Explainable Writing Scores via Fine-grained, LLM-Generated Features

James V. Bruno
Pearson
US School Assessment Technology
jimmy.bruno@pearson.com

Lee Becker
Pearson
US School Assessment Technology
lee.becker@pearson.com

Abstract

Transfomer-based models like BERT have increasingly been employed for automated essay scoring, as their high-dimensional representations of text are effective at capturing complex patterns in language. However, transformer-based representations are opaque and difficult to trace to the underlying human-defined constructs being assessed. This paper investigates the ability of LLMs to generate scores according to a rubric constructed from academic standards and evaluates the utility of these scores as features in a supervised regression model. We show that this produces a model that is reliable, construct-relevant, and interpretable. We evaluate this approach on six narrative writing items and find that, even with only 5 features, models can achieve QWKs exceeding 0.8, while also giving a concise and interpretable score explanation.

1 Background

Prior to advances in deep learning, the prevailing approach for Automated Essay Scoring (AES), relied on pairing supervised machine learning (ML) with a set of manually-crafted features (Attali and Burstein, 2006) that aimed for construct relevance. Feature engineering consisted of extracting linguistic phenomena which could serve as proxies for the underlying construct or assessed skill. For example, type-token ratio was used to capture vocabulary richness and semantic similarity measures approximated human ratings of essay cohesion (Graesser et al., 2004). In some cases, features may come from other models trained to predict a subtrait score (Somasundaran et al., 2018).

Advancements in NLP and ML have rapidly evolved the state-of-the art in automated essay scoring (AES). The shift toward dense language representations including semantic vectors (Deerwester et al., 1990), word embeddings (Mikolov et al., 2013a,b), and contextual embeddings (Peters et al., 2018) have yielded steady gains in AES performance (Foltz et al., 1999; Alikaniotis et al., 2016), typically measured using metrics like Quadratic Weighted Kappa (QWK). Transformers (Vaswani et al., 2017) and especially variants of BERT (Devlin et al., 2019) are now considered the de-facto approach for training AES systems (Mayfield and Black, 2020; Wang et al., 2022; Wang, 2024; Elmassry et al., 2025). However, these approaches are complex with features that are not directly interpretable and which number in the hundreds or thousands. Modern, deep-learning AES systems are effectively "black-box" solutions.

A growing body of research has applied explainable AI (xAI) to AES, including approaches such as attention visualization (Yang et al., 2020), multiple instance learning (Hellman et al., 2020) and post hoc explanation methods like LIME (Ribeiro et al., 2016), which surface links between regions of the text and model outputs. However, interpreting these explanations often requires subjective inference to connect model decisions to the constructs being assessed, and may lack direct construct relevance. The capacity for reasoning exhibited by Large Language Models (LLMs) presents new possibilities for explainability. LLMs can be prompted to generate auxiliary information such as rationalization of score (Li et al., 2023) or corresponding feedback (Stahl et al., 2024). While impressive, querying LLMs to provide justifications for their scoring decisions introduces the risk of self-referential explanations.

This work approaches explainability through the lens of subtrait scoring wherein the scored construct is broken down into sub-components with their own scores (Andrade-Lotero et al., 2025). Our framework is most similar to TRATES (Eltanbouly et al., 2025) which predicts rubric elements via LLM generation. Unlike TRATES, we limit the features of our models to only construct-relevant subtrait scores. By pairing LLM-generated subtrait

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 2: Works in Progress, pages 155–165

scores with linear regression, we can not only produce "dead simple" explainability, but also help build trust in the use of LLMs for essay scoring.

2 Aims

The primary aim of this study is to explore how using LLM-outputs as input features to train automated scoring models provides a straightforward path toward interpretable scores. Specifically, we explore the ability of LLMs to produce subtrait scores aligned to components of academic standards defining grade-level expectations, such as those set by Common Core State Standards Initiative (2010). We build, evaluate, and inspect simple linear models using the LLM-produced subtrait scores as features.

This work is part of an overarching goal to develop a collection of models that can assess subtraits and/or skills in support of a wide variety of writing items and rubrics (Andrade-Lotero et al., 2025). In this work we focus on assessing Common Core standards because the standards are decomposed into elements that align well with our notion of subtrait; however, nothing about the approach we illustrate here is limited to using standards.

Our specific research questions are as follows:

- How can we leverage academic standards to generate subtrait scores via an LLM?
- How does the performance of models trained with LLM-generated subtrait scores as features compare to operational models?
- Is an explainable linear regression model with only a handful of subtrait features operationally viable?
- Can the use of a simple, transparent linear regression model enhance the interpretability and trustworthiness of the LLM features?

3 Data

The experiment dataset consists of 6 eighth grade writing prompts administered as part of state-run, year-end summative assessments. Written responses and corresponding human scores come from the train and test data used to build and evaluate the operational scoring models. Responses flagged with codes such as blank, gibberish, passage copy or off-topic are excluded from this dataset. The scoring process is such that there are a minimum of 2 ratings per response in the model-building data set, with a third resolution rating as needed. We model the final score, that is, the score assigned as the end-result of the human scoring process for the item. The items in our experiment have high human agreement and extremely strong operationally deployed models, as shown in Table 1.

	Human	Deployed model
Prompt 1	0.940	0.918
Prompt 2	0.914	0.899
Prompt 3	0.897	0.893
Prompt 4	0.912	0.903
Prompt 5	0.888	0.886
Prompt 6	0.936	0.934

Table 1: Quadratic weighted kappa representing human-human agreement and human-machine agreement for operationally deployed models.

We focus on narrative writing items, as scoring in this genre is often perceived as subjective and multi-dimensional. Explainable models add transparency to this subjective process by linking assessment of narrative elements to the score. Additionally, narrative elements are not well captured by surface level features like n-grams or word count. This presents an opportunity to highlight how the deeper semantics of LLMs can flexibly accommodate a wide variety of subtraits.

We aim to take a uniform sample across score points and use 66% of the data for training and the remainder for testing, for a total of 1125 responses per score point. However, responses at the highest score points were underrepresented in the population for some prompts, and in these cases a uniform distribution is not possible. The score distribution for the experimental dataset appears in Table 2. Additionally, it was not possible to extract subtrait features from for every response, as in some cases the LLM returned malformed JSON or raised content filters. The response counts of the final train-test partitions for the 3 LLMs we use in our experiments appear in Table 3.

4 Method

As our goal is explainability, we wish to build the simplest, most interpretable model possible using the outputs from the LLM. As discussed below, we construct an LLM query with scoring instructions and a rubric based on the Common Core standard for 8th-grade narrative writing (Common Core State Standards Initiative, 2010).

	Train						Test					
	0	1	2	3	4	Total	0	1	2	3	4	Total
Prompt 1	150	150	150	150	117	717	75	75	75	75	58	358
Prompt 2	150	150	150	150	51	651	75	75	75	75	25	325
Prompt 3	150	150	150	121	28	599	75	75	75	60	14	299
Prompt 4	150	150	150	150	68	668	75	75	75	75	35	335
Prompt 5	150	150	150	103	43	596	75	75	75	52	23	300
Prompt 6	150	150	150	148	150	748	75	75	75	74	75	374
Total	900	900	900	822	457	3979	450	450	450	411	230	1991

Table 2: Score distribution of the training and test sets.

	Original Sample		GPT-4o Mini		Llama3.1 8B		Gemma 7B	
	Train	Test	Train	Test	Train	Test	Train	Test
Prompt 1	717	358	690	343	717	358	692	347
Prompt 2	651	325	643	323	651	325	640	322
Prompt 3	599	299	589	292	599	299	582	284
Prompt 4	668	335	645	328	668	335	655	327
Prompt 5	596	300	583	298	596	300	577	290
Prompt 6	748	374	743	372	748	374	733	367
Total	3979	1991	3893	1956	3979	1991	3879	1937

Table 3: Number of responses in Train/Test partitions. It was not possible to obtain LLM-based features from GPT-4o Mini or Gemma 7B for all responses due to content-filters and the LLM returning improper JSON. Therefore, the samples for some of the models tested are non-identical.

The LLM provides feature values in the form of subtrait scores, which we use to train and evaluate highly-interpretable linear regression models. The models predict the operational score using LLM-generated subtrait scores as features. We do not have ground-truth annotations for the subtrait scores.

4.1 LLM-generated features

The feature space centers around "subtrait scoring", wherein finer-grained scores reflect performance on a facet of a larger trait. Following the approach detailed in Andrade-Lotero et al. (2025), we query an LLM to score an essay given a rubric. To maintain consistency between items, we use standards-based rubrics instead of the items' original trait rubric. As we aim to make an assessment grounded in the Common Core standard for the narrative genre, we construct a 5-point rubric from the standard elements of `CCSS.ELA-LITERACY.W.8.3`, that is, the Common Core standard for 8th grade narrative writing. The standard reads "Write narratives to develop real or imagined experiences or events using effective technique, relevant descriptive details, and well-structured event sequences." The standard is further decomposed into the 5 standard elements that appear in Table 4, which we treat as subtraits.

We use an LLM to construct a rubric from these standard elements. Specifically, we embed each of the standard elements into instructions to create criteria for 5 score points, query Claude Haiku 3.5 (Anthropic, 2024) with the instructions, and manually verify the result. We choose a score range from 0 to 4 to reflect the original range on which the responses were scored. Example system instructions to create rubrics appear in Figure 1 and an example rubric appears in Figure 2 in the Appendix.

To produce subtrait scores, we submit the response and the rubric to 3 LLMs. We choose OpenAI's GPT-4o-Mini (OpenAI, 2024) and also two open-source models of similar size: Gemma 7B (DeepMind, 2024), and Llama 3.1 8B (Meta, 2024).

The LLM query to produce subtrait scores includes an instruction to provide feedback. This likely has a positive effect on the output given that eliciting reasoning is known to improve LLM results (see for example, Huang and Chang, 2023); however, we set aside an examination of this effect

W.8.3.A	Engage and orient the reader by establishing a context and point of view and introducing a narrator and/or characters; organize an event sequence that unfolds naturally and logically.
W.8.3.B	Use narrative techniques, such as dialogue, pacing, description, and reflection, to develop experiences, events, and/or characters.
W.8.3.C	Use a variety of transition words, phrases, and clauses to convey sequence, signal shifts, and show the relationships among experiences and events.
W.8.3.D	Use precise words and phrases, relevant descriptive details, and sensory language to capture the action and convey experiences and events.
W.8.3.E	Provide a conclusion that follows from and reflects on the narrated experiences or events.

Table 4: 8th grade narrative writing standard elements.

for future work. The prompt appears in Figure 3 in the Appendix.

4.2 Model training, evaluation, and explanation.

As our overarching goal is intepretability, we train a non-negative ridge regression model with L2 regularization for each item-LLM model pair, using cross-validation within the training set to tune the regularization parameter. Ridge regression is used instead of simple linear regression to minimize the effects of multicollinearity between subtrait scores, and the coefficients are required to be positive for ease of interpretability.

Models are evaluated on quadratic weighted kappa (Cohen, 1968), and we examine the coefficients for each of the 5 subtraits for the purpose of explaining the models.

5 Results

The quadratic weighted kappa for the 6 prompts and the 3 LLMs appear in Table 5. All models were able to predict the human score, with average test QWKs of 0.81, 0.78, and 0.59 for GPT-4o-Mini, Llama 3.1 8B, and Gemma 7B, respectively. GPT-4o-Mini had the best overall performance, but we are particuarly encouraged that an open source model with few parameters such as LLama 3.1 8B is viable. We note that there is only a 3-point difference between GPT-4o-Mini and LLama 3.1 8B, compared to the 22-point difference between GPT-4o-Mini and Gemma 7B.

5.1 Explainability

With respect to explainability, the simplest approach with a non-negative linear model is to examine the relative weights, normalized to 1 to make them more intuitively interpretable. This tells us what percentage of the final score is due to each of subtrait scores from the LLM. As we have 6 prompts and 3 models per prompt, we present the means and standard deviations of the relative weights in Table 6. The relative weights for all the individual models appear in Table 8 in the Appendix.

The ability to inspect the relative weights is what we regard as the main benefit of this approach. We can see, for example, that the GPT and Llama models place more weight on establishing context, the use of narrative techniques, and the fluidity of transitions; and less weight on linguistic descriptiveness and the quality of the conclusion. We hypothesize that this is related to the models' superior performance, and we highlight that the weights can be subject to examination by a subject matter expert (SME) in writing who may not have a great deal of expertise in machine learning.

The weights of the Gemma 7B model are particularly illustrative. The model weights linguistic descriptiveness very heavily at almost half of the score, and the conclusion quality is hardly part of the model at all. This might raise validity concerns for an SME scrutinizing the model and could serve as an early and easily interpretable cautionary signal before moving forward with such a model.

5.2 LLM comparison

Gemma 7B's lack of parity with the other two models in terms of the predictiveness of its subtrait scores is striking. Furthermore, we note that the standard deviations of the per-prompt models are double those of GPT-4o-Mini and Llama 3.1 8B, as shown in Table 6. This suggests that the subtrait scores produced from Gemma 7B are less stable than those of the other two models.

	Train			Test		
	GPT-4o Mini	Llama3.1 8B	Gemma 7B	GPT-4o Mini	Llama3.1 8B	Gemma 7B
Prompt 1	0.802	0.786	0.629	0.843	0.799	0.654
Prompt 2	0.770	0.742	0.616	0.761	0.718	0.603
Prompt 3	0.815	0.781	0.468	0.852	0.791	0.467
Prompt 4	0.794	0.788	0.411	0.815	0.801	0.417
Prompt 5	0.767	0.734	0.593	0.775	0.782	0.617
Prompt 6	0.830	0.815	0.618	0.831	0.804	0.712

Table 5: Quadratic weighted kappa for models trained on subtrait features from 3 LLMs.

	W.8.3.A Context	W.8.3.B Narrative Tech.	W.8.3.C Transitions	W.8.3.D Descriptiveness	W.8.3.E Conclusion
GPT-4o Mini	0.24 (0.05)	0.34 (0.05)	0.17 (0.06)	0.08 (0.06)	0.16 (0.04)
Llama3.1 8B	0.13 (0.07)	0.30 (0.02)	0.30 (0.05)	0.15 (0.03)	0.13 (0.04)
Gemma 7B	0.19 (0.12)	0.17 (0.10)	0.16 (0.11)	0.47 (0.13)	0.01 (0.01)

Table 6: Mean and standard deviation of model weights across 6 prompts per LLM, where the standard deviations are in parenthesis. The weights are normalized to 1 within each model for intuitive interpretation.

	GPT vs. Llama	GPT vs. Gemma	Gemma vs. Llama
W.8.3.A	0.711	0.545	0.550
W.8.3.B	0.758	0.575	0.603
W.8.3.C	0.678	0.408	0.489
W.8.3.D	0.688	0.573	0.562
W.8.3.E	0.679	0.428	0.462

Table 7: Pearson correlations for pairwise comparisons of subtrait scores produced by the 3 LLMs.

When we examine the Pearson correlations of the subtrait scores for each pairing of models in Table 7, we find that the subtrait scores from GPT-4o-Mini are highly correlated with the subtrait scores from Llama 3.1 8B, and less correlated with the subtrait scores from less performant Gemma 7B. We take this as evidence that GPT-4o-Mini and Llama 3.1 8B are assessing the same or similar subtraits, whereas Gemma 7B is responding to the rubrics and responses in a significantly different way.

6 Discussion

The successful models have a QWK that hovers around 0.80, roughly 10 points under the operationally deployed models. We find this an encouraging result, particularly given the high level of performance optimizations that go into achieving the maximum possible QWKs in operational scoring. The optimizations make the model more complex, and therefore less interpretable. We are able to achieve a viable, transparently explainable model, with just 5 features. Furthermore, a comparison of the train and test QWKs in Table 5 indicates that there is no overfitting.

The simplicity and transparency of the models allows for a straight-forward look into the subtraits. While we do not have ground-truth annotations for the subtrait scores, we are reassured by how well the weights from the most successful models match with our intuitions with respect to construct relevance. Both GPT-4o-Mini and Llama 3.1 8B weighted the subtrait feature `W.8.3.B` as highest. This is the standard element that, to us, reflects the heart of the narrative genre: "use narrative techniques, such as dialogue, pacing, description, and reflection, to develop experiences, events. and/or characters." A lower-weighted feature in these models was `W.8.3.D`, "use precise words and phrases, relevant descriptive details, and sensory language…"

According to the most successful models, it is of highest importance that readers are able to understand who the characters are, what happens to them, and what they do; and it is of lesser importance how vividly these things are described. This aligns with our intuitions: a vivid description matters less if readers cannot understand what happened.

In contrast, the least successful model weights `W.8.3.D` the highest, at almost half of the score. Furthermore, it hardly assesses the quality of the conclusion at all.

It may happen that our intuitions about the relative importance of the standard elements of the narrative genre are underinformed, but we would like to emphasize that being able to inspect and reflect upon the weights of a simple linear model gives us useful tools to scrutinize the LLM output. In the case at hand, the tools would lead us to abandon Gemma 7B, given that we have evidence to suggest that the outputs from GPT-4o-Mini and Llama 3.1 8B are more trustworthy with respect to construct relevance.

A final note is that the weights themselves can be useful in downstream tasks. For example, feedback systems can use the weights in algorithms to select the most valuable feedback to display to the student. One could also build models of the scores assigned by individual raters and use the weights to understand rater behavior, revealing, for example, that one rater places more emphasis on descriptiveness whereas another places more emphasis on the quality of the conclusion.

7 Conclusion and future work

We present a method to build, evaluate, and inspect simple explainable models on the basis of subtrait-scores from LLMs, where the rubric criteria for the subtrait scores are derived from the Common Core Standard elements for 8th grade narrative writing. The models using subtrait scores from GPT-4o-Mini and Llama 3.1 8B are able to predict the holistic scores with a QWK of approximately 0.80, which we find particularly noteworthy because the standard-based rubrics were not what the humans used during scoring. We are encouraged that Llama 3.1 8B, an open source model, performs to within 3 points of the GPT model, and that its subtrait scores are highly correlated with GPT's.

The models we present are generally 10-points under the QWK for the operational models; nonetheless, at 0.80 QWK, we find that these simple, transparent, linear regression models with only 5 features may be operationally viable.

We find that the model weights from GPT-4o-Mini and Llama 3.1 8B are in alignment with our own intuitions about the narrative construct, whereas the weights from Gemma 7B are not. On the basis of this evidence, together with the raw model performance, we find that we can trust the LLM output of GPT-4o-Mini and Llama 3.1 8B much more than Gemma 7B. We take this example as an illustration of how this approach allows an inspection of the model by an SME in writing who may not have a great deal of expertise in machine learning, and as a means of understanding how the output of one LLM may differ from others overall.

We would like to better understand the impact of subtrait score accuracy on these regression models. Our previous study on subtraits found low to modest agreement between human ratings of subtrait scores and LLM-produced ones (Andrade-Lotero et al., 2025). As we did not have human-labeled narrative subtrait scores, we can not speak to the accuracy of the LLM-generated scores. In future work, we would like to work with subject matter experts to validate the accuracy and to understand if the resulting weights align with their expert judgment.

This LLM-plus-regression approach also provides a framework for not only for explaining automated scores, but human ones as well. By modeling individual raters, we can glean insights into sources of rater disagreement. We save exploration of this topic for future research.

Lastly, we are encouraged by the possibilities this framework presents for operationalizing AES models in both high stakes and formative settings. The direct interpretability of features allows for improved monitoring and transparency. As reliability of LLM subtrait assessment improves, this approach opens opportunities to enable scoring for more complex constructs and writing behaviors.

8 Limitations

The first limitation to note is that we neither have ground-truth annotations of the subtrait scores nor an in-depth understanding of the subtrait scores produced by the LLM. While the high QWK and alignment of the model weights with our intuitions is highly suggestive, we have not provided strong empirical evidence with respect to the degree to which the LLM is accurately applying the rubric. Related to this, we make a large assumption that our features are indeed construct-relevant and have not explored the impact of including distractor features.

Another important limitation is that the data is from one genre and one grade level. It is not known how well our results generalize to other grades and

genres. Similarly, the human agreement is unusually high, suggesting a strong clear signal in the response text itself. It may be that LLMs are less able to assess responses of a more ambiguous nature.

The final limitation is that there were small differences in the datasets used to extract subtrait scores to train and evaluate the 3 LLMs because the sets of responses that the LLMs were able to process successfully were not identical (shown in Table 3). We believe that our sample size is large enough to overcome this limitation; but nonetheless, the cleanest experiment would make comparisons using data sets that are absolutely identical.

9 Acknowledgments

We would like to thank Scott Hellman, Josh Southerland, and Melanie Sharif for their suggestions on the analysis of the results and Scott Hellman for comments on the final draft.

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)*, pages 715–725, Berlin, Germany. Association for Computational Linguistics.

Alejandro Andrade-Lotero, Lee Becker Becker, Joshua Southerland, and Scott Hellman. 2025. Toward subtrait-level model explainability in automated writing evaluation. Paper presented at the 2025 Annual Meeting of the National Council on Measurement in Education (NCME).

Anthropic. 2024. Claude 3.5 haiku: Our fastest model, delivering advanced coding, tool use, and reasoning at an accessible price. https://www.anthropic.com/claude/haiku.

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

Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213.

Common Core State Standards Initiative. 2010. Common core state standards for english language arts & literacy in history/social studies, science, and technical subjects. https://corestandards.org/wp-content/uploads/2023/09/ELA_Standards1.pdf.

Google DeepMind. 2024. Gemma: Open models built from the same research and technology used to create gemini. https://ai.google.dev/gemma.

Scott Deerwester, Susan T Dumais, Thomas K Landauer, George W Furnas, and Robert A Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

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.

Ahmed M. Elmassry, Nazar Zaki, Negmeldin Alsheikh, and Mohammed Mediani. 2025. A systematic review of pretrained models in automated essay scoring. *IEEE Access*, 13:121902–121917.

Sohaila Eltanbouly, Salam Albatarni, and Tamer Elsayed. 2025. Trates: Trait-specific rubric-assisted cross-prompt essay scoring. In *The 63rd Annual Meeting of the Association for Computational Linguistics*, Vienna, Austria. Association for Computational Linguistics.

Peter W Foltz, Darrell Laham, and Thomas K Landauer. 1999. The intelligent essay assessor: Applications to educational technology. *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning*, 1(2):939–944.

Arthur C Graesser, Danielle S McNamara, Max M Louwerse, and Zhiqiang Cai. 2004. Coh-metrix: Analysis of text on cohesion and language. *Behavior research methods, instruments, & computers*, 36(2):193–202.

Scott Hellman, William Murray, Adam Wiemerslage, Mark Rosenstein, Peter Foltz, Lee Becker, and Marcia Derr. 2020. Multiple instance learning for content feedback localization without annotation. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 30–40, Seattle, WA, USA → Online. Association for Computational Linguistics.

Jie Huang and Kevin Chen-Chuan Chang. 2023. Towards reasoning in large language models: A survey. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1049–1065, Toronto, Canada. Association for Computational Linguistics.

Jiazheng Li, Lin Gui, Yuxiang Zhou, David West, Cesare Aloisi, and Yulan He. 2023. Distilling ChatGPT for explainable automated student answer assessment. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6007–6026, Singapore. 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.

Meta. 2024. Llama 3.1 8b. https://www.llama.com/models/llama-3/.

Tomáš Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Tomáš Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013b. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1532–1543. Association for Computational Linguistics.

OpenAI. 2024. Gpt-4o mini: Advancing cost-efficient intelligence. https://openai.com/blog/gpt-4o-mini-advancing-cost-efficient-intelligence/.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Marco Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 97–101, San Diego, California. Association for Computational Linguistics.

Swapna Somasundaran, Michael Flor, Martin Chodorow, Hillary Molloy, Binod Gyawali, and Laura McCulla. 2018. Towards evaluating narrative quality in student writing. *Transactions of the Association for Computational Linguistics*, 6:91–106.

Maja Stahl, Leon Biermann, Andreas Nehring, and Henning Wachsmuth. 2024. Exploring LLM prompting strategies for joint essay scoring and feedback generation. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 283–298, Mexico City, Mexico. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Shixiao Wang. 2024. Deberta with hats makes automated essay scoring system better. *Applied and Computational Engineering*, 52:45–54.

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

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.

Appendix

```
You are an expert in academic standards with deep knowledge of assessment and
rubric design.  You will be given a standard_id along with information about the
standard as well as parameters for the output like min_score and max_score.  Your
job is to interpret the standard and provide a set of criteria for each score
point that will help to assess the level of student writing with respect to the
standard.

Please provide the criteria in a clear and concise manner, ensuring that
they are:
1. specific to the standard
2. relevant to the grade level of the students.
3. appropriate for the type of writing being assessed
4. written in a way that guides an LLM to evaluate a student's response in a
   reliable and consistent manner.
5. (This is very important) the criteria are written to ensure non- overlapping
    behaviors to encourage the LLM to use the full score range.
```

Figure 1: System instructions that were used to create subtrait rubrics from the narrative standard elements. Additional instructions had to do with JSON formatting; definitions of the keys and values; specifications for the min and max score points, grade, and genre; and encouragements to use the full score range.

		W.8.3.A Context	W.8.3.B Narrative Tech.	W.8.3.C Transitions	W.8.3.D Descriptiveness	W.8.3.E Conclusion
GPT-4o Mini	Prompt 1	0.231	0.285	0.157	0.129	0.198
	Prompt 2	0.276	0.381	0.084	0.096	0.163
	Prompt 3	0.172	0.304	0.243	0.151	0.131
	Prompt 4	0.323	0.310	0.112	0.076	0.178
	Prompt 5	0.257	0.418	0.214	**0.000**	0.111
	Prompt 6	0.199	0.371	0.215	0.022	0.193
Llama3.1 8B	Prompt 1	0.075	0.290	0.302	0.183	0.150
	Prompt 2	0.193	0.272	0.248	0.152	0.135
	Prompt 3	0.127	0.275	0.340	0.169	0.090
	Prompt 4	0.015	0.325	0.347	0.155	0.159
	Prompt 5	0.217	0.326	0.236	0.134	0.088
	Prompt 6	0.130	0.300	0.305	0.085	0.180
Gemma 7B	Prompt 1	0.228	0.068	0.172	0.532	**0.000**
	Prompt 2	**0.000**	0.257	0.218	0.525	**0.000**
	Prompt 3	0.124	0.107	0.311	0.458	**0.000**
	Prompt 4	0.181	0.131	0.025	0.663	**0.000**
	Prompt 5	0.336	0.326	0.033	0.304	**0.000**
	Prompt 6	0.279	0.138	0.202	0.347	0.033

Table 8: Relative weights of LLM-generated subtrait features in regression models. We observe a fair amount of consistency in the weights across prompts within models, and similar weights between GPT-4o-Mini and Llama3.1 8B. We also observe that Gemma 7B's assessment of the conclusion did not offer a unique contribution to the prediction of the score.

```
{
  "4": {
    "criteria": [
      "Exceeds Expectations",
      "Masterfully establishes a complex and engaging narrative context",
      "Provides a highly sophisticated and nuanced introduction of narrator and/or
      characters",
      "Creates an exceptionally clear and compelling point of view",
      "Organizes events with remarkable logical flow and natural progression",
      "Demonstrates advanced narrative techniques that immediately capture the
      reader's interest"
    ]
  },
  "3": {
    "criteria": [
      "Meets Expectations",
      "Effectively establishes a clear narrative context",
      "Introduces narrator and/or characters with sufficient detail",
      "Presents a distinct and appropriate point of view",
      "Organizes events in a logical and coherent sequence",
      "Provides a solid foundation for the narrative that guides the reader's
      understanding"
    ]
  },
  "2": {
    "criteria": [
      "Approaching Expectations",
      "Provides a basic narrative context with some gaps or lack of clarity",
      "Partially introduces narrator and/or characters with minimal details",
      "Demonstrates an inconsistent or somewhat unclear point of view",
      "Attempts to organize events, but the sequence may have some minor logical
      inconsistencies",
      "Shows an emerging understanding of narrative introduction"
    ]
  },
  "1": {
    "criteria": [
      "Below Expectations",
      "Offers a minimal or confusing narrative context",
      "Provides little to no introduction of narrator and/or characters",
      "Lacks a clear or coherent point of view",
      "Events are poorly organized or difficult to follow",
      "Struggles to establish a meaningful narrative foundation"
    ]
  },
  "0": {
    "criteria": [
      "Insufficient",
      "No discernible narrative context",
      "No introduction of narrator or characters",
      "No identifiable point of view",
      "No coherent event sequence",
      "Fails to create any meaningful narrative structure"
    ]
  }
}
```

Figure 2: Example rubric for the `W.8.3.A` standard element subtrait.

```
Assess the student's ability to effectively introduce a narrative by
establishing a clear context, point of view, and characters. Evaluate how well
the writer sets up the story and creates a logical, natural progression of
events. Consider the sophistication of the narrative setup, the clarity of
the introduction, and the coherence of the event sequence.

Return the chosen score_point as well as up to three small excerpts from the
response as evidence, without any modification or additional reasoning. The
excerpts should only be subsets of the original response text. It is okay to
return fewer than the max amount of excerpts, if some aren't good relative
to the others. Also, don't return the same excerpt twice. If the student got
the highest score_point, you should provide feedback summarizing what they
did well. If they did not get the highest score_point, you should give
feedback with a high level suggestion on how to improve. Feedback should be
worded to communicate with a student in grade 8 and limited to the specific
criteria in the rubric. You should not mix in unrelated analyses Return only
JSON containing the score_point, optional feedback, and each optional excerpt.

You should specifically evaluate the response based only on the following
scorepoint criteria:
//{Rubric appears here}

Score and provide feedback for this response:
//{Response text appears here.}
```

Figure 3: Prompt that was used to elicit subtrait assessments from 3 LLMs.