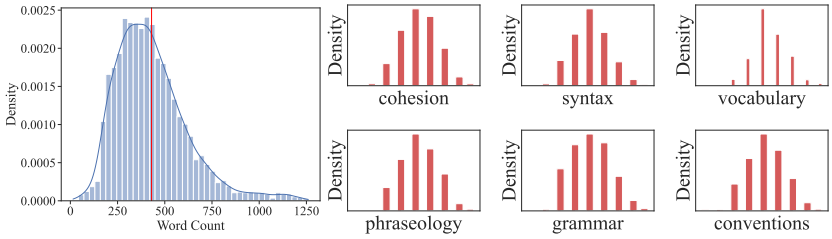
# Appendix

## ELLIPSE

To explore more about the ELLIPSE dataset, we visualized the word counts of essays, the distribution of six analytic criteria, as well as their correlations. In Figure 1a, the histogram of word counts exhibits positive skewness with a right-sided tail, suggesting there is a limited number of lengthy articles in the ELLIPSE corpus. Additionally, the majority of word counts are concentrated within the range of 250 to 500. In terms of the evaluation measure, Figure 1b illustrates that each score set conforms to an approximately normal distribution, wherein the mode, mean, and median are all centred around the value 3.

Figure 2 shows the Pearson correlation coefficients among the evaluation criteria all exceed 0.6, suggesting a substantial positive linear relationship. Specifically, a coefficient of 0.74 between Phraseology and Vocabulary indicates a notably strong positive correlation. We share the train and test datasets of ELLIPSE, as well as the subset at this repository.




(a) Text word count (b) Evaluation criteria

Fig. 1: Data distribution in ELLIPSE

Fig. 4 presents an example of the few-shot prompt we input into ChatGPT. The detailed rubrics of the ELLIPSE measures can be found in Few-shot_prompt.txt. For the zero-shot prompt, simply remove the examples that are necessary for the few-shot prompt.

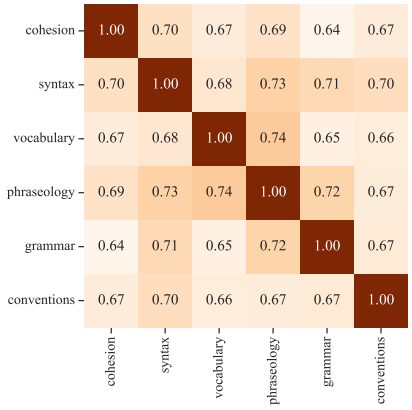



Fig. 2: Correlation of six evaluation criteria

**Grade the Cohesion, Syntax, Vocabulary, Phraseology, Grammar, and Conventions of the essay based on the rubrics. The scores range from 1.0 to 5.0 in increments of 0.5, with greater scores corresponding to greater proficiency in that measure.**

Cohesion Rubrics:
5: "Text organization consistently well controlled using ..."
...
Here are some examples of the evaluation results by human raters:

Example 1:
"*Essay*": "an essay"
"*Score*": "scores given by the human rater"
...

Fig. 3: Prompt with few-shot learning using ChatGPT

### ASAP++

ASAP++ is developed on top of ASAP and uses the same score range as the overall score range of the essays in ASAP. Table 1 presents the essay description of the ASAP dataset and Tabele 2 and Table 3 present the definitions of the measures. For a comprehensive understanding of the ASAP++ dataset, we recommend reading the paper "ASAP++: Enriching the ASAP Automated Essay Grading Dataset with Essay Attribute Scores".

### Experimental settings

We evaluate the AA models by a column-wise mean technique, i.e., for each task, we compute metric scores based on actual-predicted pairs and then average these scores across all tasks to obtain an overall assessment metric value.

**Table 1:** Description of the ASAP dataset

| Prompt ID | Essay Type | No. of Essays | Avg. Length | Score Range |
|---|---|---|---|---|
| Prompt 1 | Argumentative | 1785 | 350 | 1 - 6 |
| Prompt 2 | Argumentative | 1800 | 350 | 1 - 6 |
| Prompt 3 | Source-Dependent | 1726 | 150 | 0 - 3 |
| Prompt 4 | Source-Dependent | 1772 | 150 | 0 - 3 |
| Prompt 5 | Source-Dependent | 1805 | 150 | 0 - 4 |
| Prompt 6 | Source-Dependent | 1800 | 150 | 0 - 4 |

**Table 2:** Attributes of Argumentative / Persuasive Essays

| Attributes of Argumentative / Persuasive Essays | |
|---|---|
| Content | The quantity of relevant text present in the essay. |
| Organization | The way the essay is structured. |
| Word Choice | The choice and aptness of the vocabulary used in the essay. |
| Sentence Fluency | The quality of the sentences in the essay. |
| Conventions | Overall writing conventions to be followed, like spelling, punctuations, etc. |

**Table 3:** Attributes of Source-dependent Responses

| Attributes of Source-dependent Responses | |
|---|---|
| Content | The amount of relevant text present in the essay. |
| Prompt Adherence | A measure of how the writer sticks to the question asked in the prompt. |
| Language | The quality of the grammar and spelling in the response. |
| Narrativity | A measure of the coherence and cohesion of the response to the prompt. |

**Table 4:** Hyperparameter setting for MTAA

| Hyperparameters | Value |
|---|---|
| attention_probs_dropout_prob | 0.0 |
| hidden_act | gelu |
| hidden_dropout_prob | 0.0 |
| hidden_size | 768 |
| max_length | 512 |
| relative_attention | true |
| num_attention_heads | 12 |
| num_hidden_layers | 12 |

For DLRD setting, we initialize $S^k(t) = 1$ for $t = 1, 2$, and $\eta_0 = 1e - 5$. We set $\alpha$ to 10 and set $\gamma$ to 0.3. $n$ is set to 1, indicating the learning rate decay performs in every epoch.

For the shared encoder setting, due to the numerous parameters in the BERT style models and our limited computational resources, we empirically prioritize determining the hyperparameters with more impact on model performance.

Among these hyperparameters, we found that dropout has the most significant impact on the model and we found an optimal dropout rate of 0 for best model performance. We then focused on the learning rate and set it to 1e-5. The batch size in our experiments is set to 4. After exploring various pooling methods, Mean Pooling emerged as the best choice and thus it is employed to compress the shared representations in our system.

To ensure a fair comparison, we maintained consistent experimental settings among the BERT-style models, other than the differences in model designs themselves. We list the network configurations of the MTAA in Table 4.

### ChatGPT multi-dimensional evaluation

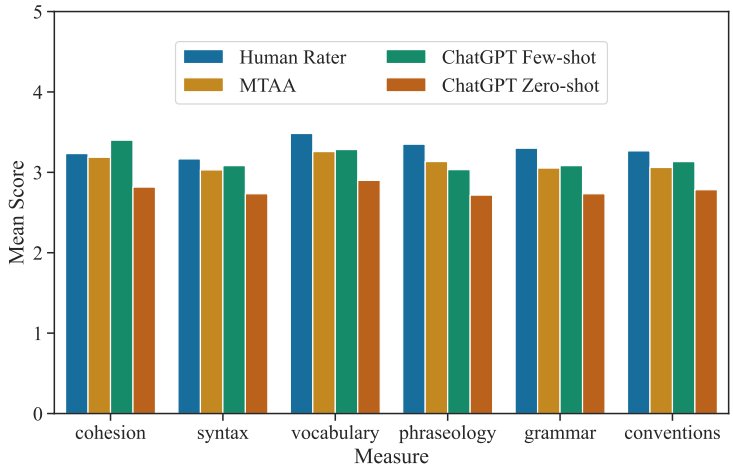



Fig. 4: Comparison of scores on measures by different methods on the ELLIPSE subset

Overall, as depicted in Fig. 4, essay scores from the human rater are generally the highest, except for the cohesion score, where ChatGPT few-shot outperforms. In terms of specific methods, the MTAA approach demonstrates the smallest discrepancy in scoring compared to the human rater. Examination of the zero-shot outputs of ChatGPT shows that the scores on all measures are less than those given by human raters, indicating that ChatGPT is a tougher grader. However, when supplied with three essay-score pairs for few-shot evaluation, the scores it generated were closer to those given by human raters.