Representation-to-Creativity (R2C): Automated Holistic Scoring Model for Essay Creativity

Deokgi Kim, Joonyoung Jo, Byung-Won On[*]
Dept. of Software Science & Engineering
Kunsan National University
{thekey1220,joon0zo1022,bwon}@kunsan.ac.kr

Ingyu Lee
Inst. of Info. & Comm.
Yeungnam University
inleeatyu@yu.ac.kr

Abstract

Despite active research on Automated Essay Scoring (AES), there is a noticeable scarcity of studies focusing on predicting creativity scores for essays. In this study, we develop a new essay rubric specifically designed for assessing creativity in essays. Leveraging this rubric, we construct ground truth data consisting of 5,048 essays. Furthermore, we propose a novel self-supervised learning model that recognizes cluster patterns within the essay embedding space and leverages them for creativity scoring. This approach aims to automatically generate a high-quality training set, thereby facilitating the training of diverse language models. Our experimental findings indicated a substantial enhancement in the assessment of essay creativity, demonstrating an increase in F_1-score up to 58% compared to the primary state-of-the-art models across the ASAP and AIHUB datasets.

1 Introduction

Technological advancements have precipitated an era of rapid transformation, rendering traditional knowledge and experience swiftly obsolete and underscoring the limitations of rote memorization. In this context, creativity education becomes crucial, equipping individuals to handle unpredictability through diverse experiences, divergent thinking, and critical thinking skills necessary for discerning valuable insights among extensive information. Writing, as a tool for cultivating creativity, facilitates the logical organization of thoughts, the generation of original ideas, and effective self-expression.

A key area of interest in natural language processing is the computational evaluation of writing, particularly known for the Automated Essay Scoring (AES) problem, which has been extensively studied. Given an essay as an input, artificial intelligence (AI) models predict its accurate score. Initially, regression (Attali and Burstein, 2006; Phandi et al., 2015) or ranking-based models (Chen and He, 2013) were proposed, followed by machine learning and neural network models (Taghipour and Ng, 2016; Dong et al., 2017; Ke and Ng, 2019). More recently, Transformer models (Yang et al., 2020; Wang et al., 2022) and Large Language Model (LLM)-based essay scoring models (Han et al., 2023) have been presented.

To tackle the AES problem, numerous AI models undergo training and evaluation using well-known and publicly available benchmark datasets such as Automated Student Assessment Prize (ASAP) (The Hewlett Foundation, 2012) and AI-HUB (The Open AI Dataset Project, 2021). The ASAP dataset comprises over 12,900 essays written by students in grades 7 to 10, covering a total of 8 prompts. The AIHUB dataset includes over 50,413 essays written by students in grades 4 to 6 in elementary school, grades 1 to 3 in middle and high school. Each essay is carefully scored by three writing experts containing a total of 14 prompts. Please, see the details in Appendix A.

Nevertheless, there exists a deficiency in evaluation metrics capable of accurately quantifying creativity within current rubrics. For example, in the ASAP dataset, the grading system is based on a coarse-grained rubric, containing multifaceted aspects such as ideas, content, organization, style, voice, and language rules related to spelling and grammar, thereby reflecting the overall score. In contrast to the ASAP dataset, the rubric employed in AIHUB adopts a relatively fine-grained approach, encompassing 11 multifaceted dimensions. This rubric places particular emphasis on evaluating two key aspects: (1) expression proficiency, which includes grammar accuracy, vocabulary usage, and skillful sentence construction, and (2) compositional ability, which involves assessing paragraph and transition coherence, structural con-

[*] Corresponding author.

Findings of the Association for Computational Linguistics: NAACL 2025, pages 5272–5290
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

Category	Trait	Description	Score
Content (C)	Novelty of ideas and content (C1)	The idea is ingenious. The argument or evidence is fresh and novel.	1-5
	Richness of content (C2)	The content is varied, including diverse evidence and examples. The content is specific.	1-5
	Logicality of content (C3)	The evidence is valid and reasonable. The author responds appropriately, taking into account expected counter arguments.	1-5
Organization (O)	Originality of structure (O1)	The narrative approach is unique. The introduction and conclusion have been impressively structured.	1-5
	Cohesiveness of structure (O2)	Paragraphs are well separated, and the structure is systematic. The text flows smoothly and is cohesive.	1-5
Expression (E)	Originality of expression (E1)	The expression is original and not clichéd, featuring creative metaphors, witty quotations, and literary expressions.	1-5
	Appropriateness of expression (E2)	The author uses accurate and objective words that fit the grammar.	1-5
Author's voice (V)	Perspective and personality (V1)	The author's new perspective on the topic is revealed. The author's personality is revealed through their writing style and other elements.	1-5
Readers' response (R)	Fun and persuasiveness (R1)	The writing is fun and interesting. The author's argument is persuasive and the reader is impressed.	1-5
	Creativity score		Average

Table 1: Proposed rubric for scoring essay creativity.

sistency, and adequacy of length. The rubrics in currently available benchmark datasets prioritize the assessment of proficient writing skills over comprehensive measures of creativity.

In our study, in collaboration with writing experts who are high school teachers and college professors, we developed a new essay rubric that includes criteria or metrics for evaluating creativity in a multifaceted manner based on the existing rubric (Han, 2015) as shown in Table 1. Subsequently, three evaluators, who had undergone training in writing assessment, assigned creativity scores to 5,048 essays across a total of 11 prompts randomly selected from the ASAP and AIHUB datasets. Throughout this paper, we refer to this dataset as the ground truth Λ, which is used to demonstrate the effectiveness of the proposed method. Further details can be found in Section 3.1.

On the other hand, creativity is a highly subjective and lacks a clear definition or standard, which makes it difficult to develop a model that directly quantifies. However, by mining essay vectors in essay embedding space, we can automatically detect creative and non-creative essays from a given prompt. We call it *Representation-to-Creativity (R2C)*. These detected essays can then be used to fine-tune various language models.

Specifically, we propose a novel self-supervised learning model to identify creative essays (as positive samples), conventional essays (as neutral samples), and non-creative essays (as negative samples), given an essay prompt from ASAP and AIHUB. Furthermore, our method quantitatively assesses the creativity scores of each essay sample within a range of 1 to 5 points. Our approach facilitates automatic generation of a high-quality training dataset, and enables accurate prediction of essay creativity scores by various state-of-the-art language models including AES models, Generative Adversarial Networks (GANs), and Transformer-based models like BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), DeBERTa (He et al., 2020), and ELECTRA (Clark et al., 2020).

Our key idea is to represent all essays as essay embedding vectors and then cluster the vectors. Afterwards, we find the cluster α that is likely to contain the most conventional essays among the output cluster set ρ. Formally, we are looking for a cluster such that $\alpha = argmax_{c_i \in \rho}|c_i|$, where $|c_i|$ is the number of vectors in c_i. One cluster $c_i \in \bar{\rho} = \rho - \alpha$ is the most creative cluster κ if c_i is the furthest from α in the vector space. To measure the distance between α and $c_i \in \bar{\rho}$, we use the cross entropy. In particular, through our experiments, we observed that the variance of the vectors in a cluster is highly correlated with the diversity of the ideas or creativity of the content. When the content of the essay is lacking diversity, the variance of the cluster is tended to be low. Therefore, based on these observations, we propose a new measure called **creativity-aware distance** that calculates cross entropy considering a normalized intra-cluster variance as a weight. To the best of our knowledge, our proposed approach is the first study to utilize cluster patterns within the essay embedding space for creativity scoring.

Our technical contributions in this study are twofold:

- To date, there have been a few studies on scoring the creativity of essays. To address this issue, we first constructed ground truth dataset comprising approximately 5,048 essays. The average Kendall score with three evaluators was 0.827 which indicates a high level of agreement among them.

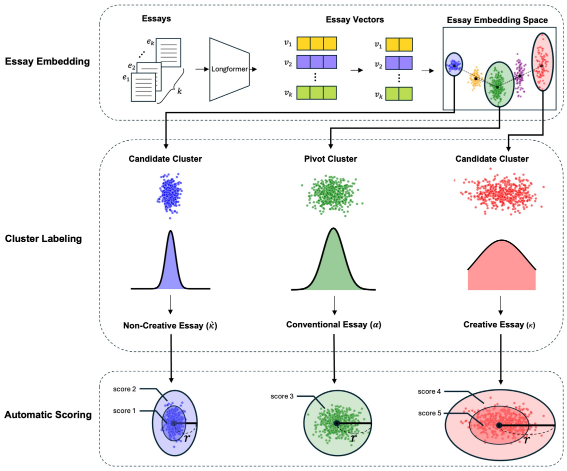


Figure 1: Overview of the proposed method.

- We propose a novel self-supervised learning model aimed at facilitating the automatic generation of a high-quality training set, thereby enabling accurate prediction of essay creativity scores by various language models. Our experimental results demonstrated a notable enhancement in the assessment of essay creativity, achieving up to a 58% increase in F_1-score compared to leading state-of-the-art methods when evaluated on ASAP and AIHUB.

2 Related Works

Various studies have been conducted focusing primarily on "Novelty Detection" in specific domains such as papers, patents, designs, and images rather than assessing creativity in essays. Ghosal et al. (Ghosal et al., 2018b) proposed a new benchmark dataset, called TAP-DLND (Document Level Novelty Detection), and applied a Relative Document Vector-based CNN model (Ghosal et al., 2018a) to detect novelty in document level. Subsequently, new techniques for event or story detection (Christophe et al., 2019), an humor and metaphor detection using a Gaussian Process (Simpson et al., 2019), a novel idea detection using machine learning (Amplayo et al., 2019) were proposed. Later, CNN model with conjunctive clause-based Tsetlin Machine for novelty detection (Nandi and Basak, 2020; Bhattarai et al., 2020), SVM model based on word embedding vectors for novelty detection in patent data (Chikkamath et al., 2020), and heuristic method for novel idea detection in academic paper (Doboli et al., 2020) have also been suggested. Particularly, the approach by Ghosal et al. (2018a) resembles our study as it encodes related and relative information of all text data at the document level and detects novelty based on this. However, CNN only considers the local context, which limits its ability to grasp the entire essay context. Moreover, the authors constructed an in-house dataset named TAP-DLND for novelty detection which makes it incomparable with our dataset.

Only a few studies explored on scoring essay creativity. Beaty and Johnson (2021) utilized a deep learning model based on diverse word embedding vectors and suggested that greater distances between word vectors indicate higher creativity. Liang et al. (2021) employed a GAN model for creative essay recommendation by calculating the distance between the generated essay and the original one to gauge creativity. Kuznetsova et al. (2013) presented statistical explorations to understand the characteristics of word combinations in order to quantitatively measure creativity. Furthermore, Lee et al. (2023) delved into the creativity of essay expression by proposing various rare token extraction methods and enhanced BERT model performance by pre-training with rare tokens. Recently, initial studies have introduced creative natural language generation using large language models (LLMs), leveraging their remarkable capabilities (Peng et al., 2023).

Overall, existing studies focus on quantifying creativity using word tokens such as word masking (Liang et al., 2021) and word representation (Beaty and Johnson, 2021; Lee et al., 2023), neglecting the importance of overall context in essay creativity. On the other hand, our study recognizes cluster patterns within the essay embedding space and leverages them for creativity scoring. The latter enables the automatic detection of creative, non-creative, and conventional essays, as well as the fine-tuning of pre-trained language models.

3 Main Proposal

In Section 3.1, we describe the process of constructing ground truth data, which is used to demonstrate the effectiveness of the proposed method. In Section 3.2, we provide a detailed explanation of the proposed method that automatically constructs high-quality training data needed to train various language models, including BERT, RoBERTa, ELECTRA, DeBERTa, GAN, and AES models.

3.1 Construction of Ground Truth Data

To construct the ground truth data, essays were randomly sampled from both ASAP and AIHUB datasets, resulting in a total of 5,048 essays sampled, with 1,800 from ASAP and 3,248 from AIHUB. To obtain detailed information about the essay samples, please refer to Table 11 in Appendix B.

Using Table 1, three evaluators, who had trained for a writing assessment, independently reviewed each essay and assigned a score ranging from 1 to 5 for each trait. The holistic creativity score of an essay is the average value of all traits. Agreement among the evaluators was reached through consensus, and in instances of disagreement, the scores assigned by the evaluators were averaged and rounded. Essays demonstrating high creativity were awarded a score of 5, while those exhibiting low creativity received a score of 1.

The average creativity score is 2.94, with the scores for prompts 1 to 11 in Table 11 ranging from 2.83 to 3.83. We also measured the Kendall correlation coefficient to ensure the reliability of the three evaluators (Kendall, 1938). The average Kendall score, indicating a very strong correlation of 0.8 or higher across all prompts, was 0.827. For further details, please refer to Appendix B.

3.2 Self-Supervised Learning Model for Automatic Construction of High-Quality Training Data on Essay Creativity Scoring

The proposed model comprises four steps, as depicted in Figure 1. In the first step, all essays are represented as essay embedding vectors. In the second step, the vectors are clustered into a set of clusters ρ. In the next step, the cluster with the most conventional essays α as well as the most creative cluster κ and the least creative cluster $\grave{\kappa}$ are automatically identified among ρ. In the final step, to train various language models, a training set is constructed from α, κ, and $\grave{\kappa}$, consisting of essays paired with their creativity scores ranging from 1 to 5.

3.2.1 Essay-to-Vector (E2V)

The purpose of E2V is to represent the i-th essay e_i in a set of essays as its corresponding essay vector v_i through Longformer (Beltagy et al., 2020) and Principal Component Analysis (PCA) (Wold et al., 1987).

Since most essays contain more than 512 tokens [1], Longformer model (Beltagy et al., 2020) is used instead of the BERT model in this work. Note that it efficiently handles long sequences up to 4,096 tokens by combining local and global attention mechanisms, enhancing its effectiveness in modeling large-scale texts within the Transformer architecture framework.

Furthermore, through Principal Component Analysis (PCA) (Wold et al., 1987), high-dimensional essay vectors represented by Longformer model are transformed into a lower-dimensional space, retaining essential information while filtering out noise. This dimensionality reduction process contributes to enhancing the accuracy of clustering essay vectors.

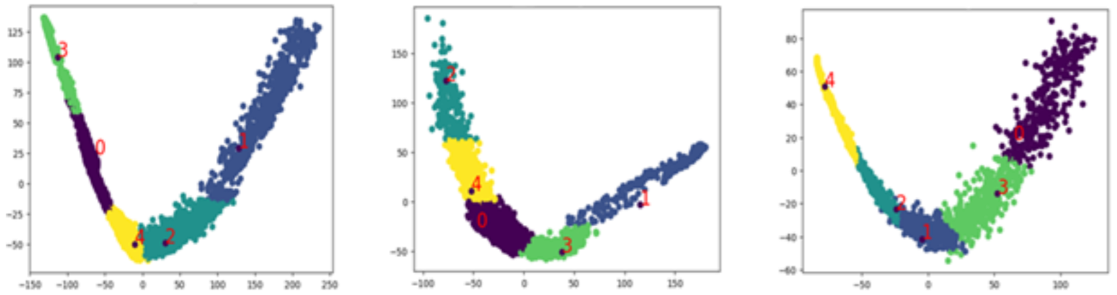
Technically, the dimensionality of vectors produced by Longformer model is 768, and the dimensionality of the lower-dimensional space is 2. To find the optimal number of dimensions, we conducted experiments by reducing the dimensionality in the following order: 500, 200, 100, 50, 10, 5, 4, 3, and 2. We found that clustering performed best when the dimensionality was reduced to 2. As the number of dimensions increased, the clustering performance worsened. This indicates that dimensionality reduction effectively removes noise without losing essential information, thereby improving clustering performance.

3.2.2 Clustering of Essay Vectors

For clustering essay vectors, we utilize the Expectation-Maximization (EM) Clustering (Dempster et al., 1977), which clusters by computing the probability of vectors v_is generated from k Gaussian mixture models. In the Expectation step (E-step), $P(c_j|v_i)$ is computed as $\frac{P(c_j)P(v_i|c_j)}{\sum_{l=1}^{k} P(c_l)P(v_i|c_l)}$. In the Maximization step (M-step), the weight $P(c_j)$, mean μ_{c_j}, and standard deviation σ_{c_j} of each cluster c_j are updated by $P(c_j) = \frac{1}{n}\sum_{i=1}^{n} P(c_j|v_i)$, $\mu_{c_j} = \frac{\sum_{i=1}^{n} v_i P(c_j|v_i)}{\sum_{i=1}^{n} P(c_j|v_i)}$, and $\sigma_{c_j} = \frac{\sum_{i=1}^{n}(v_i - \mu_{c_j})^2 P(c_j|v_i)}{\sum_{i=1}^{n} P(c_j|v_i)}$, using the updated $P(c_j|v_i)$ from the E-step.

To automatically determine the optimal number of clusters, we used the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) (Stoica and Selen, 2004) as shown in Figure 3. We set the number of clusters to 5, as this

[1] For reference, 36% of essays contained fewer than 512 tokens, while 64% had 512 tokens or more, with the maximum token count reaching 3,071.



(a) D_1 (b) D_2 (c) D_3

Figure 2: Visualization of clustering results using t-SNE. The numbers 1 through 5 correspond to clusters c_1 through c_5.

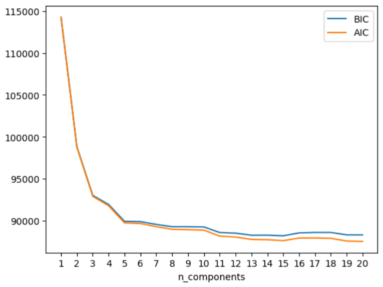


Figure 3: Estimation of the number of true clusters using AIC/BIC.

Term	Description
ρ	A set of clusters
c_i	The i-th cluster in ρ
$\lvert c_i \rvert$	# of essay vectors in c_i
α	The most conventional cluster in ρ
$\lvert\alpha\rvert$	# of essay vectors in α
κ	The most creative cluster in ρ
$\grave{\kappa}$	The least creative cluster in ρ
$\bar{\rho}$	Candidate clusters for finding κ and $\grave{\kappa}$
$ce(\alpha, c_i)$	Cross entropy between α and c_i
w_j	The j-th word
$P(w_j \in \alpha)$	Probability of w_j appearing in α
$P(w_j \in c_i)$	Probability of w_j appearing in c_i
w_α	Normalized weight of α
w_{c_i}	Normalized weight of c_i
σ^2_α	Intra-cluster variance value of α
$\sigma^2_{c_i}$	Intra-cluster variance value of c_i
e_k	The k-th essay
$P_{tf}(w_j \lvert e_k)$	Probability of w_j in e_k through Transformer
$dist(\alpha, c_i)$	Creativity-aware distance between α and c_i

Table 2: Notations for equations 1-5.

is the point where the AIC and BIC values start to level off.

3.2.3 Automatic Labeling of Conventional, Creative, and Non-creative Clusters

To create various datasets, we randomly selected 11 prompts from the ASAP and AIHUB datasets. As shown in Table 11 in Appendix B, Prompt 1 includes 1,800 essays sampled from the ASAP dataset, while the remaining prompts include 3,248 essays sampled from AIHUB. Three writing experts labeled the creativity scores for these 5,048 essays. In this work, we refer to this labeled dataset as the ground truth data Λ, which is composed of three subsets: D_1, D_2, and D_3. Specifically, D_2 consists of essays with creativity scores from AIHUB, and D_3 consists of essays with creativity scores from ASAP. D_1 is the combined set of D_2 and D_3. Each of these subsets was split into a training set and a test set in an 8:2 ratio.

Figure 2 illustrates the clustering results for each dataset. We arbitrarily assigned cluster identifiers (CIDs) to each cluster and visualized them with different colors to distinguish between different clusters. There are five clusters present in all datasets. For instance, $\rho = \{c_0, c_1, c_2, c_3, c_4\}$. Cluster patterns appear similar across all datasets.

With the notation terms from Table 2, the characteristics of the clustering results are summarized in Table 3. Interestingly, in D_1, cluster c_4 with the median value (3.08) in the Score column includes a relatively large number of essay vectors. Most essays in c_4 have creativity scores close to 3, indicating ordinary content. Similar results are observed in other datasets as well. Based on these observations, c_4 is considered as α. Formally, we define α as Equation 1.

$$\alpha = argmax_{c_i \in \rho} |c_i| \tag{1}$$

In D_1, $\bar{\rho} = \rho - \alpha = \{c_0, c_1, c_2, c_3, c_4\} - \{c_4\} = \{c_0, c_1, c_2, c_3\}$ is a set of candidate CIDs for detecting κ and $\grave{\kappa}$. Then, for $c_i \in \bar{\rho}$, $ce(\alpha, c_i)$ is measured by:

$$\begin{aligned} & ce(\alpha, c_i) \\ = & -\Sigma_{w_j} P(w_j \in \alpha) \cdot logP(w_j \in c_i) \end{aligned} \tag{2}$$

Dataset	CID	# of vectors	Score	σ^2	$dist(\alpha, c_i)$
D_1	$c_3(\grave{\kappa})$	582	1.55	677.57	3.06
	c_0	897	2.62	502.06	2.93
	$c_4(\alpha)$	1,826	3.08	204.13	-
	c_2	944	3.29	1,044.90	5.25
	$c_1(\kappa)$	783	3.57	2,965.57	7.84
D_2	$c_1(\grave{\kappa})$	438	1.60	1,085.43	3.43
	c_3	549	2.76	290.43	1.86
	$c_0(\alpha)$	1,282	3.12	288.13	-
	c_4	562	3.45	691.85	4.18
	$c_2(\kappa)$	235	4.01	1,640.87	5.66
D_3	$c_4(\grave{\kappa})$	385	1.99	457.89	2.66
	c_2	343	2.71	204.13	2.27
	$c_1(\alpha)$	417	2.93	200.49	-
	c_3	359	3.12	488.97	4.03
	$c_0(\kappa)$	296	3.56	861.56	6.26

Table 3: Characteristics of clustering results. Note that the Score column includes the average creativity scores of all essays within the cluster. The σ^2 represents the intra-cluster variance.

In Equation 2, as $P(w_j \in \alpha)$ is lower and $P(w_j \in c_i)$ is high, or vice versa, the $ce(\alpha, c_i)$ value increases. If $ce(\alpha, c_i)$ is large, then c_i is likely to be either the creative cluster or the non-creative cluster. For example, κ is either c_1 or c_3 in Figure 2(a). If c_1 is κ, then c_3 automatically becomes $\grave{\kappa}$ because both c_1 and c_3 are the farthest from α in the semantic space. To identify κ and $\grave{\kappa}$, after calculating all the cross entropy values between α and all candidates, the top-2 clusters with the highest cross entropy values are selected (i.e., c_1 and c_3 in Figure 2(a)).

Given the top-2 clusters, to accurately determine which is κ and $\grave{\kappa}$, we develop a new hypothesis. Specifically, our observations suggest that creative essays demonstrate a wealth of content and employ a varied vocabulary not typically found in conventional essays. Each creative piece exhibits its own distinctive traits. Conversely, essays lacking creativity tend to rely heavily on common, repetitive words, resulting in monotony. These essays share similar characteristics with each other. Based on these observations, our hypothesis is that *the greater the creativity within a cluster, the higher its intra-cluster variance.* For example, in Figure 2(c), c_0 corresponds to κ and c_4 corresponds to $\grave{\kappa}$. The variance of κ is much larger than that of $\grave{\kappa}$. The figures (a) and (b) present identical results.

Technically, we propose a new cross entropy measure based on normalized intra-cluster variances called creativity-aware distance. We normalize the intra-cluster variance values of the top-2 clusters as cluster weights, as shown in Equation 3.

$$w_\alpha = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_{c_i}^2}, \quad w_{c_i} = \frac{\sigma_{c_i}^2}{\sigma_\alpha^2 + \sigma_{c_i}^2} \tag{3}$$

To calculate $P(w_j \in \alpha)$ and $P(w_j \in c_i)$ in Equation 2, we first train the Transformer model using essays as input data. Then, in the inference phase, the Transformer model generates the probability distribution of word occurrences, given the k-th essay e_k, for word w_j. Based on this Transformer model, we re-formulate $P(w_j \in \alpha)$ and $P(w_j \in c_i)$ in Equation 4.

$$\begin{aligned} P(w_j \in \alpha) &= \frac{\sum_{k=1}^{l} P_{tf}(w_j|e_k)}{|\alpha|}, \\ P(w_j \in c_i) &= \frac{\sum_{k=1}^{l} P_{tf}(w_j|e_k)}{|c_i|} \end{aligned} \tag{4}$$

Finally, we define the **creativity-aware distance** between two clusters α and c_i as Equation 5.

$$\begin{aligned} &dist(\alpha, c_i) \\ &= -\Sigma_{w_j} w_\alpha P(w_j \in \alpha) \cdot w_{c_i} log P(w_j \in c_i) \end{aligned} \tag{5}$$

Table 3 also shows the distance values between α and c_i. In D_1, c_1 and c_3 are selected as the top-2 clusters using Equation 2. Because of $dist(\alpha, c_1) > dist(\alpha, c_3)$ by Equation 5, c_1 is assigned to κ, and c_3 is assigned to $\grave{\kappa}$. This creativity-aware distance does not represent an absolute value. Instead, it uses α as the pivot to quantitatively measure how far c_i is in terms of creativity. Interestingly, since $dist(\alpha, c_2)$ is also large, cluster c_2 is likely to contain many creative essays. However, c_2 can be considered a relatively less creative cluster compared to c_1.

3.2.4 Automatic Generation of Essay Creativity Scores through Cluster Labels

In the previous section, we proposed and elaborated on approaches for automatically labeling conventional, creative, and non-creative clusters. We then discuss strategies for automatically generating training data from these clusters to train various language models. Please, note that the training data consists of pairs of an essay and its creativity score ranging from 1 to 5.

Essays included in α are assigned a creativity score of 3. To assign 4 or 5 points to essays in κ, the center point κ_c of κ is calculated. If the radius of κ is r, the vectors within the circle of $\frac{r}{2}$ in κ_c are assigned to 5 points, and the vectors within $\frac{r}{2} \sim r$ are assigned to 4 points. Through the above process, 1 or 2 points are assigned to the essays in $\grave{\kappa}$.

If vectors in candidate clusters are labeled from 1 to 5, and we denote the cluster with the smallest number of vectors as c_m, where $|c_m|$ represents the number of vectors in c_m, we can sample $|c_m|$ vectors from each of the other clusters to create the final training dataset. This ensures an equal number of vectors across all clusters, preventing model bias and improving data representativeness.

4 Experimental Set-up

First, we generated essays' embedding vectors using allenai/longformer-base-4096 provided by Hugging Face. We wrote Python scripts to implement PCA and EM Clustering using scikit-learn 1.0.2. Next, we wrote a Python script that implemented Transformer model using PyTorch's nn.Module. We set d_{model} to 512, num_layers to 1, num_heads to 8, and d_{ff} to 2,048. The hyperparameters for training the Transformer model to calculate Equation 4 were as follows: The batch size was set to 8, the number of epochs was set to 10, and AdamW was utilized as the optimizer, with a learning rate of 1e-3.

Next, we wrote Python scripts to implement pre-trained language models such as BERT, RoBERTa, ELECTRA 411, and DeBERTa using open-source libraries from Hugging Face. The hyperparameters for fine-tuning the pre-trained language models are as follows: We used StratifiedKFold to accurately evaluate the models' performance on the dataset, with a train-test ratio of 8:2. We set the number of folds to 5, the batch size to 8, and the number of epochs to 5. We employed AdamW as the optimizer for fine-tuning the models, with a learning rate of 1e-6. These hyperparameters were chosen as they produced the most effective models through extensive experimentation. To implement the AES models, open-source code from GitHub was used, and experiments for Beaty and Johnson (2021) were conducted using SemDis, an open platform provided in the paper. All models were run on a high-performance workstation with an Intel Xeon Scalable Silver 4414 CPU (2.20GHz, 40 cores), 24GB RAM, and a GEFORCE RTX 3080 (11GB RAM, 4,352 CUDA cores, 7GBPS memory clock).

Model	**Supervised**			**Proposed**		
	D_1	D_2	D_3	D_1	D_2	D_3
BERT*	0.713	0.701	0.733	0.888	0.751	0.993
RoBERTa*	0.781	0.853	0.759	0.999	0.994	0.995
ELECTRA*	0.789	0.841	0.753	0.994	0.996	0.992
DeBERTa*	0.691	0.639	0.744	0.835	0.640	0.987
Wang et al. (2022)**	0.779	0.848	0.712	1.000	1.000	1.000
Xie et al. (2022)**	0.879	0.859	0.832	0.992	0.949	1.000
Average	0.772	0.790	0.756	0.951	0.888	0.995

Table 4: Results of existing supervised and proposed models for binary classification of essay creativity. * indicates language models, and ** indicates state-of-the-art AES models.

5 Experimental Results

5.1 Binary Classification of Essay Creativity

Table 4 presents the accuracy of supervised and proposed models for the D_1, D_2, and D_3 datasets. By "Supervised," we mean that existing SOTA models such as BERT, RoBERTa, ELECTRA, DeBERTa, Wang et al. (2022), and Xie et al. (2022) are trained using the ground truth data labeled by human annotation. In contrast, "Proposed" refers to training these models with the training set generated by our proposed method. Please note that the output of our proposed method is a training set where the creativity scores for each essay are automatically generated through our method.

The trained models classify a new essay e as creative or not. Across all datasets, the proposed models significantly outperform the existing supervised models. For example, in D_1, D_2, and D_3, the supervised models have average accuracies of 0.772, 0.790, and 0.756, respectively, while the proposed models achieve 0.951, 0.888, and 0.995. The proposed models improve accuracy by 23% in D_1, 13% in D_2, and 32% in D_3, indicating greater performance improvement in all datasets.

Interestingly, the state-of-the-art AES models [2] outperform main language models. For example, in D_1, the average accuracy of the supervised AES models is 0.829, whereas the supervised language models have an average accuracy of 0.744. The reason AES models outperform general language models in creativity tasks is their specialization in predicting essay scores. For instance, Wang et al. (2022) utilize multi-level essay representation (e.g., spanning words, segments corresponding to paragraphs, and documents) and propose a multi-

[2] The performance of recently proposed LLM-based AES models has been reported to be worse than that of pre-trained BERT models (Xiao et al., 2024). It is well-known that Wang et al. (2022) and Xie et al. (2022) outperform existing AES models in recent times.

Type	Model	Accuracy	Precision	Recall	F_1-score	QWK
		D_1 D_2 D_3	D_1 D_2 D_3	D_1 D_2 D_3	D_1 D_2 D_3	D_1 D_2 D_3
Supervised	BERT	0.622 0.680 0.596	0.566 0.561 0.478	0.557 0.571 0.441	0.561 0.566 0.459	0.758 0.813 0.710
	RoBERTa	0.675 0.775 0.618	0.555 0.610 0.491	0.558 0.628 0.488	0.557 0.619 0.490	0.798 0.877 0.723
	ELECTRA	0.685 0.766 0.595	0.569 0.610 0.473	0.582 0.594 0.420	0.576 0.602 0.445	0.809 0.874 0.715
	DeBERTa	0.578 0.674 0.609	0.490 0.544 0.499	0.517 0.521 0.443	0.503 0.532 0.470	0.683 0.771 0.702
	Wang et al. (2022)	0.620 0.752 0.612	0.512 0.621 0.502	0.643 0.664 0.510	0.570 0.642 0.506	0.757 0.871 0.745
	Xie et al. (2022)	0.643 0.741 0.600	0.593 0.744 0.498	0.577 0.668 0.489	0.585 0.704 0.493	0.790 0.865 0.734
	Average	0.637 0.731 0.605	0.548 0.615 0.490	0.572 0.608 0.465	0.559 0.611 0.477	0.766 0.845 0.722
Proposed	BERT	0.712 0.824 0.819	0.695 0.797 0.794	0.684 0.657 0.790	0.690 0.720 0.792	0.860 0.814 0.948
	RoBERTa	0.769 0.889 0.790	0.690 0.749 0.667	0.708 0.779 0.760	0.699 0.764 0.710	0.915 0.949 0.940
	ELECTRA	0.736 0.880 0.801	0.645 0.760 0.728	0.660 0.768 0.771	0.652 0.764 0.749	0.913 0.947 0.949
	DeBERTa	0.636 0.722 0.771	0.588 0.567 0.710	0.578 0.537 0.741	0.583 0.552 0.725	0.821 0.656 0.923
	Wang et al. (2022)	0.794 0.902 0.779	0.747 0.803 0.747	0.663 0.782 0.742	0.703 0.793 0.745	0.940 0.963 0.941
	Xie et al. (2022)	0.789 0.833 0.837	0.757 0.755 0.814	0.756 0.739 0.809	0.756 0.747 0.811	0.920 0.906 0.957
	Average	0.739 0.842 0.800	0.687 0.739 0.743	0.675 0.710 0.769	0.681 0.718 0.755	0.895 0.873 0.943

Table 5: Results of existing supervised and proposed models for scoring essay creativity on a scale of 1 to 5.

task loss function considering both regression and ranking. Xie et al. (2022) propose integrating contrastive learning to group similar essays closely together while placing dissimilar essays farther apart, coupled with a ranking technique to predict relative scores between essays and reference essays. In Proposed, Wang et al. (2022) achieve a slight improvement in accuracy compared to Xie et al. (2022). For instance, in dataset D_1, Wang et al. (2022) in Supervised score 0.779, while with the proposed method, it reaches 1.0. On the other hand, Xie et al. (2022) in Supervised yield 0.879, and with the proposed method, it achieves 0.992. With the proposed model, it appears that the multi-level essay representation learning provides greater assistance in performance enhancement compared to contrastive learning.

RoBERTa and ELECTRA models demonstrate the best performance among language models. This is attributed to RoBERTa optimizing BERT's pretraining method by incorporating large-scale data, while ELECTRA benefits from its application of the GAN concept. On the contrary, DeBERTa exhibits the lowest performance due to its specialization in handling longer sequences through improvements in BERT's attention mechanism. However, in dataset D_3, where essays are relatively longer compared to D_2, DeBERTa performs similarly to other language models.

5.2 Multi Classification of Essay Creativity

Table 5 presents the results of language models and AES models trained on essays labeled with scores ranging from 1 to 5. In Table 4, since the models are performing binary classification, their effectiveness was assessed using accuracy metrics. However, in Table 5, as it involves multi-class classification, models were evaluated using precision, recall, and F_1-score. Additionally, the Quadratic Weighted Kappa (QWK), commonly used in AES, was employed. Overall, the results are similar to those in Table 4. The proposed models improve F_1-score by 22% in D_1, 18% in D_2, and 58% in D_3.

Note that we conducted two types of experiments. Clustering was performed without distinguishing prompts in D_1 and D_2, whereas in D_3, clustering was done separately for essays from each prompt. As shown in Table 3, the variance of clusters in D_1 and D_2 is much higher compared to D_3. This high variance reflects that essays from different prompts coexist within the same cluster. Additionally, as seen in Table 5, the performance of most proposed models using D_3 significantly outperforms those using D_1 and D_2. This indicates that clustering is more effective when performed separately for essays of each prompt rather than without distinguishing prompts.

In a nutshell, after clustering essay vectors in the embedding essay space, we identify the conventional cluster, which contains the most vectors, and automatically detect the creative and non-creative clusters that are the furthest from this conventional cluster. Furthermore, we assign scores to the essays based on their distance from the center of the clusters. This self-supervised approach based on representation learning not only automatically generates training data for assessing essay creativity but also shows a strong correlation with expert-crafted ground truth data.

5.3 Comparison of Proposed Model to Existing Creativity Assessment Models

Table 6 shows the accuracy of existing computational creativity models and proposed models. Computational creativity is a highly challenging

Model	Liang (2021)	Beaty (2021)	Proposed model
BERT	-	0.794	0.993
RoBERTa	-	0.753	0.995
ELECTRA	-	0.767	0.992
DeBERTa	-	0.763	0.987
Wang et al. (2022)	-	0.798	1.000
Xie et al. (2022)	-	0.924	1.000
Average	0.653*	0.799	0.995

Table 6: Results of existing computational creativity models and proposed models for binary classification of essay creativity in D_3. *We report only the Liang model's evaluation results, as it does not utilize language models or AES models.

Model	MaskGAN	Proposed model
BERT	0.917	0.993
RoBERTa	0.968	0.995
ELECTRA	0.970	0.992
DeBERTa	0.973	0.987
Wang et al. (2022)	0.938	1.000
Xie et al. (2022)	0.968	1.000
Average	0.956	0.995

Table 7: Results of existing models with training data generated by MaskGAN and proposed method for binary classification of essay creativity in D_3.

problem, with only a small number of computational creativity models proposed to date. Even these previous models are considered relatively unsophisticated. Liang et al. (2021) involve masking parts of essays, training a GAN to predict the masked text, and analyzing the generated essays to identify creativity. Beaty and Johnson (2021) utilize diverse word embedding vectors, suggesting that greater distances between word vectors indicate higher creativity. The proposed model improves the accuracy of the models by Liang et al. (2021) and Beaty and Johnson (2021) by 52% and 25%, respectively.

5.4 Comparison of Proposed Model to GAN

Table 7 shows the accuracy of existing language models and AES models using training data generated by both the GAN model and the proposed model.

Training data can also be generated using the existing GAN model such as SeqGAN (Yu et al., 2017) and MaskGAN (Fedus et al., 2018). While SeqGAN focuses on generating new sequences that resemble the training data, MaskGAN concurrently generates sequences and fills in missing parts, enabling a better grasp of context and the creation of more novel documents. Therefore, considering the research objectives, we demonstrate the superiority of the proposed model through a comparison with MaskGAN as a baseline.

From the ground truth data Λ, essays e_is scoring 4-5 are sampled as creative essays into a set s_p. During training, using s_p, the generator generates tokens based on feedback from the discriminator, while the discriminator provides feedback by comparing the generated tokens with actual tokens. Through this process, the generator and discriminator compete with each other and learn iteratively. In the generation phase, the generator of the MaskGAN model produces a new creative essay when given m tokens. Another MaskGAN model generates non-creative essays in a similar manner.

Despite boosting the accuracy of existing language models and AES models, MaskGAN falls short compared to the proposed model, which achieves an average of 4% higher accuracy. Additionally, our proposed method automatically generates training data, whereas MaskGAN is limited by its need for manually crafted training data.

5.5 Summary of Additional Experiments and Case Studies

Due to space constraints, only the key points of the other experiments are summarized here. For detailed information, see Appendix C. First, the proposed models also demonstrate good performance compared to supervised models for each category in the essay rubric (Table 1) when predicting the creativity scores of essays. Subsequently, we discuss the details using three real examples: creative, non-creative, and conventional essays in a case study. Finally, using the GPT-4o LLM, we calculate the creativity scores for creative, non-creative, and conventional essays sampled using the proposed method and demonstrate their similarity to the actual creativity scores of the essays.

6 Conclusion

This paper presents a novel self-supervised learning model that goes beyond the actively researched problem of Automated Essay Scoring. This model aims to automatically construct high-quality training data to enable various language and AES models to accurately predict essay creativity scores. Based on the hypothesis that creative essays are distant from conventional ones in the essay embedding space and that creative essays exhibit high intra-cluster variance due to their diversity, our model demonstrates overwhelming performance across the main language and AES models.

Limitations

The purpose of this research is focused on generating creativity scores ranging from 1 to 5 for given essays, addressing the holistic scoring problem. Considerations such as trait-based scoring, cross-prompt essay scoring, feedback generation, and Large Language Model (LLM)-based AES models, which have emerged recently in Automated Essay Scoring research, fall beyond the scope of this paper. These specific areas would become important topics within our future research focus on essay creativity.

Ethics Statement

The proposed method does not consider any ethical aspects of the student essay datasets. However, the algorithm can be effected by the training data without any notice. By collecting more dataset from diverse ethic groups, we can avoid potential risks in ethics.

Acknowledgements

This work was supported in part by the National Research Foundation of Korea (NRF) Grant by Korean Government through the Ministry of Science and ICT (MSIT) under Grant RS-2022-NR069683.

References

Reinald Kim Amplayo, Seung-won Hwang, and Min Song. 2019. Evaluating research novelty detection: Counterfactual approaches. In *Proceedings of the Thirteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-13)*, pages 124–133, Hong Kong. Association for Computational Linguistics.

Yigal Attali and Jill Burstein. 2006. Automated essay scoring with e-rater® v.2. *The Journal of Technology, Learning and Assessment*, 4(3).

Roger E. Beaty and Dan R. Johnson. 2021. Automating creativity assessment with semdis: An open platform for computing semantic distance. In *Behavior Research Methods*, volume 53, pages 757–780.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer.

Bimal Bhattarai, Ole-Christoffer Granmo, and Lei Jiao. 2020. Measuring the novelty of natural language text using the conjunctive clauses of a tsetlin machine text classifier.

Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, and Chien-Sheng Wu. 2024. Art or artifice? large language models and the false promise of creativity. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–34.

Hongbo Chen and Ben He. 2013. Automated essay scoring by maximizing human-machine agreement. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1741–1752.

Renukswamy Chikkamath, Markus Endres, Lavanya Bayyapu, and Christoph Hewel. 2020. An empirical study on patent novelty detection: A novel approach using machine learning and natural language processing. In *2020 Seventh International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 1–7.

Clément Christophe, Julien Velcin, Jairo Cugliari, Philippe Suignard, and Manel Boumghar. 2019. How to detect novelty in textual data streams? a comparative study of existing methods.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators.

Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. 1977. Maximum likelihood from incomplete data via the em - algorithm plus discussions on the paper.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

Simona Doboli, Jared Kenworthy, Paul Paulus, Ali Minai, and Alex Doboli. 2020. A cognitive inspired method for assessing novelty of short-text ideas. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Fei Dong, Yue Zhang, and Jie Yang. 2017. Attention-based recurrent convolutional neural network for automatic essay scoring. In *Proceedings of the 21st Conference on Computational Natural Language Learning*, pages 153–162.

William Fedus, Ian Goodfellow, and Andrew M. Dai. 2018. Maskgan: Better text generation via filling in the. In *Proceedings of The Sixth International Conference on Learning Representations*.

Tirthankar Ghosal, Vignesh Edithal, Asif Ekbal, Pushpak Bhattacharyya, George Tsatsaronis, and Srinivasa Satya Sameer Kumar Chivukula. 2018a. Novelty goes deep. a deep neural solution to document level novelty detection. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2802–2813, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Tirthankar Ghosal, Amitra Salam, Swati Tiwari, Asif Ekbal, and Pushpak Bhattacharyya. 2018b. TAP-DLND 1.0 : A corpus for document level novelty detection. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Jieun Han, Haneul Yoo, Junho Myung, Minsun Kim, Hyunseung Lim, Yoonsu Kim, Tak Yeon Lee, Hwajung Hong, Juho Kim, So-Yeon Ahn, and Alice Oh. 2023. Fabric: Automated scoring and feedback generation for essays.

Jung-ran Han. 2015. Development of standards for creativity assessment of narrative text and persuasive text. In *Graduate School of Korea National University of Education*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention.

Zixuan Ke and Vincent Ng. 2019. Automated essay scoring: A survey of the state of the art. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pages 6300–6308.

Maurice G Kendall. 1938. A new measure of rankcorrelation. In *Biometrika*, pages 81–93.

Polina Kuznetsova, Jianfu Chen, and Yejin Choi. 2013. Understanding and quantifying creativity in lexical composition. In *Proceedings of the Conference on Emprical Methods in Natural Language Processing*, pages 1246–1258.

Youbin Lee, Deokgi Kim, Byung-Won On, and Ingyu Lee. 2023. A comparative analysis of the effectiveness of rare tokens on creative expression using ramBERT. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10063–10077, Toronto, Canada. Association for Computational Linguistics.

Guoxi Liang, Byung-Won On, Dongwon Jeong, Ali Asghar Heidari, Hyun-Chul Kim, Gyu Sang Choi, Yongchuan Shi, Qinghua Chen, and Huiling Chen. 2021. A text gan framework for creative essay recommendation. *Knowledge-Based Systems*, 232:107501.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Dipannyta Nandi and Rohini Basak. 2020. A quest to detect novelty using deep neural nets. In *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pages 1–7.

Nanyun Peng, He He, Tuhin Chakrabarty, and Vishakh Padmakumar. 2023. Tutorial on creative natural language generation. In *Proceedings of the Conference on Emprical Methods in Natural Language Processing*.

Kian Phandi, Ming A. Chai, and Hwee Tou Ng. 2015. Flexible domain adaptation for automated essay scoring using correlated linear regression. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 431–439.

Edwin Simpson, Erik-Lân Do Dinh, Tristan Miller, and Iryna Gurevych. 2019. Predicting humorousness and metaphor novelty with Gaussian process preference learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5716–5728, Florence, Italy. Association for Computational Linguistics.

Petre Stoica and Yngve Selen. 2004. Model-order selection: a review of information criterion rules. *IEEE Signal Processing Magazine*, 36(47).

Kaveh Taghipour and Hwee Tou Ng. 2016. A neural approach to automated essay scoring. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1882–1891.

The Hewlett Foundation. 2012. Training, development, and validation data for automated essay scoring (asap). In *https://www.kaggle.com/c/asap-aes/data*.

The Open AI Dataset Project. 2021. Essay writing evaluation dataset. In *https://www.aihub.or.kr*.

Yongjie Wang, Chuang Wang, Ruobing Li, and Hui Lin. 2022. On the use of bert for automated essay scoring: Joint learning of multi-scale essay representation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3416–3425, Seattle, United States. Association for Computational Linguistics.

Svante Wold, Kim Esbensen, and Paul Geladi. 1987. Principal component analysis. In *Chemometrics and Intelligent Laboratory Systems*, pages 37–52.

Changrong Xiao, Wenxing Ma, Sean Xin Xu, Kunpeng Zhang, Yufang Wang, and Qi Fu. 2024. From automation to augmentation: Large language models elevating essay scoring landscape.

Jiayi Xie, Kaiwei Cai, Li Kong, Junsheng Zhou, and Weiguang Qu. 2022. Automated essay scoring via pairwise contrastive regression. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2724–2733, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Ruosong Yang, Jiannong Cao, Zhiyuan Wen, Youzheng Wu, and Xiaodong He. 2020. Enhancing automated essay scoring performance via fine-tuning pre-trained language models with combination of regression and ranking. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 1560–1569.

Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. 2017. Seqgan: Sequence generative adversarial nets with policy gradient. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pages 2852–2858.

A AIHUB Dataset

To make artificial intelligence technology easily accessible to everyone, the National Information Society Agency (NIA), a public institution in Korea, collected a benchmark dataset for Automated Essay Scoring (AES) and released it to the public in 2022. In this paper, we refer to such a dataset as the AIHUB dataset, which includes essays written in Korean by students from 4th to 6th grade in elementary school, 1st to 3rd grade in middle and high school, whose native language is Korean.

The total number of essays is 50,413. Table 8 summarizes basic statistics about the AIHUB dataset. Unlike the ASAP dataset, each essay in the AIHUB dataset is categorized into one of five categories: “writing,” “alternative proposal,” “narrative essay,” “argumentative essay,” and “agree/disagree.” Table 9 shows the number of prompts, the total number of essays, and the range of essay scores in each category. Table 10 shows the list of essay prompts in the AIHUB dataset.

Additionally, we collected 182 essays of first-year students taking a college writing course in 2022, evaluated the essay scores using the rubric from AIHUB, and added them to the original AIHUB dataset. Such essays belong to prompts 2, 6, 7, 8, 9, and 10 in Table 11. Thus, the total number of the updated AIHUB dataset is 50,595.

B Ground Truth Dataset

In this section, we discuss the characteristics of the ground truth data.

Figure 4 shows the average numbers of word tokens and sentences per essay. Referring to Table 11, *Pt1*, ..., and *Pt11* mean Prompt 1, ..., and Prompt 11. *Total* is the combined data from *Pt1* to *Pt11*. *AE* and *NE* indicate argumentative essays and narrative essays, respectively. In *Total*, the average number of word tokens and sentences per essay are about 180 and 15.

Figure 5 shows that the average creativity scores vary by prompt, but most scores are close to 3. As shown in *Pt3*$\sim$*Pt5* from AIHUB, it appears that the essay prompts from AIHUB are more difficult than those from ASAP. Students often lack sufficient background knowledge or interest in those topics such as intellectual property rights, racism, and anonymity.

Figure 6 shows the average creativity scores by category. In Table 1, there are five categories, such as Content (C), Organization (O), Expression (E), Author’s voice (V), and Reader’s response (R). The scores for C, O, E, and V are similar, but the score for R is relatively higher compared to the others. Essays that cover personal experiences or interests seem to engage readers’ attention.

Figure 7 shows the average creativity scores by trait. In Table 1, there are nine traits, such as Novelty of ideas and content (C1), Richness of content (C2), Logicality of content (C3), Originality of structure (O1), Cohesiveness of structure (O2), Originality of expression (E1), Appropriateness of expression (E2), Perspective and personality (V1), and Fun and persuasiveness (R1). The scores for C2, C3, O2, E2, and R1 are relatively higher compared to the others. Many essays written by students in ASAP and AIHUB tend to receive higher scores for logical organization, precise expression, grammatical accuracy, rich content, and reader engagement, rather than for the originality of noble ideas, structure, and expression.

C Additional Results

In this section, we discuss three experimental results. First, we carefully analyze the experimental results of supervised and proposed models for each category in the essay rubric. Next, we qualitatively evaluate the conventional, non-creative, and creative essays in a case study. Finally, we investigate creativity scores predicted by LLM.

C.1 Results of Supervised and Proposed Models for Each Category

To provide stronger justification for collapsing multiple dimensions into one broad measure of creativity (Chakrabarty et al., 2024), we evaluate the performance on supervised and proposed models for each category in the essay rubric from Table 1 using accuracy, precision/recall/F_1-score, and QWK.

For the experiments, we use D_3, which includes 5,048 essays from both ASAP and AIHUB. As shown in Table 5, we compare the performance of the supervised and proposed methods based on the backbone models: (1) ELECTRA, which demonstrates the highest performance among existing language models, and (2) Wang et al. (2022),

School	Grade	# of essays	Average # of word tokens
Elementary	4th	5,827	270
	5th	6,761	303
	6th	6,663	384
Middle	1st	6,955	352
	2nd	6,649	436
	3rd	6,189	523
High	1st	4,812	578
	2nd	4,800	637
	3rd	1,757	608
Total		50,413	455

Table 8: Statistics of the AIHUB dataset.

Category	# of prompts	# of essays	Score range
Writing	3	5,506	0-3
Alternative proposal	3	7,005	0-3
Narrative essay	3	18,275	0-3
Argumentative essay	3	11,172	0-3
Agree/Disagree	2	8,455	0-3
Total	14	50,413	0-3

Table 9: Five categories in the AIHUB dataset.

which shows the highest performance among existing AES models. As a kindly reminder, the accuracy, F_1-score, and QWK of ELECTRA are 0.595, 0.445, and 0.715 in Supervised and are 0.801, 0.749, and 0.949 in Proposed. Additionally, the accuracy, F_1-score, and QWK of Wang et al. (2022) are 0.612, 0.506, and 0.745 in Supervised and are 0.779, 0.745, and 0.941 in Proposed.

The rubric for essay creativity scores in Table 1 consists of five categories: Content (C), Organization (O), Expression (E), Author's voice (V), and Readers' response (R). Among these, Content, Organization, and Expression, which are considered the most important when evaluating essays, are used for the experiments [3].

Table 12 summarizes the experimental results. For each category, e.g., Content, the number of essays with both a score of 1 and a score of 5 is generally very small. Therefore, we combine the essays with scores of 1 and 2, and similarly, group the essays with scores of 4 and 5. Therefore, unlike Table 5, where the essay scores range from 1 to 5, the score range is now reduced to 2 to 4, resulting in three score ranges. Since the number of predicted scores has decreased from five to three when an essay is given as input, the performance of both the supervised and proposed methods is mostly improved compared to Table 5.

The ELECTRA and Wang et al. (2022) models trained with the training data generated by the proposed model outperform the ELECTRA and Wang et al. (2022) models trained using the ground truth data labeled by human annotation. For instance, in the content category, the F_1-scores of ELECTRA is 0.562 in Supervised, while that of ELECTRA is 0.835 in Proposed. Similarly, in both organization and expression categories, the F_1-scores of ELECTRA are 0.525 and 0.526 in Supervised, while those of ELECTRA are 0.858 and 0.910 in Proposed. Similarly, in the content category, the F_1-scores of Wang et al. (2022) is 0.453 in Supervised, while that of Wang et al. (2022) is 1.000 in Proposed. Similarly, in both organization and expression categories, the F_1-scores of Wang et al. (2022) are 0.483 and 0.455 in Supervised, while those of Wang et al. (2022) are 0.971 and 1.000 in Proposed.

Note that that the average score of all essays in the ground truth data is 2.96 which is close to 3 in Figure 5(a) and the standard deviation in the content category (C) is lower than those in the or-

[3] In Table 1, each category contains multiple traits. For example, Content is further divided into three traits – Novelty of ideas and content (C1), Richness of content (C2), and Logicality of content (C3). However, we cannot evaluate the performance of the supervised and proposed methods for each trait in this work. Due to the significant difference in the number of essays between scores from 1 to 5, it is not appropriate to evaluate the performance of the supervised and proposed methods for each trait. As shown in Figure 7, the number of essays corresponding to each score varies. In other words, there is a significant imbalance in the number of essays across different scores. For instance, the number of essays corresponding to the highest creativity score of 5 in C1 is significantly fewer compared to C2 and C3.

Category	No	Prompt
Writing	1	My biography
	2	An experience of losing something precious
	3	My thoughts on the universe
Alternative proposal	1	Solutions for resolving gender conflicts
	2	Strategies for improving prejudice against people with disabilities
	3	My thoughts on the issue of constructing hate facilities
Narrative essay	1	My effort for my career path
	2	Movie or book review
	3	An example of when I didn't respect others
Argumentative essay	1	My thoughts on racial discrimination
	2	My thoughts on intellectual property rights
	3	My opinion on evaluation
Agree/Disagree	1	My thoughts on singlehood
	2	My thoughts on anonymity

Table 10: Essay prompts corresponding per category in the AIHUB dataset.

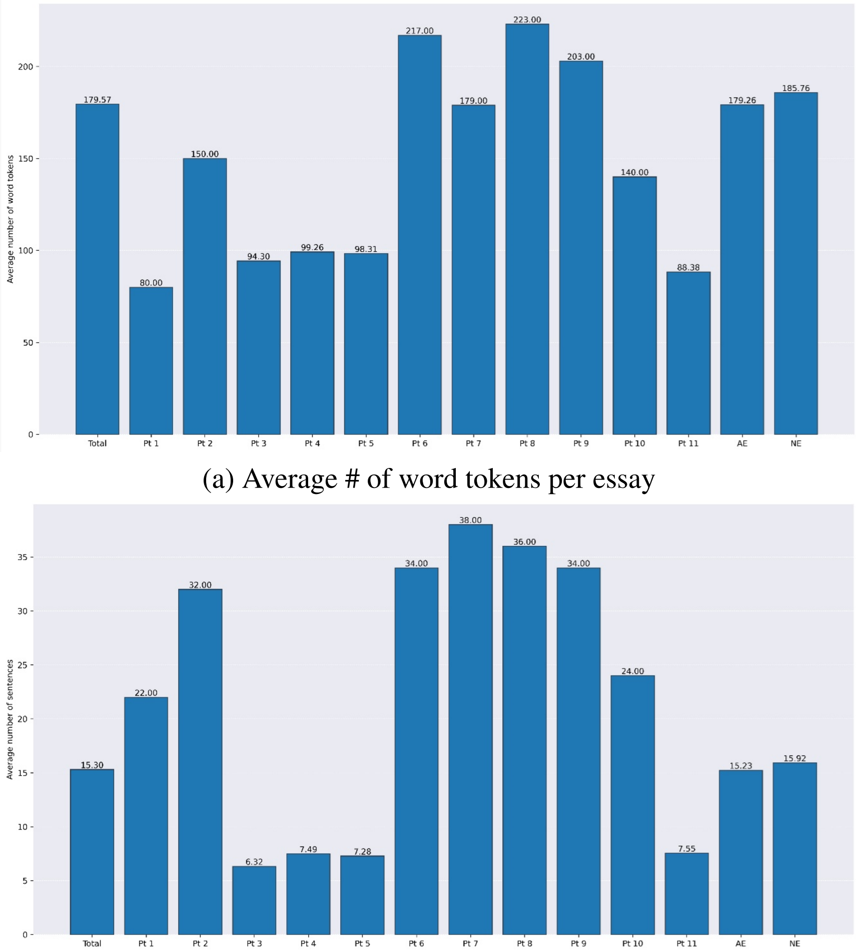
No	Type	Prompt	# of essays
1	Argumentative essay	Library censorship	1,800
2		Will AI benefit or harm future humans?	64
3		My thoughts on intellectual property rights	753
4		My thoughts on racial discrimination	757
5		My thoughts on anonymity	773
6	Narrative essay	My success or failure story	38
7		Introduction to my favorite cultural content (YouTube, Webcomics, Music, etc.)	16
8		One day without smart devices	6
9		What I am doing and can do in the age of environment pollution	9
10		Python programming language	33
11		My effort for my career path	783
		Total	5,048

Table 11: Prompts and numbers of essays in the ground truth data.

ganization and expression categories (O and E) as shown in Figure 6(b). These observations indicate that the number of essays labeled with a score of 3, which corresponds to conventional essays in the content category, is relatively higher compared to both organization and expression categories. On the other hand, in the organization and expression categories, there are more essays closed to scores of 2 or 4 than in the content category. Since there are many conventional essays in the content category, the supervised models tend to predict essays with a score of 3. As a result, the accuracy is higher compared to both the organization and expression categories. The fact that these supervised methods show high accuracy but significantly lower precision and recall suggests that they tend to predict most input essays with a score of 3.

In contrast, the proposed methods achieve similar accuracy, precision, and recall, regardless of whether it is applied to ELECTRA or Wang et al. (2022). This indicates that the proposed approach based on our proposed two hypotheses about creativity, i.e., (1) as far from α as possible and (2) through intra-cluster variances, not only automatically constructs high-quality training set for assessing essay creativity but also produces a well distinguishable distribution on predicting the creativity scores of essays.

Interestingly, the experimental results show that ELECTRA demonstrate the best performance in the expression category, while Wang et al. (2022) perform best in the expression and content categories. The reason is that both ELECTRA and Wang et al. (2022) are Transformer-based models. ELECTRA is trained using Replace Token Detection, which is similar to GAN, while Wang et al. (2022) utilize a Transformer to capture contextual information at various levels, such as words, segments, and documents. The Transformer model uses the Self-Attention mechanism to capture relationships between words and efficiently process interactions between them based on context, which results in better performance in Expression compared to Organization.



(a) Average # of word tokens per essay

(b) Average # of sentences per essay

Figure 4: Essay characteristics in the ground truth data.

C.2 A Case Study of Creative, Non-creative, and Conventional Essays

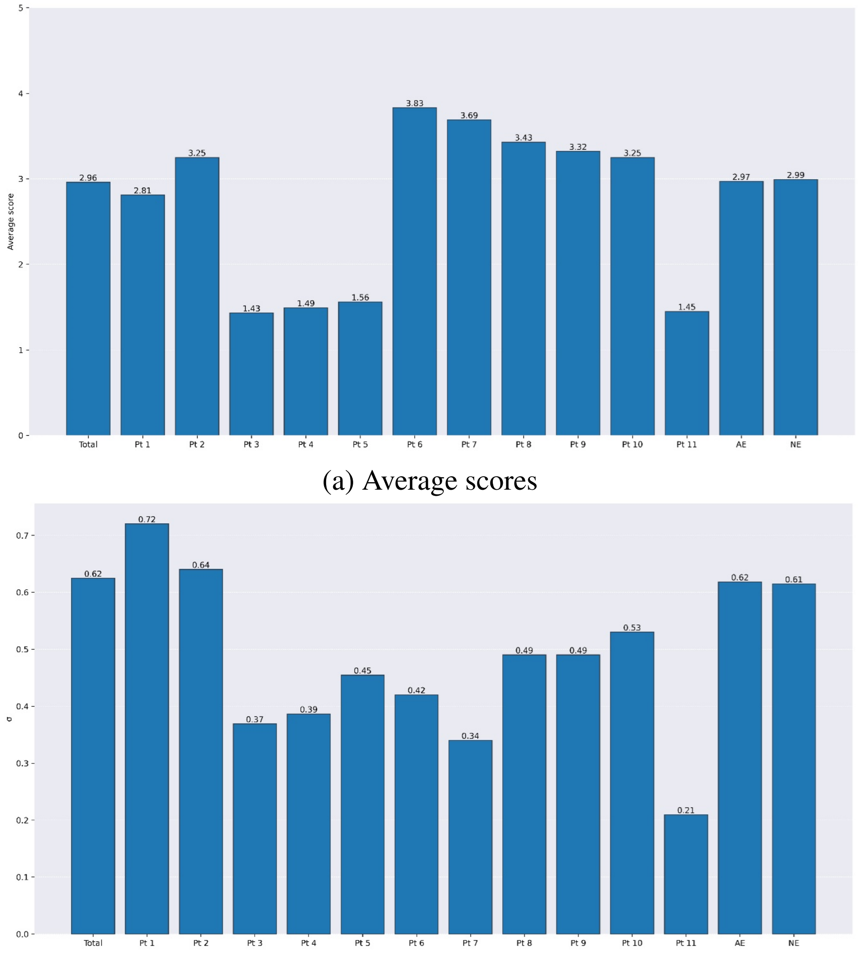
We show three real examples: creative, non-creative, and conventional essays. One example is a randomly chosen essay from the creative cluster, another is from the conventional cluster, and the third is from the non-creative cluster.

Tables 13-15 show exemplary essays e_1, e_2, and e_3 from the non-creative, conventional, and creative clusters. Prompt 1 of the ASAP is about library censorship. Access to certain materials or information is restricted for reasons such as protecting youth, societal aversion to specific topics, or political reasons. However, there is also an argument against censorship, as it limits the free access to knowledge and can hinder beneficial discussions on sensitive topics. Therefore, censorship should not be enforced.

To write a good essay on library censorship, students should demonstrate a clear understanding of the reasons and definitions of library censorship, address the perspectives of citizens with diverse opinions, discuss the actual state of library censorship and its resulting issues, and propose alternatives or policies to address the censorship problem. This showcases the writer's level of creative thinking.

In Table 13, e_1 presents a viewpoint that critical materials regarding specific races may be subject to censorship. While it provides examples relevant to the topic, the argument is somewhat cliché and lacks a robust logical structure for the writer's claims. There is little evidence of deep contemplation regarding the issue, and it does not introduce new content.

On the other hand, e_2 in Table 14 argues that library censorship is necessary to eliminate factors that may discomfort children or have a sexually negative impact, asserting that only music and educational books that can encourage students should be provided. However, it has richer expressions and a more logical structure compared to e_1, but



(a) Average scores

(b) Standard deviation

Figure 5: Average creativity scores by prompts in the ground truth data.

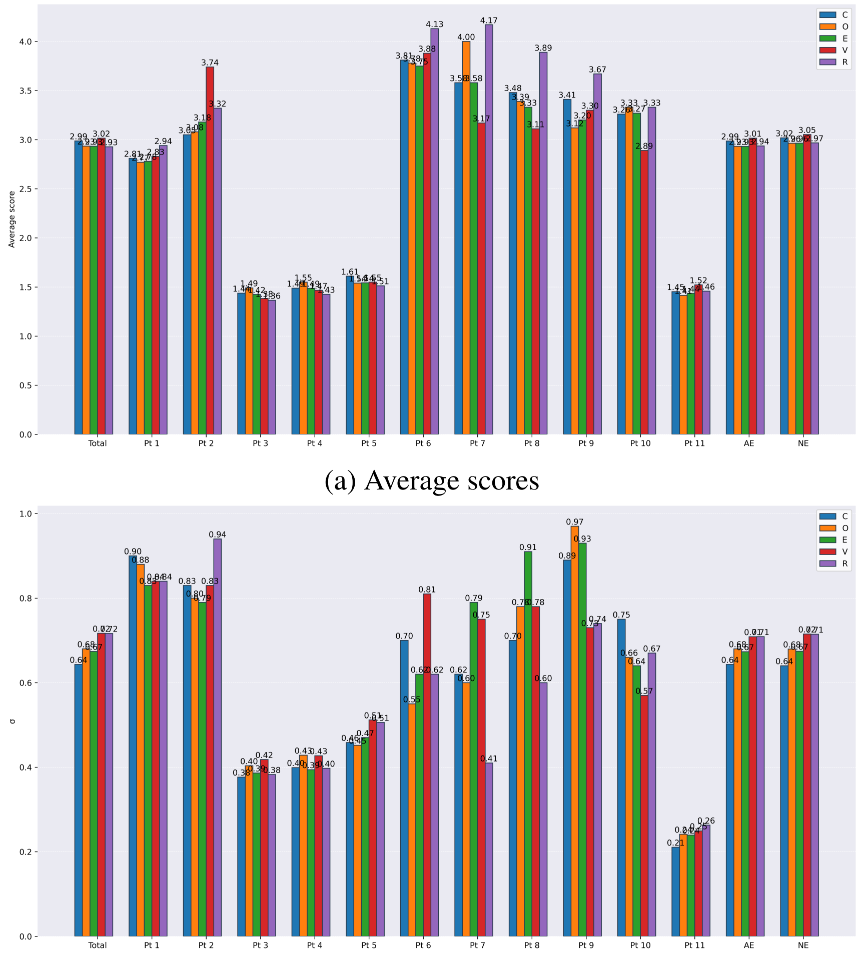
lacks unique insights or deep thinking.

In contrast, e_3 in Table 15 argues that the library censorship could undermine the core values and systems of libraries, violate citizens' constitutional rights, and damage the essence of what a library is. It strengthens the author's claims by discussing the power of knowledge, the unconstitutionality of censorship, the lack of a clear definition of offense, and the true nature of libraries from various angles. This reflects a deeper level of thinking, as well as thoughtful insights into the problem at hand.

Table 16 shows the scores for essays e_1, e_2, and e_3 on nine creativity metrics, as evaluated and agreed upon by three evaluators. The final creativity score for each essay is calculated as the average of the nine metrics. The creativity scores for them are 1.89, 3, and 4.78, respectively. The creative essay consistently scores higher than the others across all criteria in the rubric.

C.3 Creativity Scores Predicted by LLM

Table 16 also presents the creativity scores predicted by LLM, specifically GPT-4o used in our experiment. In the table, e_1, e_2, and e_3 represent non-creative, conventional, and creative essays, respectively. The average creativity scores for e_1 were 1.89 as rated by humans and 1.33 as predicted by GPT-4o. For e_2, the average scores were 3.00 by humans and 1.89 by GPT-4o. Finally, for e_3, the average creativity scores were 4.78 by humans and 4.33 by GPT-4o. The creativity scores between humans and GPT-4o were generally similar, but across all traits, human creativity scores were either higher than or equal to those of GPT-4o. Furthermore, the less creative the essay, the greater the difference in creativity scores between humans and GPT-4o. For example, in the non-creative essay e_1, for the traits of "cohesiveness of structure" and "fun and persuasiveness", the creativity score given by humans was 3, whereas GPT-4o's score was 1.



(a) Average scores

(b) Standard deviation

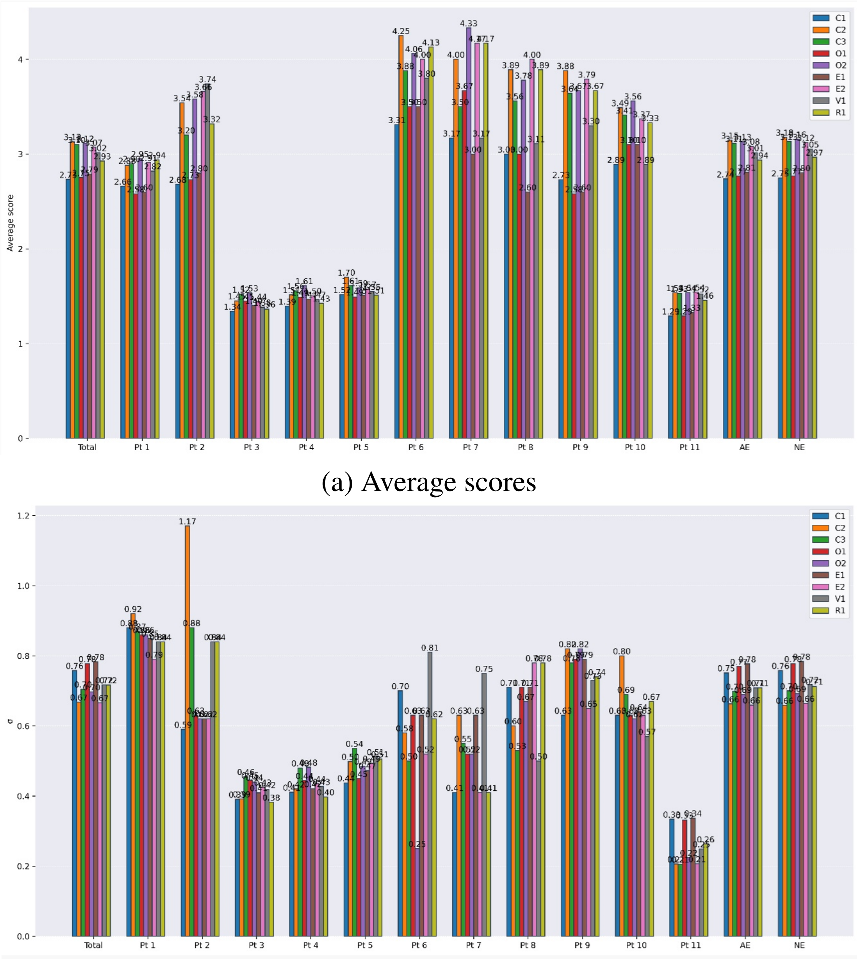
Figure 6: Average creativity scores by category in the ground truth data.

Type	Model	Category	Accuracy	Precision	Recall	F_1-score	QWK
Supervised	ELECTRA	Content	0.686	0.621	0.514	0.562	0.399
		Organization	0.525	0.530	0.520	0.525	0.436
		Expression	0.522	0.531	0.521	0.526	0.428
	Wang et al. (2022)	Content	0.661	0.414	0.501	0.453	0.216
		Organization	0.430	0.434	0.545	0.483	0.326
		Expression	0.477	0.434	0.479	0.455	0.300
Proposed	ELECTRA	Content	0.830	0.841	0.830	0.835	0.864
		Organization	0.848	0.867	0.848	0.858	0.875
		Expression	0.907	0.912	0.907	0.910	0.934
	Wang et al. (2022)	Content	1.000	1.000	1.000	1.000	1.000
		Organization	0.970	0.970	0.972	0.971	0.977
		Expression	1.000	1.000	1.000	1.000	1.000

Table 12: Results of supervised and proposed models for each category in the essay rubric from Table 1.

Many people believe that magazines, books, movies, music, posters, plus some other kind of bad stuff that parents don't want there @CAPS1 to see. Some of the stuff they might have in the are books about different colored people, maybe there might also have some terrible things about that colored skin heritage, @CAPS1 that are to young to be looking at stuff that is in these books. A lot of the @CAPS2 Libraries have these things speared around all over the store. Some of these people that get these books might need to have a good heart about what the book is saying about the skin color. If they don't like what the book and what it has to say then don't get that book. Because you can't change the way the author wrote the story, music, movie, poster. But most the people that I know that are different colored are pretty hilarious and sometimes make jokes about it. Maybe there might also be some pretty terrible things said in the book.

Table 13: Essay e_1 in the non-creative cluster.



(a) Average scores

(b) Standard deviation

Figure 7: Average creativity scores by trait in the ground truth data.

Lots of people come to the library to check out books, music, movies, and other things for entertainment. Mainly children use the library more than any other age. The library is a place where you can be humble and relaxed. No one wants to check out thinks that are offensive in any type of way. Not even books that will influence bad things. Therefore certain things shouldn't be allowed to be checked out in libraries. Certain movies shouldn't be in the library. For example if a child go to library they @MONTH1 see a movie that contains sexual activity and decided to check the movie out. The kid @MONTH1 be underage and the movie would be considered inappropriate. The child @MONTH1 get in trouble by there guardian. In reality it would be the librarian's fault for letting the child check the movie out. That is why the movies should be censored. On the cover of many magazines women as well as men bodies are exposed. That @MONTH1 cause children to be attracted to the magazine. They @MONTH1 even decided to check the magazine out. Children are influenced by the things they see. Seeing naked bodies would be considered setting a bad example. In today's world all children listen to music. The music you listen to has a huge impact on your life. The music in libraries should be censored. Libraries should have music that encourage you to go out and be something in life. In my opinion only influential music should be able to be rented out. Reading is the key to success. Libraries should have books that will help you increase your vocabulary as well as educate you. Not the books that will talk about sexual things or even make you wonder about sex. Libraries should only have history, math, and educational books. In conclusion everything in the library should be censored. A lot of children use the library. Therefore it should be age appropriate. No child needs to be put under a bad influence. Certain thinks should be removed from the shelves to keep down less confusion.

Table 14: Essay e_2 in the conventional cluster.

The common phrase ‘knowledge is power’ accurately paints the situation at hand. Through censorship in libraries, which are meant to allow free access to all media-related materials, we take away the opportunity for people to empower themselves through knowledge. The reasons for not censoring media in libraries are numerous and include, but are not limited to, unconstitutionality of censorship, the lack of proper definition for ‘offensive’, and the contradiction censorship poses to the core of what a library is. The first, and most blatant, reason for not allowing the censorship of materials in libraries is the unconstitutionality of the matter. Freedom of speech and freedom of press are both protected by the constitution upon which our country was founded. Censorship of said materials will effectively undermine both of these rights. Authors, illustrators, journalists and the slew of other individuals who compose the materials available in libraries are exercising their right to free speech through their works. They are constitutionally protected in the things they say. Notice that @CAPS1 have a right to free speech and not a right to free speech so long as no one is offended. The freedom of press is the means by which they share their ideas and concepts with the world. They are utilizing their free speech by means of press. Censoring library materials would be limiting the means by which artists can share their ideas with citizens of the @LOCATION1. This would be a clear failure to uphold the rights of the citizens. For this reason, we cannot censor libraries. The constitutional concept of equality of man plays into the situation as well. There is no adequate legal definition of ‘offend’. People could be offended by something as serious as a personal attack on their beliefs and personal entity or something as light as a personal distaste for certain mindsets and ideas. So, because all men are equal, we would have to respect the ‘offense’ taken by all people and to all degrees. This leaves every single work of art, whether it be a book, movie, magazine or any other form provided by libraries, open to the fickle definition of offensive and vulnerable to the personal feelings of every individual. If said offense taken by individuals to works is able to mark a work as ‘offensive’ and thus make it open to removal from libraries, we are effectively allowing citizens to undermine the rights of other citizens. This is something that cannot be allowed in our libraries. The final core issue is the essence of censorship versus the essence of a library. Censorship is meant to create a politically correct and non-offensive environment through the limitation of exposure to materials. Libraries, however, are meant to allow public access to works that stretch and challenge knowledge, beliefs, notions and all ideas held by people through the works of others. Censorship cannot be implemented without a clash with the essence of a library’s purpose. If censorship were to be enacted, a library could not provide new information if someone else did not like what was being taught. A library could not provide works that challenge and stretch individual beliefs if someone was offended by the means by which the stretching occurred. Pre-conceived notions could not be defeated with the presence of factual knowledge if someone did not like the truth. In all of these ways, and many more, a library’s core ideals and purpose could not be upheld with the induction of a system of censorship. All in all, we can see that censorship could only hope to destroy the system libraries abide by. The constitutional rights of citizens would be infringed upon, the fickle nature of humans and the lack of definition for ‘offensive’ would allow people to undermine the rights of others, and the essence of what a library really is would be ravaged. We cannot, as @CAPS1 with rights, employ a system of censorship.

Table 15: Essay e_3 in the creative cluster.

Category	**Evaluation criterion**	**Score**					
		e_1		e_2		e_3	
		Human	GPT-4o	Human	GPT-4o	Human	GPT-4o
Content	Novelty of ideas and content	2	2	3	2	4	4
	Richness of content	1	2	3	2	5	5
	Logicality of content	1	1	4	2	5	4
Organization	Originality of structure	1	1	3	2	5	3
	Cohesiveness of structure	3	1	2	2	5	5
Expression	Originality of expression	2	1	3	2	4	4
	Appropriateness of expression	2	1	3	1	5	5
Author’s voice	Perspective and personality	2	2	3	2	5	5
Reader’s response	Fun and persuasiveness	3	1	3	2	5	4
	Creativity score	1.89	1.33	3	1.89	4.78	4.33

Table 16: Creativity scores of essays e_1, e_2, and e_3 in Tables 13-15.