Towards Robust Cross-Prompt Essay Trait Scoring: A Generative Model Framework with Ranking Loss

Anonymous ACL submission

Abstract

Automated Essay Scoring (AES) aims to evaluate the overall quality of essays, while essay trait scoring provides a detailed assessment by assigning separate scores to specific traits. Prompt-specific AES models have shown success, but their application to "unseen" prompts remains challenging due to limited prompt and essay diversity, hindering the generalization ability. This paper introduces GenAES, a generative model framework for cross-prompt essay trait scoring, leveraging large language models (LLMs) to augment prompts and essays. GenAES further develops a prompt encoder to manage representations of unseen prompts and introduces a ranking loss to evaluate the similarity of unlabeled generated essays with the source essays. Experimental results demonstrate that GenAES significantly improves generalization, achieving state-of-the-art performance on the ASAP++ dataset. It shows improvements of 6.5% and 7.3% in average QWK scores across prompts and traits, respectively. The generated prompts and essays are released to facilitate future research.

1 Introduction

Automated Essay Scoring (AES) is a complex task that involves predicting a holistic score for a given student essay. This task requires evaluating various aspects of the essay, including its coherence, structure, quality, and relevance to the given prompt. Early studies (Taghipour and Ng, 2016; Dong and Zhang, 2016; Alikaniotis et al., 2016; Yang et al., 2020) have shown the effectiveness of supervised learning for prompt-specific tasks, but these models require substantial same-prompt training data, limiting their applicability to new prompts. Thus, recent work (Jin et al., 2018; Mayfield and Black, 2020; Ridley et al., 2020) has shifted focus towards developing cross-prompt AES models, which are trained and tested on essays from different prompts to improve generalization.

To provide comprehensive feedback on overall quality as well as specific elements of essays, recent studies (Mathias and Bhattacharyya, 2018, 2020; Ridley et al., 2021; Chen and Li, 2023; Do et al., 2023) have aimed at scoring essays on different traits, such as content, organization, style, and conventions. ProTACT (Do et al., 2023) represents the current state-of-the-art cross-prompt essay trait scoring systems, which enhance joint learning of traits by recognizing prompt and trait similarities through prompt-specific encoding and attention mechanisms. In this paper, we focus on the cross-prompt essay trait scoring setting.

We identify two key challenges in this task. First, the effectiveness of cross-prompt scoring systems is limited when rating essays for unseen prompts during inference, as models struggle to apply appropriate rubrics. Second, sparse labeled essays for a given prompt make learning its continuous score distribution challenging.

To address these challenges, we propose GenAES: a generative model framework for cross-prompt essay trait scoring. GenAES evaluates both holistic and trait scores of essays, focusing on robust performance for unseen prompts. To address the first challenge, we develop an attribute prompt generator using large language models (LLMs) to produce diverse prompts aligned with rubrics, enriching the training dataset. Additionally, we introduce a prompt encoder that utilizes contrastive learning to learn prompt representation, enabling the projection of unseen prompt representations into the prompt category representation space during inference. For the second challenge, an essay generator synthesizes high-quality essays to learn from a densely populated essay distribution, enhancing evaluation granularity. Furthermore, we introduce a ranking loss mechanism to ensure consistent relative relations between essays when measuring the similarity of unlabeled generated essays to labeled essays. The results indicate that our

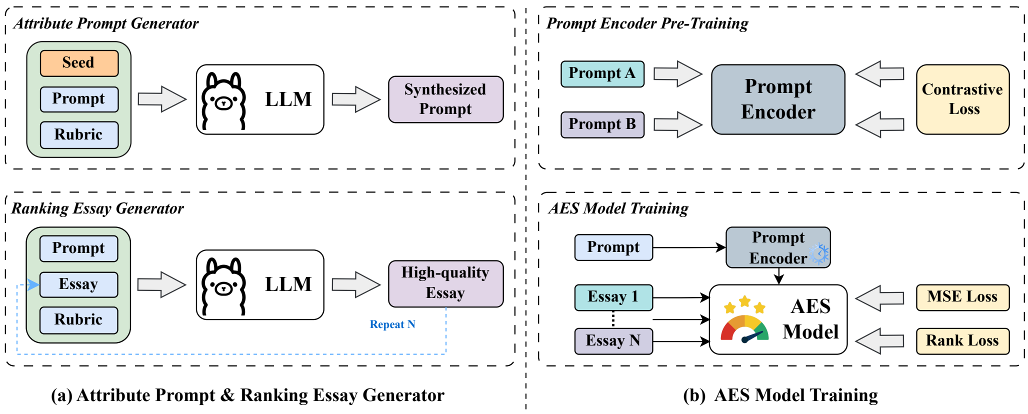

Figure 1: The workflow of GenAES.

proposed method, GenAES, improves generalization for data-scarce and unseen prompts, increasing average QWK scores by 6.5% over traits and 7.3% over prompts, achieving state-of-the-art performance on the ASAP++ dataset.

2 Method

Figure 1 presents an overview of our method, including the data generators for synthesized prompts and essays, as well as the training phases for the prompt encoder and the overall AES model.

2.1 Prompt and Essay Generators

Recently, LLMs have been successfully demonstrated as effective training data generators (Meng et al., 2022). Leveraging this capability, our method employs LLMs to create variations of given essay prompts under relevant topics and high-quality essays of given source essays.

To guide LLMs in generating task-specific content, we follow the methodology of AttrPrompt (Yu et al., 2023) to create attribute-specific prompts. Initially, thematic words are manually extracted from the original prompts as "attributes." Using LLMs, we generate a list of topic-relevant words based on these attributes as seeds (examples are listed in Appendix A.1). Next, an essay prompt, its corresponding scoring rubrics, and these seed words are inputted into the LLMs. The LLMs then produce variations of the original prompt by substituting thematic words with conceptually similar terms that align with the provided rubrics. This method ensures that the new prompts retain the original essence while covering diverse yet related topics. The prompt used for attribute prompts is shown in Appendix A.2.

To enhance the model's sensitivity to subtle score differences, we present a progressive method to improve essay quality by generating additional high-quality essays. Initially, the essay prompt, source essay, and corresponding scoring rubrics guide the LLM in refining the source essay. The improved essay, along with the original prompt and rubrics, is iteratively reintroduced to the LLM to generate further refined essays. This iterative process allows for increasing the number of iterations as necessary to expand the dataset, thereby refining the quality gradations between essays. The prompt used for generating essays is shown in Appendix A.3.

2.2 AES model

Our AES model extends upon the ProTACT (more details are in the Appendix A.4), integrates the proposed prompt encoder (Section 2.2.1), and is trained using our generated prompts and essays with a combination of regression and ranking loss (Section 2.2.2).

2.2.1 Prompt Encoder

To process unseen prompts during inference, our proposed prompt encoder learns prompt representations based on prompt categories. For example, argumentative essays are grouped closely together in the representation space, while essays from different categories are farther apart. We draw inspiration from Gao et al. (2022) to pre-train the generated prompts, and employ contrastive learning to project prompts into similar prompt categories and their associated rubrics. The contrastive loss is:

Models	P1	P2	P3	P4	P5	P6	P7	P8	Avg.
PAES (Ridley et al., 2020)	0.605	0.522	0.575	0.606	0.634	0.545	0.356	0.447	0.536
CTS (Ridley et al., 2021)	0.629	0.543	0.596	0.620	0.614	0.546	0.382	0.501	0.554
ProTACT (Do et al., 2023)	0.647	**0.587**	**0.623**	0.632	0.674	0.584	**0.446**	0.541	0.592
- w/o topic coherence features	0.638	0.559	0.595	0.624	0.615	0.567	0.397	0.531	0.566
ProTACT (our implementation)	0.648	0.570	0.623	0.613	0.669	0.573	0.466	0.450	0.576
GenAES (ours)	0.666	0.585	0.616	**0.656**	0.669	0.600	0.412	**0.620**	**0.603**
- w/o Essay ranking loss	**0.668**	0.577	0.612	0.623	**0.680**	0.578	0.420	0.610	0.596
- w/o Prompt contrastive loss	0.649	0.586	0.609	0.623	0.668	**0.606**	0.400	0.617	0.595

Table 1: Performance comparison across 8 prompts of ASAP dataset in the cross-prompt setting.

$$\ell_i = -\log \frac{e^{\text{sim}(h_i, h_i^+)/\tau}}{\sum_{j=1}^{N} e^{\text{sim}(h_i, h_j^+)/\tau} + e^{\text{sim}(h_i, h_j^-)/\tau}} \tag{1}$$

where h_i, h_i^+, and h_j^- are the embeddings of a prompt, a positive prompt within the same category, and a negative prompt within a different category, respectively. τ is a scaling hyperparameter, and e denotes the exponential function.

2.2.2 Ranking Loss

GenAES integrates MSE loss for the original essays and the proposed ranking loss for generated essays (Section 2.1). To learn the dense score distribution from the generated essays, the proposed pair-wise ranking loss (hinge loss) is:

$$\mathcal{L}_{\text{rank}} = \max(0, 1 - (\text{logits}_a - \text{logits}_b)) \tag{2}$$

where logits_a and logits_b denote the predicted scores for an augmented essay (essay a) and a training sample (essay b), respectively. The loss function penalizes cases where logits_a is not significantly higher than logits_b, given that essay a is intended to be of higher quality than essay b.

Three advantages of the proposed ranking loss are: First, this approach enforces a margin between high- and low-quality essay scores, enhancing generalization and prediction. Second, It mitigates noise from generated essays, resulting in more robust training. Third, it addresses the lack of ground truth for generated essays by using relative scores instead of absolute values.

3 Experiment

3.1 Experimental Setup

Datasets. In this study, we utilized ASAP (Kaggle, 2012) and ASAP++ (Mathias and Bhattacharyya, 2018) for evaluations. The Automated Student Assessment Prize (ASAP) competition dataset (Kaggle, 2012) consists of 13,000 essays categorized into argumentative, response, and narrative types across 8 prompts, widely used for evaluating AES systems. ASAP++ is an extension of the ASAP dataset that includes additional trait scores for each prompt. We followed ProTACT to partition the dataset and report results on test sets.

Metrics. We evaluate our model using Quadratic Weighted Kappa (QWK) to assess agreement between ground truth and predictions. In accordance with the cross-prompt setting, we employ a leave-one-out strategy, training on seven sets and testing on the remaining set.

Baselines. We benchmark our approach against established cross-prompt essay trait scoring systems: PAES (Ridley et al., 2020), CTS (Ridley et al., 2021) and ProTACT (Do et al., 2023). PAES (Ridley et al., 2020) uses a hierarchical CNN-LSTM with POS embeddings and linguistic features, while CTS (Ridley et al., 2021) employs trait-specific attention mechanisms. Additionally, we evaluate ProTACT without topic coherence features to compare with our implementation.

Implementation details. We follow and replicate the ProTACT implementation as our backbone. It is important to note that ProTACT does not include detailed procedures for the topic coherence (TC) features; hence, our implementation is built upon the same architecture but without the TC features. For LLMs, we use Command R+ (104B), an open-source model comparable to OpenAI's ChatGPT-4. Our LLM backend is powered by Ollama, utilizing the 4-bit quantized version of Command R+. For the prompt encoder, we perform contrastive learning on a pre-trained BERT model, replacing ProTACT's original prompt encoder. To preserve pre-trained knowledge, we freeze the prompt encoder during training. The statistical description and examples of the generated prompts and essays are provided in Appendix A.5.

Models	Traits									Avg.
	Overall	Content	Org	WC	SF	Conv	PA	Lang	Nar	
CTS	0.670	0.551	0.459	0.562	0.556	0.413	0.568	0.533	0.610	0.547
ProTACT	0.674	0.596	0.518	0.599	0.585	0.450	**0.619**	0.596	0.639	0.586
- w/o TC feature	0.671	0.565	0.477	0.582	0.574	0.435	0.573	0.550	0.618	0.561
ProTACT (our implementation)	0.633	0.581	0.510	0.583	0.589	0.474	0.578	0.560	0.623	0.570
GenAES (ours)	0.676	**0.612**	**0.545**	0.610	**0.614**	0.497	0.618	**0.600**	**0.644**	**0.602**
- w/o Essay ranking loss	**0.682**	0.601	0.520	**0.612**	0.602	0.481	0.616	0.580	0.636	0.592
- w/o Prompt contrastive loss	0.659	0.609	0.540	0.607	0.600	**0.500**	0.612	0.583	**0.644**	0.595

Table 2: The average QWK scores over all prompts for each **trait** (WC: Word Choice; PA: Prompt Adherence; Nar: Narrativity; Org: Organization; SF: Sentence Fluency; Conv: Conventions; Lang: Language).

3.2 Results

Table 1 presents the QWK performance of all compared methods tested across 8 prompts from the ASAP dataset in a cross-prompt setting. On average, GenAES achieves the highest performance, particularly excelling in inference for prompts 4 and 8, while ProTACT shows relatively lower agreement but performs well for prompts 2, 3, and 7. Moreover, the essay distribution for prompt 8 is markedly different from the other prompts. Specifically, essays for prompt 8 have an average length of 620 words and a score range of 0-60, compared to average lengths ranging from 150 to 350 words and score ranges from 0 to 30 for the other prompts. Observing the significant results on prompt 8, this indicates that GenAES possesses superior generalization ability for essay scoring.

It should be noted that while our implementation of ProTACT lacks topic coherence features, GenAES still demonstrates a promising improvement from 0.576 to 0.603 in QWK score. This suggests that despite the absence of certain features, our approach shows significant enhancement in performance. Furthermore, when GenAES operates without either the ranking loss for generated essays or the contrastive loss for prompt construction, its performance decreases. This indicates the critical role of these components in enhancing essay scoring accuracy.

Table 2 presents the average QWK scores for each trait across all prompts from the ASAP++ dataset. GenAES outperforms the compared systems on most traits, demonstrating its robustness and effectiveness in capturing various aspects of essay quality. However, GenAES shows slightly lower performance in the Conventions (Conv) and Prompt Adherence (PA) traits. This suggests that while GenAES is highly effective in assessing the overall quality and structure of essays due to the design of attribute prompts and high-quality essays, there is room for improvement in ensuring strict adherence to writing conventions and the specific requirements of the prompt, particularly for long-length prompts in the response category.

Visualization for prompt representation We use t-SNE (van der Maaten and Hinton, 2008) to visualize the prompt encoder representations after contrastive learning, shown in Figure 2. Notably, Prompts 3, 4, 5, and 6, which belong to the response category, and Prompts 1 and 2, which belong to the argumentative essay category, are respectively clustered closely together in the representation space. This clustering indicates that the contrastive learning-trained prompt encoder effectively captures inherent similarities between prompts of the same type, which is crucial for developing robust cross-prompt essay scoring systems.

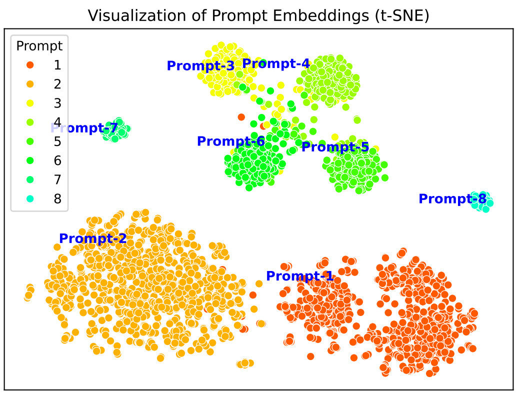

Figure 2: Visualization of Prompt Encoder Clustering After Contrastive Training.

4 Conclusion

In this work, we present GenAES, a generative framework for cross-prompt essay trait scoring aimed at enhancing generalization to unseen prompts. Our experimental results demonstrate that the mechanisms implemented in GenAES significantly improve the ability to evaluate essays and traits, achieving state-of-the-art performance on the ASAP++ dataset.

5 Limitations

While our proposed methods introduce innovative techniques for generating new prompts and training a prompt encoder to address this issue, they are constrained by the narrow scope and biased distribution of the ASAP dataset, resulting in limited and potentially skewed training samples. Despite promising results, our approach introduces potential noise and variability due to the quality of generated data, necessitating meticulous tuning and validation to ensure consistent performance. Additionally, we observed an intriguing phenomenon: due to the selective training data used by large language models (LLMs), generating new low-quality essays from a given essay is challenging. This phenomenon warrants further exploration in future research.

References

Dimitrios Alikaniotis, Helen Yannakoudakis, and Marek Rei. 2016. Automatic text scoring using neural networks. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

Yuan Chen and Xia Li. 2023. PMAES: Prompt-mapping contrastive learning for cross-prompt automated essay scoring. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1503, Toronto, Canada. Association for Computational Linguistics.

Heejin Do, Yunsu Kim, and Gary Geunbae Lee. 2023. Prompt- and trait relation-aware cross-prompt essay trait scoring. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1538–1551, Toronto, Canada. Association for Computational Linguistics.

Fei Dong and Yue Zhang. 2016. Automatic features for essay scoring – an empirical study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1072–1077, Austin, Texas. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2022. Simcse: Simple contrastive learning of sentence embeddings. *Preprint*, arXiv:2104.08821.

Cancan Jin, Ben He, Kai Hui, and Le Sun. 2018. TDNN: A two-stage deep neural network for prompt-independent automated essay scoring. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1088–1097, Melbourne, Australia. Association for Computational Linguistics.

Kaggle. 2012. The hewlett foundation: Automated essay scoring.

Sandeep Mathias and Pushpak Bhattacharyya. 2018. ASAP++: Enriching the ASAP automated essay grading dataset with essay attribute scores. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Sandeep Mathias and Pushpak Bhattacharyya. 2020. Can neural networks automatically score essay traits? In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 85–91, Seattle, WA, USA → Online. Association for Computational Linguistics.

Elijah Mayfield and Alan W Black. 2020. Should you fine-tune BERT for automated essay scoring? In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 151–162, Seattle, WA, USA → Online. Association for Computational Linguistics.

Yu Meng, Jiaxin Huang, Yu Zhang, and Jiawei Han. 2022. Generating training data with language models: Towards zero-shot language understanding. *Advances in Neural Information Processing Systems*, 35:462–477.

Robert Ridley, Liang He, Xin-yu Dai, Shujian Huang, and Jiajun Chen. 2021. Automated cross-prompt scoring of essay traits. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13745–13753.

Robert Ridley, Liang He, Xinyu Dai, Shujian Huang, and Jiajun Chen. 2020. Prompt agnostic essay scorer: A domain generalization approach to cross-prompt automated essay scoring. *Preprint*, arXiv:2008.01441.

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, Austin, Texas. Association for Computational Linguistics.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605.

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 *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1560–1569, Online. Association for Computational Linguistics.

Yue Yu, Yuchen Zhuang, Jieyu Zhang, Yu Meng, Alexander Ratner, Ranjay Krishna, Jiaming Shen, and Chao Zhang. 2023. Large language model as attributed training data generator: A tale of diversity and bias. *Preprint*, arXiv:2306.15895.

A Appendix

A.1 Sample seeds

Table 4 presents a subset of the seeds used in our study. We adopted the methodology from (Yu et al., 2023) to identify key attributes for each prompt set. Attribute seeds were then generated using OpenAI GPT-4, ensuring that the prompts aligned with the original criteria and enhanced model performance through attributed prompt engineering.

A.2 Prompt for generating attribute prompt

The prompts used to generate new prompts are shown in Figure 3 respectively. We require the LLM to generate new prompts based on the original prompt, rubric guidelines, and an additional seed list.

A.3 Prompt for generating high-quality essays

The prompts used to generate essay are shown in Figure 4. The yellow box contains the prompt rubric guidelines provided by the ASAP dataset. These guidelines are primarily intended to guide evaluators in scoring. We used these rubric guidelines, along with rules on style, length, and other factors, to help the LLM understand how to revise the essays to achieve higher scores.

A.4 ProTACT

The ProTACT model employs a hierarchical architecture using part-of-speech embeddings, convolutional layers, and attention pooling to craft nuanced sentence and document-level representations. It incorporates pre-trained GloVe embeddings and multi-head self-attention to enhance prompt understanding and align essay content with prompts effectively. The model integrates these advanced textual features with non-prompt-specific traits and unique topic coherence features, supported by a trait attention mechanism that improves assessment precision by focusing on trait interrelationships. Scoring is performed via a sigmoid-functioned linear layer for precise and reliable trait evaluation.

A.5 Generated Data

Table 3 presents the statistical description of our generated prompts and essays. Please note that the generated data was length-restricted to ensure consistency with the ASAP dataset. Table 5 and Table 6 respectively present examples for our generated essays and prompts.

Set	Prompts	Avg Len.	Essays	Avg Len.
1	750	62	3566	352
2	960	53	3600	351
3	200	66	3452	177
4	192	71	3540	155
5	184	65	3610	184
6	161	69	3600	173
7	150	30	3138	234
8	140	41	1446	443

Table 3: The statistical description of our generated prompts and essays.

The provided prompt primarily encourages students to write a letter to their local newspaper, expressing their views on the impact of computers on humanity. Please generate 10 new prompts based on the following requirements:
1. Replace the computer theme mentioned in the provided prompt with {seed}.
2. Ensure the narrative style, vocabulary, and structure of the new prompt are distinct from those in the provided prompt.
3. Develop the prompt according to the provided rubric guidelines.
4. The essays written based on the prompt typically have an average length of {num_avg_word} words.
5. Please ensure that the every generated prompt is between {min_word} and {max_word} words long.
6. Please ensure that the output format adheres to the instructions specified in the output template.

Prompt:
{prompt}

Rubric guidelines:
{rubric}

Output template:
1. <generated prompt>
2. <generated prompt>
...
9. <generated prompt>
10. <generated prompt>

Figure 3: Prompt used for generating new prompts.

Type of essay:
Persuasive/Narrative/Expository
Grade level: 8
Average length of essays: 350 words

Rubric Guidelines:
Score Point 1: An undeveloped response that may take a position but offers no more than very minimal support. Typical elements:
· Contains few or vague details.
· Is awkward and fragmented.
...

You are an English student proficient in English. Please follow the provided prompt, rubric guidelines, and essay, to revise the essay for a slightly higher score. Please only revise the essay and do not include any unrelated content.

Prompt:
{prompt}

Rubric guidelines:
{rubric}

Essay:
{essay}

Figure 4: Prompt used for generate high-quality essays.

Set	Attribute	Example
1	impact	1. SmartphonesSocial media platforms (Facebook, Instagram, Twitter, etc.) 2. Cloud serving, Internet of Things (IoT) 3. Artificial intelligence (AI) 4. Autonomous vehicles (self-driving cars) 5. Virtual reality (VR) 6. Biometric security (fingerprint, iris scans) 7. Drone delivery services 8. Teletherapy services 9. Video blogs (vlogs) 10. Online language learning platforms (Duolingo) 11. Digital wallets
2	censorship	1. National healthcare systems 2. Public education systems 3. National taxation systems 4. Social security systems 5. Immigration policies 6. Climate change policies 7. Freedom of information laws 8. Internet neutrality laws 9. E-voting systems 10. Public referendums and initiatives 11. Jury system in the legal process 12. Parole and probation services
3 ∼6	books, novels	1. "The Adventures of Huckleberry Finn" by Mark Twain 2. "To Kill a Mockingbird" by Harper Lee 3. "Animal Farm" by George Orwell 4. "Frankenstein" by Mary Shelley 5. "The Picture of Dorian Gray" by Oscar Wilde6. "The Time Machine" by H.G. Wells 7. "The War of the Worlds" by H.G. Wells 8. "The Call of the Wild" by Jack London 9. "White Fang" by Jack London 10. "Robinson Crusoe" by Daniel Defoe 11. "Little Women" by Louisa May Alcott 12. "Animal Farm: A Fairy Story" by George Orwell
7	personal experience	1. A Memorable Summer Camp Adventure 2. The Day I Overcame My Greatest Fear 3. An Unforgettable Family Road Trip 4. A special birthday surprise 5. A Life-Changing Volunteer Experience
8	creative idea	1. Discovering a secret entrance to a magical world at school 2. Waking up one day with the ability to talk to animals 3. Participating in a mysterious science experiment and gaining superpowers 4. Finding out that the school playground transforms into a fantasy adventure park after school 5. Realizing that the school bus is actually an alien spaceship

Table 4: Sample seeds for generating attributed prompts.

Iterative step	Essay
Original essay	In the story “ Winter Hibiscus,” the author concludes the story with this paragraph for many reasons. This paragraph shows that Saeng will not give up, she will wait for next time to come around to take the test. It also means she will be well prepared and will have a list of confidence in passing this test. It gives the reader an idea that Saeng will not give up and she will complete the test. Her mother will be proud of her for it and it shows Saeng will never give up. Just like when she said “let’s plant it, right now,” This example shows that even though Saeng had failed the test and disappointed her mother, she will never give up and she still has hope in the plant just like how her mother has hope for Saeng.
High-quality generated essay 1	In the story “Winter Hibiscus,” the author concludes with a powerful paragraph that underscores Saeng’s determination and resilience. Despite her initial failure, Saeng is committed to retaking the test in the spring when the hibiscus blooms, symbolizing a fresh start and renewed hope. Her eagerness to plant the hibiscus, even in winter, mirrors her resolve to succeed despite setbacks. This final paragraph emphasizes her unwavering spirit and the support of her mother, who believes in her. The hibiscus serves as a metaphor for Saeng’s journey, suggesting that, like the plant, she will flourish after overcoming challenges.
High-quality generated essay 2	In ‘Winter Hibiscus,’ the author concludes by emphasizing the main character’s resilience in the face of adversity through a powerful paragraph focusing on Saeng’s tenacity. Despite setbacks, Saeng displays an unwavering resolve to persevere and confidently prepare for future challenges—a testament to her indomitable spirit. This conclusion provides a deeper understanding of Saeng’s character beyond mere words; it showcases her determination, reflecting the same nurturing hope she displayed when planting the hibiscus during winter despite its seemingly slim chances of survival. Just as Saeng fosters belief in the plant’s future bloom, so too does her mother unwaveringly believe in Saeng’s ability to thrive and succeed against all odds. Through this final paragraph, the narrative offers insight into a profound message of resilience that extends beyond the story’s ending—a lasting impression of Saeng’s steadfast determination that resonates long after the final page is turned.

Table 5: An example for the generated essay of Prompt 4.

Prompt 1	Data
Original Prompt	More and more people use computers, but not everyone agrees that this benefits society. Those who support advances in technology believe that computers have a positive effect on people. They teach hand-eye coordination, give people the ability to learn about faraway places and people, and even allow people to talk online with other people. Others have different ideas. Some experts are concerned that people are spending too much time on their computers and less time exercising, enjoying nature, and interacting with family and friends. Write a letter to your local newspaper in which you state your opinion on the effects computers have on people. Persuade the readers to agree with you.
Generated Prompt	With public spaces increasingly offering free Wi-Fi connectivity, debates emerge regarding its impact on our daily lives and relationships. Do these technological advancements enhance community bonds by providing new avenues for connection and collaboration? Or do they distract us from cultivating deeper, more meaningful interactions offline? Write a letter to your local newspaper, sharing your perspective on the social implications of widespread Wi-Fi adoption.

Table 6: An example for generated prompts of Prompt 1.