Unleashing Large Language Models' Proficiency in Zero-shot Essay Scoring

Sanwoo Lee[1,2] **Yida Cai**[1,3] **Desong Meng**[1,2] **Ziyang Wang**[1,3] **Yunfang Wu**[1,2]*
[1]National Key Laboratory for Multimedia Information Processing, Peking University
[2]School of Computer Science, Peking University
[3]School of Software and Microelectronics, Peking University
{sanwoo, wuyf}@pku.edu.cn, {caiyida, 2100013162, wzy232303}@stu.pku.edu.cn

Abstract

Advances in automated essay scoring (AES) have traditionally relied on labeled essays, requiring tremendous cost and expertise for their acquisition. Recently, large language models (LLMs) have achieved great success in various tasks, but their potential is less explored in AES. In this paper, we show that our zero-shot prompting framework, Multi Trait Specialization (MTS), elicits LLMs' ample potential for essay scoring. In particular, we automatically decompose writing proficiency into distinct traits and generate scoring criteria for each trait. Then, an LLM is prompted to extract trait scores from several conversational rounds, each round scoring one of the traits based on the scoring criteria. Finally, we derive the overall score via trait averaging and min-max scaling. Experimental results on two benchmark datasets demonstrate that MTS consistently outperforms straightforward prompting (Vanilla) in average QWK across all LLMs and datasets, with maximum gains of 0.437 on TOEFL11 and 0.355 on ASAP. Additionally, with the help of MTS, the small-sized Llama2-13b-chat substantially outperforms ChatGPT, facilitating an effective deployment in real applications.

1 Introduction

Automated essay scoring (AES) aims at evaluating and scoring essays with machine learning (Dikli, 2006). AES is a promising alternative to costly and laborious human assessment, greatly resolving rater fatigue and inter-rater inconsistency. AES systems have been widely deployed in classroom settings (Dikli and Bleyle, 2014) and high-stakes tests such as TOEFL (Attali and Burstein, 2006).

Previous studies highly matched human ratings via developing supervised models tailored to a specific prompt (Yannakoudakis et al., 2011; Taghipour and Ng, 2016; Dong et al., 2017), assuming essays from the train and test sets belong to the same prompt. However, prompt-specific models struggled when confronted with essays written for unseen prompts (Jin et al., 2018; Cozma et al., 2018). Hence AES is progressing towards reflecting more real-world scenarios, exemplified by cross-prompt approaches (Cao et al., 2020; Jiang et al., 2023b) which reinforce domain transferability of the supervised models.

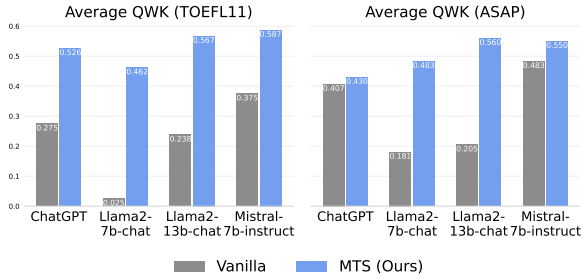


Figure 1: Comparison of our MTS zero-shot prompting framework and Vanilla baseline across different types of LLMs and datasets, measured on average QWK.

Recently, advances in large language models (LLMs) (Brown et al., 2020; Ouyang et al., 2022; OpenAI, 2023; Touvron et al., 2023) have led to a paradigm shift in which LLMs excel across a wide range of downstream tasks via zero-shot or few-shot instructions (Yuan et al., 2023; Zhang et al., 2023a). In many cases, careful prompt design plays a crucial role in unlocking LLMs' potential. For instance, chain-of-thought (CoT) prompting (Wei et al., 2022; Kojima et al., 2022) improves LLMs' performance on complex reasoning benchmarks by externalizing the reasoning process.

The development of LLM-based chatbots aligned with human preferences (Ouyang et al., 2022; Rafailov et al., 2023) has given rise to zero-shot AES, allowing us to move beyond the cross-prompt setting. However, leveraging LLMs for zero-shot AES is less explored, in contrast to pro-

* Corresponding author.

Findings of the Association for Computational Linguistics: EACL 2024, pages 181–198
November 12-16, 2024

liferating studies harnessing LLMs to serve as an evaluation metric for machine-generated text (Chiang and Lee, 2023; Zheng et al., 2023; Liu et al., 2023). Initial works of zero-shot AES (Mizumoto and Eguchi, 2023; Yancey et al., 2023) prompt LLMs to assign the overall score within a single step, which demonstrates suboptimal agreement with human raters.

In this paper, we present **MTS** (**M**ulti **T**rait **S**pecialization), a zero-shot prompting framework to elicit essay scoring capabilities in LLMs, inspired by supervised models that explicitly predict trait scores and improve the overall scoring (Ridley et al., 2021; Kumar et al., 2022; Do et al., 2023). In particular, we exploit ChatGPT (OpenAI, 2022) to decompose the writing quality into multiple traits and generate scoring criteria for each trait. Next, an LLM engages in several rounds of conversation, each round evaluating with respect to one of the traits. During the conversation, the LLM is instructed to retrieve quotes to provide faithful evaluation on the essay, then assign a score based on the given scoring criteria. Finally, the overall score is derived by averaging and min-max scaling the trait scores, in combination with the outlier clipping mechanism.

We evaluate MTS on ASAP (Hamner et al., 2012) and TOEFL11 (Blanchard et al., 2013) with different LLMs, including ChatGPT, Llama 2 (Touvron et al., 2023) and Mistral 7b (Jiang et al., 2023a). We take the Vanilla approach as a primary baseline which asks LLMs to produce rationales followed by an overall score. As illustrated in Figure 1, MTS consistently outperforms Vanilla in average Quadratic Weighted Kappa (QWK) across all combinations of LLMs and datasets, with maximum gains of 0.437 ($0.025 \rightarrow 0.462$) on TOEFL11 and 0.355 ($0.205 \rightarrow 0.560$) on ASAP. In addition, the small-sized Llama2-13b-chat substantially surpasses ChatGPT with the help of MTS, enabling a more effective deployment.

In essence, our contributions are as follows:

- Our framework is free from training as well as labeling essays, which can be readily applied to new essay prompts (domains).
- We automatically decompose the AES task into more specific subtasks with respect to diverse traits, thereby significantly boosting the agreement with human raters.
- We utilize min-max scaling with outlier clipping, effectively addressing LLMs' scoring bias and contributing to robust performance.
- MTS achieves promising results, largely exceeding Vanilla, also outperforming ChatGPT with the small-sized Llama2-13b-chat.

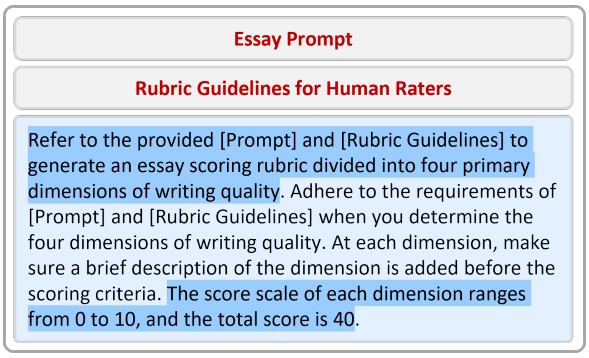


Figure 2: Illustration of the prompt for multi trait decomposition used for ASAP. The contents to be filled are denoted in red. See Appendix A for the templates used for ASAP and TOEFL11.

2 Method

We formalize the definition of zero-shot AES as follows: Given a dataset consisting of unlabelled essays $\mathcal{D} = \{x^{(i)}\}_{i=1}^{N_\mathcal{D}}$, the goal is to output an overall score $\hat{y}^{(i)}$ for every essay $x^{(i)}$ from a set of predefined scores $\mathcal{Y}$ where $\hat{y}^{(i)} \in \mathcal{Y}$.

Multi trait specialization encourages LLMs to assess the essay from diverse aspects of writing quality. It consists of three steps: (1) decomposing writing proficiency into multiple traits and generating scoring criteria; (2) assigning a trait-specific score by step-by-step evaluation specialized to the trait; (3) deriving the final score via trait aggregation and scaling. The overall architecture of MTS is illustrated in Figure 3.

2.1 Multi Trait Decomposition

A straightforward way of zero-shot AES with LLM would be asking it to score an essay in a single response. Despite its simplicity, this approach cannot guarantee that the LLM employs the same scoring criteria across essays, leading to inconsistent evaluation. Moreover, scoring an essay should ideally be based on a comprehensive analysis of various dimensions of writing quality, while the LLM may be overloaded to do so in a single response.

We address this issue by decomposing the writing proficiency into several key traits and defining trait-specific scoring criteria which will be fixed

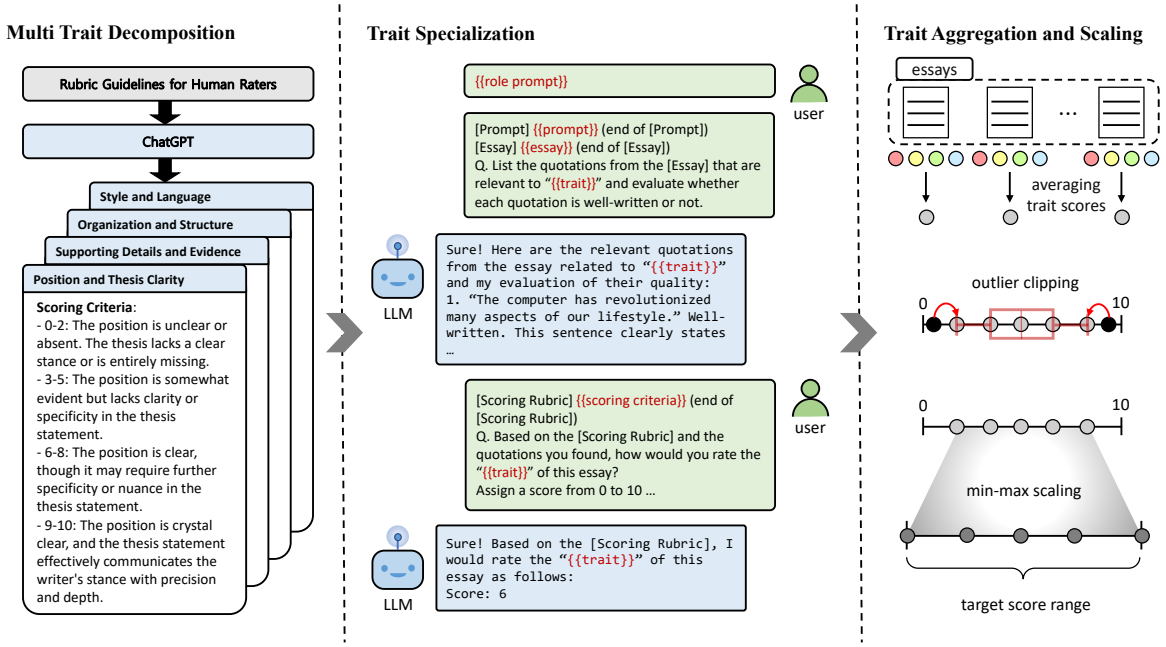


Figure 3: The illustration of Multi Trait Specialization framework. The parts to be filled with specific contents are substituted with comments between double curly braces and colored in red.

during the assessment (details in Section 2.2), contributing to a consistent scoring behavior of LLM. We automate this process via instructing ChatGPT to condense the rubric guidelines used by human raters into several key traits of writing quality and generate scoring criteria for each trait, using the prompt outlined in Figure 2. This procedure ensures isolating trait-specific scoring criteria from rubric guidelines that mix multiple traits as a whole. An example of the generation result is depicted in the left part of Figure 3 (see Appendix B for more results).

2.2 Trait Specialization

Prompt design plays a crucial role in unlocking the emergent abilities of LLMs. One of the key findings is that their reasoning ability benefits from sub-problem decomposition of the complex promblem (Zhou et al., 2022). We hypothesize that LLMs provide more reasonable assessment of essays when specialized to perform step-by-step evaluation restricted to one aspect of writing quality, inspiring us to design the following steps of prompting (See Appendix C.1 for the full template):

1. **Trait-specific Conversation**: For an essay, a number of independent conversations are initiated with each conversation specialized to one of the traits. Within the conversation, LLM is given a role prompt (i.e., system prompt) to focus solely on evaluating the specific trait.

2. **Quote Retrieval**: Each conversation consists of two turns. In the first turn, LLM is required to retrieve quotes relevant to the trait and provide verbal evaluations for each quote.

3. **Scoring**: In the second turn, LLM is asked to score the essay with respect to the trait, referring to both the previous turn and the given scoring criteria.

In Step 1, leveraging trait-specific conversations not only simplifies the AES task as a form of sub-problem decomposition but also prevents evaluations on each trait from being influenced by the other traits. In Step 2 and 3, the quote retrieval task is followed by the main scoring task, as depicted in the middle part of Figure 3. The quote retrieval task allows LLM to adhere to the details of the essay and avoid producing generic evaluation, whereas the scoring task transforms the verbal evaluation into the score based on the predefined scoring criteria.

2.3 Trait Aggregation and Scaling

Parsing the output of Section 2.2 yields trait scores $\{\hat{y}_j^{(i)}\}_{j=1}^{N_T}$ for an essay $x^{(i)}$, where N_T represents the number of predefined traits. The trait scores

should be transformed to an overall score $\hat{y}^{(i)}$ that falls under the target score range which may vary across different prompts (as in Table 1).

Based on the trait scores, we devise a simple yet effective trait aggregation and scaling strategy. First, the trait scores are aggregated by taking their average, and the outliers among the averaged scores are clipped using Q1 and Q3, i.e. the first and the third quartiles of the averaged scores, that is, the clipped score $\hat{y}^{(i)}_{agg}$ is computed as follows:

$$\hat{y}^{(i)}_{agg} = \min(\max(\frac{1}{N_T}\sum_{j=1}^{N_T}\hat{y}^{(i)}_j, v_{min}), v_{max}) \quad (1)$$

where $v_{min} = \text{Q1} - 1.5(\text{Q3} - \text{Q1})$ and $v_{max} = \text{Q3} + 1.5(\text{Q3} - \text{Q1})$. The value 1.5 here is commonly used for outlier detection (Seo, 2006). Next, the clipped scores are mapped to the target range $[a, b]$ via min-max scaling:

$$\hat{y}^{(i)} = a + \frac{(\hat{y}^{(i)}_{agg} - \hat{y}_{min})(b-a)}{\hat{y}_{max} - \hat{y}_{min}} \quad (2)$$

where $\hat{y}_{min} = \min_i \hat{y}^{(i)}_{agg}$ and $\hat{y}_{max} = \max_i \hat{y}^{(i)}_{agg}$. In this way, the clipping alleviates the sensitivity of min-max scaling to the outliers.

3 Experimental Settings

3.1 LLMs

To verify the effectiveness of our proposed method is universal across LLMs, we choose different types of LLMs which are not variants of one another: **ChatGPT**, **Llama 2** and **Mistral 7b**. In detail, we use their instruction-tuned models which are **gpt-3.5-turbo-0613**, **Llama2-7b-chat**, **Llama2-13b-chat** and **Mistral-7B-Instruct-v0.2**. The temperature is set to 0.1 for all LLMs and the repetition penalty is set to 1.1 for all LLMs but ChatGPT. Other hyperparameters for sampling follow the defaults. Experiments are run once with a fixed random seed due to limits of computational resources.

3.2 Datasets and Evaluation Metric

We conduct experiments on two datasets, ASAP[1] and TOEFL11. ASAP (Automated Student Assessment Prize) comprises 12,978 essays provided by students spanning in grade level from 7 to 10. The essays were written in response to 8 prompts of varying genres and score ranges. TOEFL11 consists of 12,100 essays written by test takers of TOEFL iBT which measures the academic English proficiency of non-native English speakers. The statistics of the two datasets are shown in Table 1.

Dataset	Prompt	#Essay	Genre	Avg Len	Range
ASAP	1	1783	ARG	427	2-12
	2	1800	ARG	432	1-6
	3	1726	RES	124	0-3
	4	1772	RES	106	0-3
	5	1805	RES	142	0-4
	6	1800	RES	173	0-4
	7	1569	NAR	206	0-30
	8	723	NAR	725	0-60
TOEFL11	1	1656	ARG	342	l/m/h
	2	1562	ARG	361	l/m/h
	3	1396	ARG	346	l/m/h
	4	1509	ARG	340	l/m/h
	5	1648	ARG	361	l/m/h
	6	960	ARG	360	l/m/h
	7	1686	ARG	339	l/m/h
	8	1683	ARG	344	l/m/h

Table 1: Statistics of ASAP and TOEFL11. **Genre**: ARG (argumentative), RES (source-dependent), NAR (narrative). **Avg Len**: Average essay length in words. **Range**: Score range (l/m/h for low/medium/high).

For ASAP, 10% of the essays from each prompt are randomly sampled for testing. There is no significant difference in the average essay score between the sample and the population in all prompts (Z-test, $\alpha = 0.05$). For TOEFL11 dataset, we adopt the test split from the 2013 Native Language Identification Shared Task (Tetreault et al., 2013), which consists of 1,100 essays collected from 8 prompts.

We use Quadratic Weighted Kappa (QWK) to measure the agreement between groundtruth scores and predicted scores. QWK is commonly used in AES research (Taghipour and Ng, 2016; Dong et al., 2017; Cao et al., 2020; Xie et al., 2022).

3.3 Implementation Details

Scoring Criteria. As outlined in Section 2.1, ChatGPT generates scoring criteria based on the rubric guidelines for human raters. For ASAP, we use the dataset's original rubric guidelines for each prompt. For TOEFL11, we follow Mizumoto and Eguchi (2023) and choose IELTS Task2 Writing Band Descriptor as the rubric guidelines instead of TOEFL Independent Writing Rubrics since the former provides more fine-grained descriptions. Note that each prompt in ASAP has a distinct scoring criteria whereas all prompts in TOEFL11 share the same scoring criteria.

Scoring Strategy. For ASAP, the predicted score $\hat{y}$ is rounded to integers. For TOEFL11, $\hat{y}$ is

[1] https://www.kaggle.com/c/asap-aes/data

Dataset	LLM	Method	P1	P2	P3	P4	P5	P6	P7	P8	Avg.
ASAP	ChatGPT	Vanilla	0.032	0.220	0.476	0.597	0.479	0.637	**0.289**	**0.527**	0.407
		MTS	**0.138**	**0.443**	**0.502**	**0.611**	**0.662**	**0.668**	0.261	0.157	**0.430**
	Llama2-7b-chat	Vanilla	0.163	**0.468**	0.016	0.000	0.117	0.304	0.187	0.192	0.181
		MTS	**0.371**	0.466	**0.504**	**0.378**	**0.673**	**0.507**	**0.563**	**0.409**	**0.483**
	Llama2-13b-chat	Vanilla	0.158	0.189	0.069	0.004	0.280	0.393	0.333	0.213	0.205
		MTS	**0.591**	**0.541**	**0.552**	**0.591**	**0.620**	**0.590**	**0.483**	**0.511**	**0.560**
	Mistral-7b-instruct	Vanilla	0.206	**0.512**	0.516	0.587	0.457	0.601	0.624	**0.304**	0.483
		MTS	**0.545**	0.455	**0.550**	**0.691**	**0.540**	**0.657**	**0.672**	0.289	**0.550**
TOEFL11	ChatGPT	Mizumoto and Eguchi (2023)	0.308	0.165	0.252	0.182	0.181	0.336	0.318	0.318	0.258
		Vanilla	0.215	0.240	0.337	0.332	0.227	0.306	0.237	0.306	0.275
		MTS	**0.495**	**0.447**	**0.651**	**0.595**	**0.489**	**0.496**	**0.500**	**0.536**	**0.526**
	Llama2-7b-chat	Mizumoto and Eguchi (2023)	0.009	0.047	0.085	0.140	0.133	0.027	0.032	0.023	0.062
		Vanilla	0.000	-0.007	0.026	-0.006	0.041	0.015	0.111	0.020	0.025
		MTS	**0.545**	**0.395**	**0.540**	**0.472**	**0.497**	**0.388**	**0.419**	**0.437**	**0.462**
	Llama2-13b-chat	Mizumoto and Eguchi (2023)	0.125	0.132	0.400	0.130	0.462	0.176	0.113	0.123	0.208
		Vanilla	0.196	0.156	0.285	0.268	0.165	0.329	0.249	0.257	0.238
		MTS	**0.580**	**0.373**	**0.703**	**0.557**	**0.612**	**0.457**	**0.620**	**0.630**	**0.567**
	Mistral-7b-instruct	Mizumoto and Eguchi (2023)	0.227	0.218	0.383	0.350	0.222	0.129	0.132	0.196	0.232
		Vanilla	0.486	0.259	0.355	0.344	0.431	0.456	0.286	0.383	0.375
		MTS	**0.637**	**0.510**	**0.654**	**0.587**	**0.516**	**0.554**	**0.564**	**0.677**	**0.587**

Table 2: Zero-shot evaluation results in QWK. **P1-8** denotes Prompt 1-8. The best measures in each LLM are in bold. Negative value indicates that the predictions are worse than random.

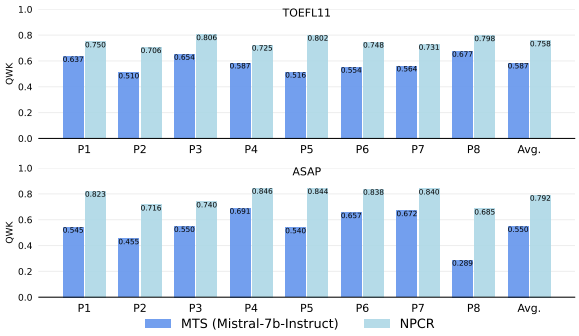


Figure 4: Comparison between MTS (Mistral-7b-instruct) and supervised SOTA (NPCR).

first scaled to $[1, 5]$ (Blanchard et al., 2013), then mapped to low/medium/high with respect to the thresholds of 2.25 and 3.75.

3.4 Comparisons with Other Methods

We compare MTS against two LLM-based zero-shot baselines as well as a supervised SOTA model.

Rubric Scoring (Mizumoto and Eguchi, 2023) is a zero-shot approach designed for TOEFL11 which provides the LLM with human rubrics and asks it to assign an overall score within the target range.

Vanilla is a zero-shot approach that asks the LLM to assign an overall score within the target range and provide rationales before the score to elicit CoT. Requiring rationales has been shown to improve zero-shot scoring performance on the CEFR scale (Yancey et al., 2023). See Appendix C.2 for the full template.

NPCR (Xie et al., 2022) is the state-of-the-art supervised prompt-specific model which predicts the score difference of two input essays based on pairwise ranking objective. For a valid comparison, we re-implement NPCR to ensure it is evaluated on the same test set as MTS. See Appendix D for the training details.

4 Main Results

The zero-shot evaluation results are shown in Table 2. Building upon asking the LLM to assign a score, providing human rubrics (Mizumoto and Eguchi, 2023) and requiring rationales (Vanilla) yield similar average QWK on TOEFL11. MTS not only leverages the scoring rubrics but also retrieves evidence in trait-specific manner, consistently and significantly outperforming the baseline(s) in average QWK across all LLMs, on both datasets. In general, MTS achieves great performance gains over the baselines for Llama 2 series and moderate gains for ChatGPT and Mistral-7b-Instruct. For instance, MTS greatly improves over Vanilla using Llama2-13b-chat, with gains of 0.355 (0.205 $\rightarrow$ 0.560) on ASAP and 0.329 (0.238 $\rightarrow$ 0.567) on TOEFL11. This reassures the importance of careful prompt design to elicit LLM's ample potential to perform AES. Moreover, the superiority of MTS remains solid across diverse settings of AES reflected in the datasets, including variations in essay genre and the first language (L1) backgrounds of the test takers.

By comparing different LLMs, we observe that Mistral-7b-instruct achieves the most competitive performance under different prompting methods overall, reaching average QWK of 0.550 on ASAP and 0.587 on TOEFL11. Interestingly, ChatGPT underperforms the small-sized Mistral-7b-Instruct across all methods but Rubric Scoring, implying that the model size might not be a decisive factor in the performance. Nevertheless, the larger model tends to perform better within the same model type, evidenced by the comparison between Llama2-7b-chat and Llama2-13b-chat.

In terms of the comparison with the supervised SOTA, the gap in average QWK between MTS and NPCR can be narrowed down to 0.171 on TOEFL11 and 0.242 on ASAP using Mistral-7b-instruct, as shown in Figure 4. This can be promising given that NPCR consumes approximately a thousand labeled essays per prompt to reach the SOTA. We expect further reduction in the gap with more powerful LLMs.

5 Analysis

We conduct experiments to derive insights into the success of MTS by analyzing each of the modules in MTS.

5.1 Analysis of Multi Trait Decomposition

MTS requires LLMs to follow the predefined traits and scoring criteria (i.e., the **guidance**), the quality of which is important to elicit their potential. In this section, we examine the effect of different guidance generation methods.

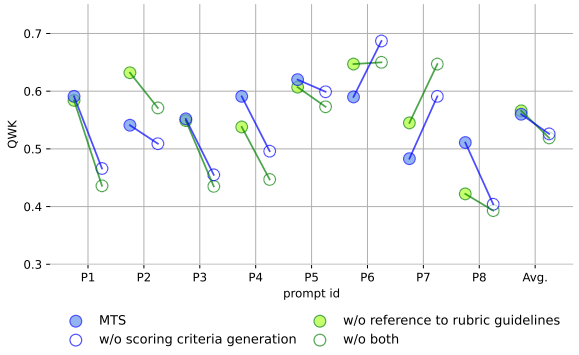
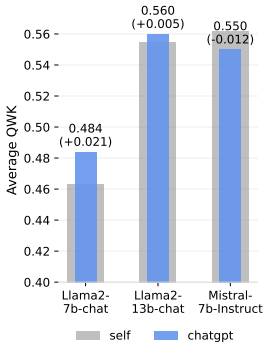


Figure 5: Ablation over (1) reference to rubric guidelines and (2) scoring criteria generation. The QWKs of Llama2-13b-chat on ASAP are reported.

Role of Rubric Guideline and Scoring Criteria. When decomposing the writing proficiency into multiple traits, MTS (1) refers to rubric guidelines designed by human raters, and (2) generates scoring criteria along with each trait. We conduct ablation study over both options to investigate impact on the performance.

Figure 5 shows that the reference to the rubric guidelines has negligible impact on the average QWK. Yet skipping the reference leads to higher standard deviation ($0.044 \rightarrow 0.066$) of the QWKs, indicating a fluctuating performance across the prompts. Next, the average QWK drops greatly after discarding the scoring criteria, both with and without the reference to human standards. One possible reason could be that the scoring criteria regulates the LLMs' behavior, encouraging them to adhere to the predefined criteria for better consistency.



(a) Impact of the source of the guidance on average QWK.

Llama2-13b-chat

Scoring Criteria:
- 0-2 points: The writer's structure is confusing, disorganized, or absent.
- 3-5 points: The writer's structure is somewhat clear and organized, but may lack coherence and flow.
- 6-8 points: The writer's structure is largely effective and easy to follow, with clear transitions between sections.
- 9-10 points: The writer's structure is highly effective and engaging, with a clear introduction, well-developed body paragraphs, and a satisfying conclusion.

ChatGPT

Scoring Criteria:
- 0-2: The essay lacks organization and structure, making it challenging to follow or understand.
- 3-5: Shows minimal organization but lacks a coherent structure or transitions.
- 6-8: Demonstrates satisfactory organization with some coherence, though transitions may be weak in connecting ideas.
- 9-10: Strong organization with clear and smooth transitions, presenting ideas logically and coherently.

(b) The scoring criteria for "Organization and Structure" generated by Llama2-13b-chat and ChatGPT, respectively.

Figure 6: Comparison of guidance (traits and scoring criteria) generated by ChatGPT and LLM itself.

Leveraging ChatGPT for Guidance Generation. MTS leverages ChatGPT to generate guidance for all LLMs used for actual scoring. We assess the significance of the guidance generated by ChatGPT over the one generated by the LLM itself. Figure 6(a) shows that ChatGPT-generated guidance slightly outperforms self-generated ones for Llama2-7b-chat and Llama-13b-chat, whereas the opposite holds for Mistral-7b-Instruct. Figure 6(b) further reveals that both ChatGPT and Llama2-13b-chat produce reasonable scoring criteria. Therefore, while ChatGPT-generated guidance is readily applicable to various LLMs, it is still valid for the LLMs to use the self-generated guidance instead.

5.2 Ablation Study of Trait Specialization

With the scoring guidance given, the structure of Trait Specialization has evolved over Vanilla with

Method	P1	P2	P3	P4	P5	P6	P7	P8	Avg.
Vanilla	0.158	0.189	0.069	0.004	0.280	0.393	0.333	0.213	0.205
Scoring All Traits Sequentially	0.267	0.237	0.354	0.280	0.292	0.396	**0.512**	0.111	0.306
Scoring Each Trait Independently	0.489	0.487	0.478	0.467	0.568	0.549	0.471	0.430	0.492
+Quote Retrieval and Scoring	**0.591**	**0.541**	**0.552**	**0.591**	**0.620**	**0.590**	0.483	**0.511**	**0.560**

Table 3: Ablation study of Trait Specialization. The QWKs of Llama2-13b-chat measured on ASAP are reported. The best performance in each column is in bold. See Appendix C.3 and C.4 for the templates of 2nd and 3rd method.

a sequence of incremental improvements:

Scoring All Traits Sequentially. The LLM is required to read through the entire guidance and generate the evaluation-score pairs for all traits sequentially in a single turn of a conversation. Trait scores assigned in the fixed range of $[0, 10]$ are aggregated and scaled the same way as MTS.

Scoring Each Trait Independently. For each trait, a new trait-specific conversation is initiated where the LLM reads through the guidance restricted to the specific trait and generate the evaluation-score pair for the trait in a single turn. Trait scores are aggregated and scaled the same way as MTS.

Quote Retrieval and Scoring. On the basis of scoring each trait independently, we divide each conversation into **two turns** where the quote retrieval task precedes the scoring task, instead of generating evaluation-score pair in a single turn. This constitutes our proposed MTS.

Table 3 demonstrates that all of the above design choices have positive impact on the average QWK. Building on Vanilla, we have considered two ways of integrating the predefined guidance into the conversation: scoring all traits sequentially and scoring each trait independently. While both methods are beneficial, independently scoring each trait proves to be much more effective, ensuring the evaluation on one trait is not influenced by the other traits. In addition, quote retrieval and scoring further enhances performance on all prompts, highlighting the importance of an in-depth analysis of the essay's content prior to assigning the score.

A characteristic these three strategies share in common is to decompose a complex problem of assigning an overall score into simpler subproblems and more specific tasks. We observe that this idea succeeds in zero-shot AES, in addition to complex reasoning tasks (Zhou et al., 2022).

5.3 Analysis of Trait Aggregation and Scaling

Merit of Diversified Assessment. MTS averages multiple trait scores $\{\hat{y}_j^{(i)}\}_{j=1}^{N_T}$ and scales the averaged scores to the target score range. We examine whether aggregating more trait scores improves performance by selecting all subsets of the original four traits with varying cardinality of $n \in \{2, 3, 4\}$ and evaluating the average QWK for each cardinality, as shown in Figure 7. We observe a clear tendency where a higher number of traits leads to elevated performance, suggesting that different traits are complementary to each other. In other words, MTS takes advantages of diversified assessment of writing proficiency.

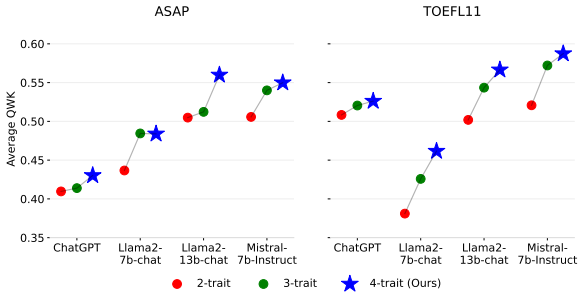


Figure 7: QWK under different numbers of traits, averaged over 8 prompts. n-trait denotes further averaging the performance over all combinations ($\binom{4}{n}$) of n traits.

	ASAP				TOEFL11			
	C	L7	L13	M7	C	L7	L13	M7
Vanilla+	**0.455**	0.189	0.303	0.522	0.463	0.164	0.196	0.496
MTS	0.430	**0.484**	**0.560**	**0.550**	**0.526**	**0.462**	**0.567**	**0.587**

Table 4: Average QWK of the overall (**Vanilla+**) and diversified (**MTS**) assessment. **C**: ChatGPT; **L7/13**: Llama2-7/13b-chat; **M7**: Mistral-7b-instruct. Best QWK in each column is in bold.

To further inspect if diversified scoring brings more benefit than the overall evaluation, we consider a new baseline called **Vanilla+** which predicts the *overall* score in the same range as MTS (from 0 to 10) and applies the same scaling method as MTS[2]. As shown in Table 4, the average of trait scores leads to better estimates of the writing quality than the overall score in most cases.

[2]Since the overall scores given by Vanilla+ lack diversity, applying outlier clipping may result in all scores being clipped to a single value for some prompts, in which case we set their QWKs to zero.

	ASAP				TOEFL11			
	C	L7	L13	M7	C	L7	L13	M7
fixed	0.350	0.254	0.477	0.520	0.357	0.071	0.445	0.385
minmax	0.405	0.477	0.553	0.529	**0.526**	0.438	0.420	0.499
+clipping	**0.430**	**0.484**	**0.560**	**0.550**	**0.526**	**0.462**	**0.567**	**0.587**

Table 5: Average QWK under different scaling methods. **clipping**: outlier clipping. Best QWK in each column is in bold.

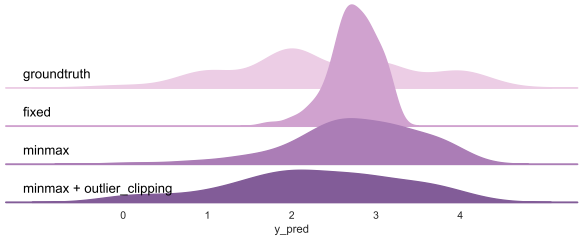


Figure 8: Distributions (KDE) of $\hat{y}$, estimated on ASAP Prompt 5 using Llama2-7b-chat.

Significance of Min-max Scaling and Outlier Clipping. We conduct a comparative analysis of various scaling methods. We consider a simple baseline of **fixed scaling** where $\hat{y}_{min}$ and $\hat{y}_{max}$ are fixed to 0 and 10 during min-max scaling. As shown in Table 5, fixed scaling mostly fails, especially with Llama2-7b-chat showing significantly degraded performance. Conversely, min-max scaling greatly outperforms fixed scaling in average QWK for most cases, and clipping outliers further brings consistent and sometimes crucial improvements over min-max scaling.

Figure 8 shows Kernel Density Estimation (KDE) of $\hat{y}$, providing insights into how the scaling mechanism of MTS works: (1) The LLM displays a bias toward predicting within a specific and concentrated interval, as shown in fixed scaling; (2) Min-max scaling effectively addresses the scoring bias by spreading its predictions across the target score range; (3) Outlier clipping further alleviates the distortion of the distribution caused by the outliers, contributing to robust performance.

6 Related Work

Automated Essay Scoring. Discovering essay representation discriminative of writing qualities has been a major concern in AES. Early works had explored extracting hand-crafted linguistic features (Yannakoudakis et al., 2011; Persing and Ng, 2013; Chen and He, 2013), learning features via neural networks (Taghipour and Ng, 2016; Dong et al., 2017; Tay et al., 2018), as well as combining the two (Dasgupta et al., 2018; Uto et al., 2020). Particularly, recent works (Yang et al., 2020; Xie et al., 2022) exhibited high level of agreement with human raters in the prompt-specific setting.

Despite the success of prompt-specific models, they experienced significant performance drops when adapted to unseen prompts (Jin et al., 2018; Cozma et al., 2018), limiting their applicability in practice. To tackle this challenge, recent works introduce domain adaptation (Phandi et al., 2015; Cao et al., 2020) and generalization (Ridley et al., 2020; Jiang et al., 2023b) techniques to AES. Nevertheless, these methods still consume considerable amount of labelled essays of the source domain(s).

The advent of LLM and its versatility on a wide range of downstream tasks (Wei et al., 2022; Kojima et al., 2022) has raised attention to its potential in essay scoring, giving rise to LLM-based zero-shot AES. For instance, Mizumoto and Eguchi (2023) prompt text-davinci-003 to only respond with the essay's overall score referring to a given scoring rubric. Yancey et al. (2023) reveal that LLMs benefit from referring to a detailed rubric or generating a rationale prior to the score in zero-shot setting. However, GPT-4 with these strategies still exhibits suboptimal performance similar to length-only classifier which relies solely on essay's character length for its prediction.

LLM-based Evaluators. Recent studies have investigated the use of LLM as a reference-free evaluation metric for natural language generation (NLG) tasks (Fu et al., 2023; Chiang and Lee, 2023). Zheng et al. (2023) showed that strong LLMs (e.g., GPT-4) highly matched human preferences when evaluating the quality of AI assistant's respones in open-ended QA tasks. Furthermore, LLMs achieved state-of-the-art agreement with human judgements in translation quality assessment (Kocmi and Federmann, 2023) and summarization task (Wang et al., 2023), verifying LLM evaluators' scalability to a variety of NLG tasks.

In addition to the variations in prompt design to fit different tasks, sophisticated prompting strategies have been devised to further elicit LLMs' potential as evaluators: Liu et al. (2023) prompt LLM with auto-generated evaluation steps to elicit CoT, achieving superior correlation with human ratings on text summarization and dialogue generation task; Chan et al. (2023); Zhang et al. (2023b) devised a multi-agent cooperation framework where LLMs with diverse role descriptions synergize to refine the evaluation result.

7 Conclusion

We present MTS (Multi Trait Specialization), an LLM-based zero-shot prompting framework for AES. In essence, MTS leverages trait-specific conversations with LLM to derive the overall score from diverse aspects of writing proficiency. Experimental results show that MTS consistently outperforms Vanilla approach in average QWK across all LLMs and datasets, while also substantially reducing the gap with fully supervised SOTA model. Our analysis reveals key insights into the success of MTS: (1) Providing the LLM with predefined scoring criteria regulates its scoring behavior, contributing to improved performance; (2) MTS benefits from subproblem decomposition such as independent trait-specific conversations and separation of quote retrieval task and scoring task; (3) Outlier clipping and min-max scaling effectively map the predictions to arbitrary target score range, alleviating the LLM's scoring bias as well as min-max scaling's sensitivity to the outliers.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (62076008) and the Key Project of Natural Science Foundation of China (61936012).

Limitations

First, MTS consumes significantly longer inference time than Vanilla approach due to multiple rounds of conversation. In addition, the pre-generated scoring criteria is included in the conversations for every essay, further increasing the computational cost. We believe that distilling the predictions to small models such as BERT (Devlin et al., 2019) would be a promising direction for cost-effective inference. Second, our analysis still demands more detailed illustrations of the LLMs' scoring behavior. For instance, LLMs produce verbal evaluations on the quotes they found in the essay, but it is unclear whether there is a faithful relation between the evaluations and the score. Moreover, it should be examined if the inclusion of scoring criteria truly leads to a more consistent scoring behavior, or it merely shifts the average of predicted scores closer to the groundtruth score while maintaining the same degree of inconsistency. Third, outlier clipping using Q1 and Q3 may not make min-max scaling completely resistant to outliers. We have not conducted extensive experiments with more robust ways of addressing outliers.

Ethics Statement

Potential Risks Our method does not guarantee fair evaluations, that is, MTS might reinforce LLMs' scoring tendency of favoring certain social groups. For example, it is possible that the predictive outcomes assign higher scores to a group with certain L1 (first language) background than the others. In addition, the datasets (ASAP and TOEFL11) we use might disproportionately represent certain demographic group, potentially leading to a biased conclusion. We partially address this concern by selecting TOEFL11 test dataset that contains equal number of samples from each L1. For ASAP, there is no demographic information provided.

Use of Scientific Artifacts We use the open source scikit-learn package (v1.0.2) (Pedregosa et al., 2011) for the calculation of QWK. We conduct experiments with ASAP (Hamner et al., 2012) and TOEFL11 (Blanchard et al., 2013) datasets, which are available for non-commercial research purposes. ASAP have anonymized personally identifying information from the essays by replacing them with symbols. TOEFL11 only includes essays with the author's permission for research use. As for the LLMs used in our study, OpenAI authorizes exploring its LLMs including ChatGPT (OpenAI, 2022) through its API for research publication (see OpenAI's sharing and publication policy). Llama 2 (Touvron et al., 2023) and Mistral 7b (Jiang et al., 2023a) are licensed under the Llama 2 Community license and Apache-2.0 license, respectively, both permitting research use.

Computational Budget We use a single NVIDIA A40 for each model inference including Llama2-7b-chat, Llama2-13b-chat and Mistral-7b-Instruct-v0.2. With only one sample in each batch, running MTS on ASAP test set takes approximately one GPU day for Llama2-7b-chat and Mistral-7b-Instruct-v0.2 and two GPU days for Llama2-13b-chat. Running time of MTS for ChatGPT through OpenAI API was similar to that of Llama2-13b-chat.

References

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

Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. 2013. Toefl11: A corpus of non-native english. *ETS Research Report Series*, 2013(2):i–15.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Yue Cao, Hanqi Jin, Xiaojun Wan, and Zhiwei Yu. 2020. Domain-adaptive neural automated essay scoring. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1011–1020.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*.

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

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.

Mădălina Cozma, Andrei Butnaru, and Radu Tudor Ionescu. 2018. Automated essay scoring with string kernels and word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 503–509, Melbourne, Australia. Association for Computational Linguistics.

Tirthankar Dasgupta, Abir Naskar, Lipika Dey, and Rupsa Saha. 2018. Augmenting textual qualitative features in deep convolution recurrent neural network for automatic essay scoring. In *Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications*, pages 93–102, Melbourne, Australia. Association for Computational Linguistics.

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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Semire Dikli. 2006. An overview of automated scoring of essays. *The Journal of Technology, Learning and Assessment*, 5(1).

Semire Dikli and Susan Bleyle. 2014. Automated essay scoring feedback for second language writers: How does it compare to instructor feedback? *Assessing writing*, 22:1–17.

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, 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 (CoNLL 2017)*, pages 153–162, Vancouver, Canada. Association for Computational Linguistics.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166*.

Ben Hamner, Jaison Morgan, lynnvandev, Mark Shermis, and Tom Vander Ark. 2012. The hewlett foundation: Automated essay scoring.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023a. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Zhiwei Jiang, Tianyi Gao, Yafeng Yin, Meng Liu, Hua Yu, Zifeng Cheng, and Qing Gu. 2023b. Improving domain generalization for prompt-aware essay scoring via disentangled representation learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12456–12470, Toronto, Canada. Association for Computational Linguistics.

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.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Rahul Kumar, Sandeep Mathias, Sriparna Saha, and Pushpak Bhattacharyya. 2022. Many hands make light work: Using essay traits to automatically score essays. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1485–1495, Seattle, United States. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Atsushi Mizumoto and Masaki Eguchi. 2023. Exploring the potential of using an ai language model for automated essay scoring. *Research Methods in Applied Linguistics*, 2(2):100050.

OpenAI. 2022. Introducing chatgpt. `https://openai.com/blog/chatgpt`, Last accessed on 2024-01-15.

OpenAI. 2023. Gpt-4 technical report.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Isaac Persing and Vincent Ng. 2013. Modeling thesis clarity in student essays. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 260–269, Sofia, Bulgaria. Association for Computational Linguistics.

Peter Phandi, Kian 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.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*.

Robert Ridley, Liang He, Xin-yu Dai, Shujian Huang, and Jiajun Chen. 2021. Automated cross-prompt scoring of essay traits. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 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. *arXiv preprint arXiv:2008.01441*.

Songwon Seo. 2006. *A review and comparison of methods for detecting outliers in univariate data sets*. Ph.D. thesis, University of Pittsburgh.

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.

Yi Tay, Minh Phan, Luu Anh Tuan, and Siu Cheung Hui. 2018. Skipflow: Incorporating neural coherence features for end-to-end automatic text scoring. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Joel Tetreault, Daniel Blanchard, and Aoife Cahill. 2013. A report on the first native language identification shared task. In *Proceedings of the eighth workshop on innovative use of NLP for building educational applications*, pages 48–57.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Masaki Uto, Yikuan Xie, and Maomi Ueno. 2020. Neural automated essay scoring incorporating handcrafted features. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6077–6088, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. Is ChatGPT a good NLG evaluator? a preliminary study. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 1–11, Hybrid. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

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.

Kevin P Yancey, Geoffrey Laflair, Anthony Verardi, and Jill Burstein. 2023. Rating short l2 essays on the

cefr scale with gpt-4. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 576–584.

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.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A new dataset and method for automatically grading ESOL texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 180–189, Portland, Oregon, USA. Association for Computational Linguistics.

Chenhan Yuan, Qianqian Xie, and Sophia Ananiadou. 2023. Zero-shot temporal relation extraction with ChatGPT. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 92–102, Toronto, Canada. Association for Computational Linguistics.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B Hashimoto. 2023a. Benchmarking large language models for news summarization. *arXiv preprint arXiv:2301.13848*.

Xinghua Zhang, Bowen Yu, Haiyang Yu, Yangyu Lv, Tingwen Liu, Fei Huang, Hongbo Xu, and Yongbin Li. 2023b. Wider and deeper llm networks are fairer llm evaluators. *arXiv preprint arXiv:2308.01862*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V Le, et al. 2022. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations*.

A Instructions for Multi Trait Decomposition

For multi trait decomposition, we prompt ChatGPT with different instructions for ASAP and TOEFL11. While ASAP's rubric guidelines do not necessarily consist of four traits, TOEFL11's rubric guideline (here we use IELTS Task2 Writing Band Descriptor, see Section 3.3) explicitly divides writing proficiency into four traits. Therefore, for ASAP, we ask ChatGPT to first determine the four traits and generate scoring criteria for each trait. In contrast, for TOEFL11, we ask it to generate scoring criteria based on the traits determined by the rubric guideline. Specifically, we use the following instructions for ASAP and TOEFL11 where the contents to be filled are substituted with comments between double curly braces `{{ }}`:

ASAP

```
[Excerpt]
{{excerpt (specific to prompt3-6)}}
(end of [Excerpt])
[Prompt]
{{prompt}}
(end of [Prompt])
[Rubric Guidelines]
{{rubric guidelines}}
(end of [Rubric Guidelines])
Refer to the provided [Prompt] and
[Rubric Guidelines] to generate an
essay scoring rubric divided into
four primary dimensions of writing
quality.  Adhere to the requirements
of [Prompt] and [Rubric Guidelines]
when you determine the four dimensions
of writing quality.  At each dimension,
make sure a brief description of the
dimension is added before the scoring
criteria.  The score scale of each
dimension ranges from 0 to 10, and the
total score is 40.
```

TOEFL11

```
[Scoring Rubric]
{{IELTS Task2 Writing Band Descriptor,
restricted to one trait}}
(end of [Scoring Rubric])
Refer to [Scoring Rubric] to generate
a scoring criteria with score
ranging from 0 to 10, following the
instruction below:
1.  Briefly describe ‘‘{{trait}}’’ with
one sentence.
2.  Divide the score range [0-10] into
5 appropriate intervals.
3.  For each interval, summarize its
characteristics.
```

B Guidance from Multi Trait Decomposition

We present guidance from multi-trait decomposition, comprising: (1) trait, (2) trait description, and (3) scoring criteria. MTS uses (1) and (2) for the role prompt (system message) and (1) and (3) for conversations. ASAP uses different rubric guidelines for each of its eight prompts, generating distinct guidance for each, while TOEFL11 applies the same rubric for all eight, resulting in identical guidance. Due to length, we show only the guidance for ASAP Prompt 1 and all prompts for TOEFL11.

ASAP Prompt 1

Position and Thesis Clarity
This dimension evaluates how clearly the writer takes a stance on the effects of computers on people and how effectively this stance is conveyed in the thesis statement.
Scoring Criteria:
- 0-2: The position is unclear or absent. The thesis lacks a clear stance or is entirely missing.
- 3-5: The position is somewhat evident but lacks clarity or specificity in the thesis statement.
- 6-8: The position is clear, though it may require further specificity or nuance in the thesis statement.
- 9-10: The position is crystal clear, and the thesis statement effectively communicates the writer's stance with precision and depth.

Supporting Details and Evidence
This dimension assesses the quality and relevance of the supporting details and evidence used to back the writer's position.
Scoring Criteria:
- 0-2: Very minimal or no supporting details provided.
- 3-5: General or vague supporting details with minimal relevance to the thesis.
- 6-8: Adequate supporting details offered, although some lack specificity or relevance.
- 9-10: Rich and specific supporting details effectively back the thesis, providing compelling evidence and relevance.

Organization and Structure
This dimension evaluates the overall coherence, logical progression, and structural framework of the essay.
Scoring Criteria:
- 0-2: The essay lacks organization and structure, making it challenging to follow or understand.
- 3-5: Shows minimal organization but lacks a coherent structure or transitions.
- 6-8: Demonstrates satisfactory organization with some coherence, though transitions may be weak in connecting ideas.
- 9-10: Strong organization with clear and smooth transitions, presenting ideas logically and coherently.

Style, Language, and Audience Awareness
This dimension assesses the writer's language use, style, and their ability to engage the audience while demonstrating an awareness of the target readers.
Scoring Criteria:
- 0-2: Language use is awkward, and there's no evident awareness of the audience.
- 3-5: Language use is basic, and there's little attempt to engage the audience or demonstrate awareness.
- 6-8: Language is somewhat engaging, with occasional attempts to connect with the audience.
- 9-10: Language is engaging, sophisticated, and consistently demonstrates an acute awareness of the audience, effectively connecting with them.

TOEFL11

Task Response
This dimension evaluates how well the prompt is understood, addressed, and developed within the response.
0-2:
- Barely relevant or unrelated content to the given prompt.
- Lack of identifiable position or comprehension of the question.
- Minimal or no development of ideas; content may be tangential or copied.
3-4:
- Partially addresses the prompt but lacks depth or coherence.
- Discernible position, but unclear or lacking in support.
- Ideas are difficult to identify or irrelevant with some repetition.
5-6:
- Addresses main parts of the prompt but incompletely or with limited development.
- Presents a position with unclear or repetitive development.

- Some relevant ideas but insufficiently developed or supported.
7-8:
- Adequately addresses the prompt with clear and developed points.
- Presents a coherent position with well-extended and supported ideas.
- Some tendencies toward over-generalization or lapses in content, but mostly on point.
9-10:
- Fully and deeply explores the prompt with a clear, well-developed position.
- Extensively supported ideas relevant to the prompt.
- Extremely rare lapses in content or support; demonstrates exceptional depth and insight.

Coherence and Cohesion
This criterion assesses how well ideas are logically organized and connected within a written response.
0-2:
- Lack of coherence; response is off-topic or lacking in relevant message.
- Minimal evidence of organizational control or logical progression.
- Virtually absent or ineffective use of cohesive devices and paragraphing.
3-4:
- Ideas are discernible but arranged incoherently or lack clear progression.
- Unclear relationships between ideas, limited use of basic cohesive devices.
- Minimal or unclear referencing, inadequate paragraphing if attempted.
5-6:
- Some underlying coherence but lacks full logical organization.
- Relationships between ideas are somewhat clear but not consistently linked.
- Limited use of cohesive devices, with inaccuracies or overuse, and occasional repetition.
- Inconsistent or inadequate paragraphing.
7-8:
- Generally organized with a clear overall progression of ideas.
- Cohesive devices used well with occasional minor lapses.
- Effective paragraphing supporting coherence, though some issues in sequencing or clarity within paragraphs.
9-10:
- Effortless follow-through of ideas with superb coherence.
- Seamless and effective use of cohesive devices with minimal to no lapses.
- Skilful paragraphing enhancing overall coherence and logical progression.

Lexical Resource
This dimension evaluates the range, precision, and appropriateness of vocabulary used within a written response.
0-2:
- Minimal to no resource evident; extremely limited vocabulary or reliance on memorized phrases.
- Lack of control in word formation, spelling, and recognition of vocabulary.
- Communication severely impeded due to the absence of lexical range.
3-4:
- Inadequate or limited resource; vocabulary may be basic or unrelated to the task.
- Possible dependence on input material or memorized language.
- Errors in word choice, formation, or spelling impede meaning.
5-6:
- Adequate but restricted resource for the task.
- Limited variety and precision in vocabulary, causing simplifications and repetitions.
- Noticeable errors in spelling/word formation, with some impact on clarity.
7-8:
- Sufficient resource allowing flexibility and precision in expression.
- Ability to use less common or idiomatic items, despite occasional inaccuracies.
- Some errors in spelling/word formation with minimal impact on communication.
9-10:
- Full flexibility and precise use of a wide range of vocabulary.
- Very natural and sophisticated control of lexical features with rare minor errors.
- Skilful use of uncommon or idiomatic items, enhancing overall expression.

Grammatical Range and Accuracy
This dimension assesses the breadth of grammatical structures used and the precision in applying them within written communication.
0-2:
- Absence or extremely limited evidence of coherent sentence structures.

- Lack of control in grammar, minimal to no use of sentence forms.
- Language largely incomprehensible or irrelevant to the task.
3-4:
- Attempts at sentence forms but predominantly error-laden.
- Inadequate range of structures with frequent grammatical errors.
- Limited coherence due to significant errors impacting meaning.
5-6:
- Limited variety in structures; attempts at complexity with faults.
- Some accurate structures but with noticeable errors and repetitions.
- Clear attempts at complexity but lacking precision and fluency.
7-8:
- Adequate variety with some flexibility in using complex structures.
- Generally well-controlled grammar but occasional errors.
- Clear attempts at complexity and flexibility in sentence structures.
9-10:
- Extensive range with full flexibility and precision in structures.
- Virtually error-free grammar and punctuation.
- Exceptional command with rare minor errors, showcasing nuanced and sophisticated language use.

C Prompt Templates

In this section, we provide the exact templates of the prompts used for Vanilla and MTS. Our prompt design consists of three components: system message, user message and assistant message. Contents to be filled are placed between double curly braces `{{ }}`. Contents specific to ASAP are enclosed with *`ASAP(·)`* and those specific to TOEFL11 are enclosed with *`TOEFL11(·)`*, both in italic font.

C.1 Template for MTS

System Message

```
You are a member of the English essay
writing test evaluation committee.
Four teachers will be provided with
a [Prompt] and an [Essay] written
by a student in response to the
[Prompt]. Each teacher will score the
essays based on different dimensions
of writing quality. Your specific
responsibility is to score the essays
in terms of ‘‘{{trait}}". {{a brief
description of the trait}} Focus on
the content of the [Essay] and the
[Scoring Rubric] to determine the
score.
```

User Message

```
[Prompt]
{{prompt}}
(end of [Prompt])
ASAP([Note]
I have made an effort to remove
personally identifying information
from the essays using the Named
Entity Recognizer (NER). The relevant
entities are identified in the text
and then replaced with a string such
as "{PERSON}", "{ORGANIZATION}",
"{LOCATION}", "{DATE}", "{TIME}",
"{MONEY}", "{PERCENT}", ‘‘{CAPS}’’ (any
capitalized word) and ‘‘{NUM}’’ (any
digits)[3]. Please do not penalize the
essay because of the anonymizations.
(end of [Note]))
[Essay]
{{essay}}
(end of [Essay])
Q. List the quotations from the
[Essay] that are relevant to
‘‘{{trait}}’’ and evaluate whether each
quotation is well-written or not.
```

Assistant Message

```
{{a response from the LLM}}
```

User Message

```
[Scoring Rubric]
**{{trait}}**:
{{scoring criteria}}
(end of [Scoring Rubric])
Q. Based on the [Scoring Rubric] and
the quotations you found, how would
you rate the ‘‘{{trait}}’’ of this
essay? Assign a score from 0 to 10,
strictly following the [Output Format]
```

[3] In the original ASAP dataset, all named entities are marked in the format of "@named entity". We convert this format to "{named entity}" in order to make the boundaries of the named entities more explicit.

below.
[Output Format]
Score: <score>insert ONLY the numeric score (from 0 to 10) here</score>
(End of [Output Format])

C.2 Template for Vanilla

System Message

As an English teacher, your primary responsibility is to evaluate the writing quality of essays written by *ASAP(middle school students) TOEFL11(second language learners on an English exam)*. During the assessment process, you will be provided with a prompt and an essay. First, you should provide comprehensive and conrete feedback that is closely linked to the content of the essay. It is essential to avoid offering generic remarks that could be applied to any piece of writing. To create a compelling evaluation for both the student and fellow experts, you should reference specific content of the essay to substantiate your assessment. Next, your evaluation should culminate in assigning an overall score to the student's essay, *ASAP(measured on a scale from {{minimum score value}} to {{maximum score value}}, where higher score should reflect a higher level of writing quality. It's crucial to tailor your evaluation criteria to be well-suited for middle school level writing, taking into account the developmental stage and capabilities of these students.) TOEFL11(on a three level scale of "low", "medium" and "high". It's crucial to tailor your evaluation criteria to be well-suited for second language learners, taking into account their expected abilities.)*

User Message

[Prompt]
{{prompt}}
(end of [Prompt])
ASAP([Note]
I have made an effort to remove personally identifying information from the essays using the Named Entity Recognizer (NER). The relevant entities are identified in the text and then replaced with a string such as "{PERSON}", "{ORGANIZATION}", "{LOCATION}", "{DATE}", "{TIME}", "{MONEY}", "{PERCENT}", "{CAPS}" (any capitalized word) and "{NUM}" (any digits). Please do not penalize the essay because of the anonymizations. (end of [Note]))
[Essay]
{{essay}}
(end of [Essay])
Strictly follow the format below to give your answer. Other formats are NOT allowed. Evaluation: <evaluation>insert evaluation here</evaluation>
ASAP(Score: <score>insert score ({{minimum score value}} to {{maximum score value}}) here</score>)
TOEFL11(Score: <score>insert score (choose one of "low", "medium", and "high") here</score>)

C.3 Template for Scoring All Traits Sequentially

System Message

You are an English teacher who is responsible for rating essays. You will be provided with a prompt and a student's essay written in response to the prompt. Follow the provided [Evaluation Steps] and assign a score to the essay in the specified format.

User Message

[Prompt]
{{prompt}}
(end of [Prompt])
ASAP([Note]
I have made an effort to remove personally identifying information from the essays using the Named Entity Recognizer (NER). The relevant entities are identified in the text and then replaced with a string such as "{PERSON}", "{ORGANIZATION}", "{LOCATION}", "{DATE}", "{TIME}", "{MONEY}", "{PERCENT}", "{CAPS}" (any capitalized word) and "{NUM}" (any

digits). Please do not penalize the essay because of the anonymizations. (end of [Note]))
[Essay]
{{essay}}
(end of [Essay])
[Evaluation Steps]
{{evaluation steps which include the entire guidance generated from multi trait decomposition (see Appendix B)}}
(end of [Evaluation Steps])
Q. For each step in [Evaluation Steps], assign a score from 0 to 10, strictly following the [Output Format] below.
[Output Format]
Step 1
- Evaluation: <evaluation>insert evaluation here</evaluation>
- Score: <score>insert ONLY the numeric score (from 0 to 10) here</score>
Step 2
- Evaluation: <evaluation>insert evaluation here</evaluation>
- Score: <score>insert ONLY the numeric score (from 0 to 10) here</score>
Step 3
- Evaluation: <evaluation>insert evaluation here</evaluation>
- Score: <score>insert ONLY the numeric score (from 0 to 10) here</score>
Step 4
- Evaluation: <evaluation>insert evaluation here</evaluation>
- Score: <score>insert ONLY the numeric score (from 0 to 10) here</score>
(end of [Output Format])

C.4 Template for Scoring Each Trait Independently

System Message
You are a member of the English essay writing test evaluation committee. Four teachers will be provided with a [Prompt] and an [Essay] written by a student in response to the [Prompt]. Each teacher will score the essays based on different dimensions of writing quality. Your specific responsibility is to score the essays in terms of "{{trait}}". {{a brief description of the trait}} Focus on the content of the [Essay] and the [Scoring Rubric] to determine the score.

User Message
[Prompt]
{{prompt}}
(end of [Prompt])
ASAP([Note]
I have made an effort to remove personally identifying information from the essays using the Named Entity Recognizer (NER). The relevant entities are identified in the text and then replaced with a string such as "{PERSON}", "{ORGANIZATION}", "{LOCATION}", "{DATE}", "{TIME}", "{MONEY}", "{PERCENT}", "{CAPS}" (any capitalized word) and "{NUM}" (any digits). Please do not penalize the essay because of the anonymizations. (end of [Note]))
[Essay]
{{essay}}
(end of [Essay])
[Scoring Rubric]
{{trait}}:
{{scoring criteria}}
(end of [Scoring Rubric])
Q: From the above [Scoring Rubric], how would you rate the "{{trait}}" of this essay? Respond a reasoning followed by a score from 0 to 10, strictly following the [Output Format] below:
[Output Format]
Reasoning: <reasoning>insert your reasoning which will justify your decision on the score</reasoning>
Score: <score>insert ONLY the numeric score (from 0 to 10) here</score> (End of [Output Format])

D Details of Re-implementation of NPCR

NPCR (Xie et al., 2022) was originally implemented on ASAP. For our re-implementation on

ASAP, while we leave out the same test set as MTS, the remaining data is randomly divided into train set and validation set by $4:1$. We re-implement NPCR on TOEFL11 as well with minimal adjustments: we use the train, dev and test split provided by Tetreault et al. (2013). The test set is identical to that of MTS. The predicted scores are scaled to $[1, 5]$ and mapped to low/medium/high with respect to the thresholds $[2.25, 3.75]$, which is consistent with Section 3.3.

For both datasets, the number of epochs is reduced from 80 to 20 so that the time for training and inference is kept in an acceptable range. Other settings are identical to the original implementation.