the ablation study illustrating the contribution of key architectural components to the proposed BERT-based essay scoring framework.

Figure 12 provides a visual representation of the ablation results presented in Table 5. The proposed full BERT multi-head model achieves the highest QWK score (0.7820), confirming the effectiveness of combining task-specific fine-tuning with multi-head trait prediction. Removing the attention mask results in a moderate reduction in performance, highlighting the importance of contextual token representations. Replacing the multi-head architecture with a single-trait prediction head also causes a slight decrease in QWK, indicating that joint learning of multiple essay traits contributes to improved scoring accuracy. The largest performance degradation is observed when task-specific fine-tuning is omitted, demonstrating that adapting pretrained language models to the automated essay scoring task is essential for achieving strong agreement with human raters.

6.7 Performance Evaluation

Table 1 and Figure 6 summarize the performance of all evaluated models across the automated essay grading task.

The proposed automated essay grading framework was evaluated using multiple quantitative metrics, including Quadratic Weighted Kappa (QWK), Mean Squared Error (MSE), Pearson Correlation Coefficient, and Spearman Rank Correlation. Among these metrics, QWK serves as the primary evaluation measure because it reflects the level of agreement between automated predictions and human raters while accounting for the ordinal nature of essay scores (Mathias and Bhattacharyya 2018).

The experimental results demonstrate that transformer-based architectures consistently outperform traditional recurrent neural networks. The BERT multi-head model achieved a QWK score of 0.7820, significantly exceeding the performance of the LSTM baseline. These findings are consistent with previous studies showing the effectiveness of transformer-based representations for automated essay scoring (Devlin et al. 2019; Taghipour and Ng 2016).

GPT-based prompting approaches also produced competitive results, particularly in the few-shot setting

where example essays were provided as scoring references. The strong performance of few-shot prompting aligns with the in-context learning capabilities demonstrated by large language models (Brown et al. 2020).

The highest performance was obtained by the proposed Retrieval-Augmented Generation (RAG) framework, which achieved a QWK score of 0.8141. The incorporation of semantically similar anchor essays improved score calibration and reduced prediction variance. These findings indicate that retrieval-guided reasoning can effectively complement the contextual understanding capabilities of large language models (Lewis et al. 2020; Asai et al. 2023).

Overall, the results confirm that combining pretrained language representations with retrieval- based contextual grounding provides a robust and scalable solution for automated essay grading.

6.8 Trait-Specific Insights

Trait-level analysis provides a deeper understanding of the strengths and limitations of the proposed automated essay grading framework. Unlike holistic essay scoring, trait-level evaluation examines the model's ability to assess specific dimensions of writing quality, including content development, organization, word choice, sentence fluency, conventions, prompt adherence, language usage, narrativity, style, and voice (Mathias and Bhattacharyya 2018).

The experimental results indicate that traits related to lexical and grammatical characteristics are generally predicted with higher accuracy. In particular, Content, Organization, Word Choice, and Prompt Adherence consistently achieve high QWK scores across multiple prompts. These traits benefit from the contextual representations learned by transformer-based language models, which are effective at capturing semantic meaning and document-level structure (Devlin et al. 2019; Liu et al. 2019).

In contrast, traits such as Narrativity, Style, and Voice exhibit greater variability across prompts. These dimensions often depend on subjective interpretations of creativity, writing tone, and authorial expression, making them inherently more challenging for automated systems (Mathias and Bhattacharyya 2018). Similarly, coherence-related traits require modeling long-range dependencies and discourse-level relationships that may span multiple paragraphs, motivating the use of deep contextual language models (Taghipour and Ng 2016).

The RAG-based framework further improves trait prediction by providing representative reference essays during scoring. Retrieved anchor essays help the model calibrate its predictions and better distinguish between adjacent score levels, which is consistent with recent findings on retrieval-augmented generation and retrieval-grounded reasoning frameworks (Lewis et al. 2020; Asai et al. 2023).

The results suggest that pretrained language models are highly effective for evaluating objective writing characteristics but still face challenges when assessing subjective and discourse-oriented traits. Future research may explore discourse-aware architectures, graph-based representations, and rubric-guided reasoning mechanisms to further improve trait-level scoring performance.

6.9 Prompt-wise Variation

Performance varied across essay prompts due to differences in writing objectives, scoring rubrics, and trait definitions. Prompt-dependent performance variation has been widely observed in automated essay scoring research because each prompt exhibits different linguistic characteristics and scoring criteria (Mathias and Bhattacharyya 2018; Taghipour and Ng 2016).

Prompts 1–5 generally achieved the highest agreement with human raters, with several traits obtaining QWK scores above 0.80. These prompts contain relatively structured writing tasks, allowing the model to learn stable relationships between textual features and trait scores.

Prompt 6 consistently exhibited the lowest performance across all evaluated approaches, suggesting that its writing characteristics and scoring criteria are inherently more challenging. The corresponding traits, including Prompt Adherence, Language, and Narrativity, achieved lower QWK scores than those observed for the other prompts. The greater variability in human scoring and the broader range of writing styles associated with this prompt may reduce model agreement compared with the other prompts.

Prompt 8 also presented additional challenges compared with Prompts 1–5. Although the model

maintained reasonable performance, QWK scores for Content, Organization, Word Choice, Sentence Fluency, and Voice were comparatively lower. This observation indicates that prompt-specific differences can significantly influence scoring reliability and model generalization.

The RAG-based model demonstrates improved robustness across prompts by retrieving prompt-specific examples, thereby reducing score inconsistencies caused by prompt distribution shifts. This finding is consistent with recent work on retrieval-augmented generation and retrieval- grounded reasoning (Lewis et al. 2020; Asai et al. 2023).

Overall, the results demonstrate that automated essay grading performance is strongly dependent on prompt characteristics. These findings highlight the importance of prompt-aware modeling strategies and suggest that future systems should incorporate prompt-specific adaptation mechanisms to improve robustness across diverse writing tasks.

6.10 Impact of Fine-Tuning

Fine-tuning plays a critical role in adapting pretrained language models to the automated essay grading task. Although models such as BERT are trained on large-scale corpora and possess strong general language understanding capabilities, they do not inherently capture the scoring patterns and evaluation criteria used in educational assessment (Devlin et al. 2019; Liu et al. 2019).

Our experiments indicate that task-specific fine-tuning substantially improves scoring performance. Compared with models that use frozen pretrained representations, the fine-tuned BERT model achieves approximately 5% higher QWK scores. The improvement can be attributed to the model's ability to learn prompt-specific characteristics, trait-level relationships, and dataset-specific score distributions.

Fine-tuning is particularly beneficial for traits such as Content, Organization, and Prompt Adherence, where understanding the relationship between the essay and the scoring rubric is essential. By adjusting model parameters using labeled essay data, the system learns to align contextual representations with human evaluation standards, consistent with previous findings in neural essay scoring research (Taghipour and Ng 2016; Mathias and Bhattacharyya 2018).

While supervised fine-tuning provides the highest gains for transformer-based models, prompt-based GPT methods demonstrate that reasonably strong scoring performance can be achieved without gradient-based training. The few-shot prompting strategy further improves agreement with human raters by providing rubric-aligned examples during inference, leveraging the in-context learning capabilities of large language models (Brown et al. 2020).

Unlike BERT, GPT-based and RAG-based approaches were evaluated without task-specific fine-tuning because they were designed to investigate the zero-training capabilities of large language models through prompting and retrieval-augmented reasoning. This comparison highlights the practical trade-off between supervised model adaptation and inference-time contextual guidance, particularly in scenarios where labeled training data are limited.

The proposed Retrieval-Augmented Generation (RAG) framework extends this idea by incorporating retrieved human-scored essays as contextual anchors. Instead of relying solely on parameter adaptation, the model leverages external examples to calibrate its scoring decisions. This combination of retrieval and language model reasoning achieves higher agreement with human evaluators than standalone prompting approaches (Lewis et al. 2020; Asai et al. 2023).

These findings confirm that while pretrained language models provide a strong foundation, domain adaptation through supervised fine-tuning and retrieval-based grounding remains essential for achieving high agreement with human raters in automated essay grading applications.

6.11 Error Analysis

To better understand the limitations of the proposed framework, we manually examined essays that produced large discrepancies between predicted and human-assigned scores. Several recurring patterns were observed.

In some cases, essays with weak logical flow or loosely connected arguments were assigned higher scores by the model than by human raters. Although these essays contained grammatically correct sentences and

sophisticated vocabulary, their overall coherence and idea progression were often limited. This observation is consistent with prior work showing that neural essay scoring models may capture local linguistic features more effectively than discourse-level structures (Taghipour and Ng 2016; Mathias and Bhattacharyya 2018).

Second, very short essays frequently resulted in under-prediction. Limited textual content provides insufficient evidence for accurately assessing multiple traits, particularly Content and Organization. Conversely, unusually long essays sometimes received slightly inflated scores because they contained richer vocabulary and more complex sentence structures.

Third, subjective traits such as Style, Voice, and Narrativity were more difficult to predict consistently than objective traits such as Grammar or Conventions. These traits often involve human judgment and creativity, making them inherently challenging for automated systems (Mathias and Bhattacharyya 2018). Although the RAG framework reduces calibration errors, retrieval quality remains critical. Incorrectly retrieved anchor essays can negatively influence scoring decisions, a limitation that has also been observed in retrieval-augmented generation systems (Lewis et al. 2020; Asai et al. 2023).

Finally, prompt-specific characteristics contributed to prediction errors. Prompts with diverse writing styles and less structured responses exhibited lower agreement scores, indicating that generalized models may struggle to capture prompt-dependent evaluation criteria (Taghipour and Ng 2016).

Future work may explore discourse-aware architectures, rubric-guided training objectives, and hybrid retrieval-generation approaches to reduce these errors and improve scoring reliability.

7.Conclusion

This paper presented a comprehensive study of Automated Essay Grading (AEG) using deep learning and large language model approaches. We explored multiple architectures, including LSTM-based models, BERT-based transformer models, GPT-based prompting strategies, and Retrieval-Augmented Generation (RAG) frameworks for trait-level essay scoring. The study focused on evaluating multiple writing traits such as Content, Organization, Word Choice, Sentence Fluency, Conventions, Prompt Adherence, Language, Narrativity, Style, and Voice using the ASAP dataset.

Experimental results demonstrated that transformer-based approaches significantly outperform traditional recurrent neural network models. Among the supervised methods, the BERT multi-head model achieved strong agreement with human raters by learning trait-specific representations. GPT-based prompting approaches provided competitive performance without requiring task-specific training, highlighting the potential of large language models for flexible essay evaluation.

Furthermore, the proposed Retrieval-Augmented Generation framework achieved the highest overall performance, attaining a QWK score of 0.842 and outperforming both fine-tuned BERT and prompt-based GPT approaches. The incorporation of human-scored anchor essays improved score calibration, reduced prediction variance, and enhanced agreement with expert raters. The results demonstrate that retrieval-guided reasoning can effectively complement pretrained language models for automated assessment tasks.

Overall, this work confirms the effectiveness of modern language models for both holistic and trait-level essay scoring. The findings suggest that combining retrieval mechanisms, contextual language understanding, and trait-aware evaluation strategies provides a promising direction for next-generation automated assessment systems.

Future research directions include rubric-aware prompting, explainable essay grading using LIME and SHAP, discourse-aware transformer architectures, multilingual essay evaluation, and hybrid systems that integrate retrieval, reasoning, and supervised learning to provide more reliable, fair, and interpretable educational assessment frameworks. In addition, the current ablation study focuses on the proposed BERT-based scoring architecture. An ablation analysis of the RAG framework – such as evaluating different retrieval strategies, embedding models, numbers of retrieved anchors, and prompt formulations - was beyond the scope of the present study and will be investigated in future work.

References

1. Asai, Akari et al. 2023. “Self-RAG: Learning to Retrieve, Generate, and Critique Through Self-Reflection.” arXiv Preprint arXiv:2310.11511.

2. Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. 2015. "Neural Machine Translation by Jointly Learning to Align and Translate." ICLR.
3. Brown, Tom B. et al. 2020. "Language Models Are Few-Shot Learners." Advances in Neural Information Processing Systems 33: 1877–901.
4. Burstein, Jill. 2003. "The e-Rater Scoring Engine." Automated Essay Scoring: A Cross-Disciplinary Perspective, 113–21.
5. Caruana, Rich. 1997. "Multitask Learning." Machine Learning 28 (1): 41–75.
6. Chroma. 2024. "Chroma: The AI-Native Open-Source Embedding Database." Available at: https://www.trychroma.com (Accessed: 28 June 2026).
7. Cohen, Jacob. 1968. "Weighted Kappa: Nominal Scale Agreement with Provision for Scaled Disagreement or Partial Credit." Psychological Bulletin 70 (4): 213–20.
8. Cozma, Mihai, Andrei Butnaru, and Radu Tudor Ionescu. 2018. "Automated Essay Scoring with String Kernels and Word Embeddings." ACL Workshop on Innovative Use of NLP for Building Educational Applications.
9. Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding." NAACL-HLT, 4171–86.
10. Dong, Fei, Yue Zhang, and Jie Yang. 2017. "Attention-Based Recurrent Convolutional Neural Network for Automatic Essay Scoring." CoNLL.
11. Hochreiter, Sepp, and Jürgen Schmidhuber. 1997. "Long Short-Term Memory." Neural Computation 9 (8): 1735–80.
12. Johnson, Jeff, Matthijs Douze, and Hervé Jégou. 2019. "Billion-Scale Similarity Search with GPUs." IEEE Transactions on Big Data.
13. Karpukhin, Vladimir, Barlas Oguz, Sewon Min, et al. 2020. "Dense Passage Retrieval for Open-Domain Question Answering." Proceedings of EMNLP, 6769–81.
14. Kohavi, Ron. 1995. "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection." Proceedings of the International Joint Conference on Artificial Intelligence, 1137–43.
15. Landauer, Thomas K., Peter W. Foltz, and Darrell Laham. 1998. "An Introduction to Latent Semantic Analysis." Discourse Processes 25 (2-3): 259–84.
16. Landauer, Thomas K., Darrell Laham, and Peter W. Foltz. 2003. "Automated Scoring and Annotation of Essays with the Intelligent Essay Assessor." In Automated Essay Scoring: A Cross-Disciplinary Perspective. Lawrence Erlbaum Associates.
17. Lewis, Patrick et al. 2020. "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks." NeurIPS.
18. Li, Y. et al. 2020. "BERT-Based Automated Essay Scoring." Proceedings of the International Conference on Educational Data Mining.
19. Liu, Pengfei, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. "Pre-Train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing." ACM Computing Surveys 55 (9): 1–35.
20. Liu, Yinhan et al. 2019. "RoBERTa: A Robustly Optimized BERT Pretraining Approach." arXiv Preprint arXiv:1907.11692.
21. Loshchilov, Ilya, and Frank Hutter. 2019. "Decoupled Weight Decay Regularization." International Conference on Learning Representations (ICLR).
22. Mathias, Sandeep, and Pushpak Bhattacharyya. 2018. "Why Stop at Two? Multi-Target Automated Essay Scoring." Proceedings of LREC.
23. OpenAI. 2023. "GPT-4 Technical Report." arXiv Preprint arXiv:2303.08774.
24. OpenAI. 2024. "OpenAI Embeddings Guide." Available at: https://platform.openai.com/docs/guides/embedding (Accessed: 28 July 2026).
25. Page, Ellis B. 1966. "The Imminence of Grading Essays by Computer." Phi Delta Kappan 47 (5): 238–43.
26. Pearson, Karl. 1895. "Notes on Regression and Inheritance in the Case of Two Parents." Proceedings of the Royal Society of London 58: 240–42.
27. Pennington, Jeffrey, Richard Socher, and Christopher Manning. 2014. "GloVe: Global Vectors for Word Representation." EMNLP, 1532–43.
28. Reimers, Nils, and Iryna Gurevych. 2019. "Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks." EMNLP-IJCNLP, 3982–92.
29. Shermis, Mark D., and Ben Hamner. 2012. "Contrasting State-of-the-Art Automated Scoring of Essays." Annual National Council on Measurement in Education.
30. Spearman, Charles. 1904. "The Proof and Measurement of Association Between Two Things." American Journal of Psychology 15 (1): 72–101.
31. Taghipour, Kaveh, and Hwee Tou Ng. 2016. "A Neural Approach to Automated Essay

Scoring." Proceedings of EMNLP, 1882–91.
32. Vaswani, Ashish et al. 2017. "Attention Is All You Need." NeurIPS.
33. Wu, Yonghui et al. 2016. "Google's Neural Machine Translation System: Bridging the Gap Between Human and Machine Translation." arXiv Preprint arXiv:1609.08144.