Evaluating Transformer Model Family Representations Through Automated Essay Scoring

Akchay Ozten, Rodrigo Wilkens
University of Exeter
{ao478, r.wilkens}@exeter.ac.uk

Abstract

Large Language Models have become central to Automated Essay Scoring (AES), typically through fine-tuned transformer encoders or prompt-based applications of decoder models. However, the representational capacity of decoder models as frozen embedding extractors remains largely unexplored. In this paper, we present a controlled comparison between encoder and decoder transformer embeddings for prompt-agnostic AES. Using regression models, we evaluate frozen representations across two English datasets. We analyzed scaling effects and the impact of integrating explicit linguistic features in hybrid configurations. Our results show that decoder embeddings consistently outperform encoder embeddings in embedding-only settings, with gains generalizing across holistic essay scoring and proficiency prediction. Scaling effects are modest, and hybrid models that combine contextual embeddings with linguistic features yield further improvements. Notably, frozen decoder embeddings achieve performance competitive with a fine-tuned BERT. These findings highlight the importance of representation-level properties in essay scoring.

Keywords: Automated Essay Scoring, Autoregressive Models, Transformer Representations, CEFR Classification, Hybrid NLP Models

1. Introduction

Automated Essay Scoring (AES) is the use of machine learning (ML) and natural language processing (NLP) techniques to evaluate and grade written essays automatically (Taghipour and Ng, 2016; Shermis and Burstein, 2013). AES systems provide consistent, objective, and scalable assessments, reducing the workload of human graders while potentially offering rapid feedback to students (Klebanov and Madnani, 2022). Essays are typically written in response to specific prompts, and scoring requires assessing multiple dimensions of writing quality, including coherence, fluency and grammatical accuracy.

Early AES systems relied on manually engineered features such as lexical diversity, syntactic complexity, and surface-level readability metrics (Page, 1966; Attali and Burstein, 2006; Foltz et al., 1999; Zesch et al., 2015). While these approaches achieved moderate correlations with human raters, they struggled to capture deeper semantic and discourse-level properties and were vulnerable to superficial manipulation (Perelman, 2014; Shermis and Burstein, 2013). Subsequent neural approaches based on CNNs and RNNs reduced reliance on feature engineering and improved representation learning (Taghipour and Ng, 2016; Wang et al., 2018), yet they remained limited in modeling longer essays.

The introduction of the transformer architecture transformed Natural Language Processing (NLP) by enabling efficient modeling of long-range dependencies through self-attention (Vaswani et al., 2017). Encoder-based models have since become dominant in AES research. Fine-tuning pre-trained encoders on scoring datasets has yielded strong performance (Rodriguez et al., 2019; Mayfield and Black, 2020), and hybrid approaches combining contextual embeddings with hand-crafted features have further improved results (Dasgupta et al., 2018; Uto et al., 2020). Such hybrid architectures leverage both deep contextual representations and explicit linguistic signals.

Decoder-based generative models have gained attention in AES through prompt-based evaluation strategies (Mizumoto and Eguchi, 2023; Yancey et al., 2023; Stahl et al., 2024). These models generate scores conditioned on rubrics or other prompted information. While prompt-based methods have shown promising performance, they are limited to generative behavior.

Despite the growing influence of generative models, their internal embeddings have been largely overlooked in NLP and AES research. In particular, there has been limited investigation into whether frozen decoder representations, used independently of prompting, encode signals relevant to writing quality. This study investigates whether embeddings extracted from decoder models, when used as fixed representations or in a hybrid framework, can enhance the performance of AES systems. To evaluate robustness, we conduct a prompt-agnostic comparison across two assessment settings. We assess whether any observed representational advantages generalize across related evaluation tasks. The study is guided by the following research questions:

RQ1 Do frozen decoder embeddings improve per-

Proceedings of the Joint Workshop on Readability and Text Simplification (READIxTSAR) @ LREC 2026, pages 142–150
11 May 2026.

formance over encoder embeddings in prompt-agnostic AES?

RQ2 How does model size influence the effectiveness of decoder representations for scoring?

RQ3 Do linguistic features remain complementary when combined with decoder embeddings?

The main contribution of this study is a systematic comparison of encoder and decoder embeddings as frozen representations for AES and related assessment tasks. Our results show that decoder embeddings consistently outperform encoder embeddings in this setting and that these gains generalize across datasets. Furthermore, we demonstrate that hand-crafted linguistic features provide complementary information when integrated with transformer-based representations.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 details the methodology. Section 4 presents the results, followed by discussion in Section 5 and concluding remarks in Section 6.

2. Related Work

2.1. Transformer-Based Approaches

Transformer architectures have become central to Automated Essay Scoring due to their ability to model long-range dependencies and contextual interactions (Vaswani et al., 2017). Encoder-based models, such as BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), XLNet (Yang et al., 2019), and DeBERTa (He et al., 2020, 2021), have been widely adopted in AES (Devlin et al., 2018; Liu et al., 2019; He et al., 2020, 2021). Fine-tuning pre-trained encoders for essay scoring has consistently demonstrated improvements over earlier neural approaches (Rodriguez et al., 2019; Mayfield and Black, 2020). In these systems, scoring is formulated as a supervised regression or classification task, with a task-specific head trained on top of contextualized document representations.

A second line of work integrates encoder representations within hybrid architectures that concatenate contextual embeddings with linguistic features (Dasgupta et al., 2018; Uto et al., 2020). These features typically capture lexical richness, syntactic complexity or discourse-level statistics. Empirical results suggest that hybrid systems often outperform purely neural models. However, both fine-tuning and hybrid approaches have predominantly relied on encoder-derived document representations, typically extracted via pooled outputs (e.g., the *[CLS]* token).

2.2. Prompt-Based and Decoder Architecture

From a different perspective, large autoregressive decoder models, such as GPT (Radford et al., 2018), Mistral (Jiang et al., 2023), Qwen (Bai et al., 2023), and DeepSeek (Bi et al., 2024), have achieved strong performance in generative language tasks. Their flexibility through prompting has led to growing interest in applying them to assessment settings (Liu et al., 2023). Mizumoto and Eguchi (2023) evaluated GPT-3 for rubric-based essay scoring, Yancey et al. (2023) examined GPT-3.5 and GPT-4, and Stahl et al. (2024) investigated prompting strategies using Mistral 7B. These works report performance competitive with traditional supervised AES systems on datasets such as TOEFL11, demonstrating the potential of generative inference for automated scoring.

Methodologically, prompt-based AES differs fundamentally from encoder fine-tuning. Rather than training a regression head over fixed document embeddings, decoder models are evaluated in generative inference mode, conditioned on prompts, scoring rubrics, or example essays. Consequently, scoring performance may depend on prompt formulation and decoding strategies (Liu et al., 2023). This makes it difficult to isolate intrinsic representational capacity from inference-time adaptation effects.

Encoder and decoder architectures are also trained under distinct pre-training objectives. Encoder models rely on masked language modeling (MLM) (Devlin et al., 2018), whereas decoder models are optimized using autoregressive next-token prediction (Radford et al., 2018; Brown et al., 2020). These objectives impose different structural constraints on learned representations and have been shown to induce distinct inductive biases and internal geometries (Rogers et al., 2020; Belinkov and Glass, 2019).

Despite the increasing use of decoder models in prompt-based AES, prior research has largely examined them in generative inference settings rather than as sources of document-level embeddings for supervised scoring. To our knowledge, the use of decoder-derived representations as fixed embedding features in AES remains largely unexplored.

3. Methodology

To address the research questions outlined in Section 1, we adopt a controlled comparative framework in which the only systematically varied component is the source of transformer-based representations. Specifically, we compare embeddings extracted from encoder-based and decoder-based transformer models under identical downstream

conditions.

3.1. Corpora

This study employs two English datasets representing related but distinct assessment settings. Corpus statistics are summarized in Table 1.

The first dataset is the AES2 benchmark (Crossley et al.), released as part of the Learning Agency Lab Automated Essay Scoring 2.0 competition. It contains 17,307 essays scored on a 1-6 ordinal scale. Unlike the earlier ASAP dataset[1], AES2 does not provide explicit prompt labels, encouraging prompt-agnostic evaluation and broader generalization.

The second corpus is the Common European Framework of Reference for Languages (CEFR) European Language Grid (ELG) dataset in English (Breuker, 2023). It is available as part of the UniversalCEFR project (Imperial et al., 2025) in HuggingFace[2]. It contains 712 essays labeled according to CEFR proficiency levels. To mitigate class imbalance, adjacent sublevels (e.g., A1-/A1+, A2-/A2+) are merged, resulting in six levels aligned with the AES scoring scale. While AES2 evaluates holistic essay quality, the ELG dataset captures the result of a descriptor-based method.

Level	#Essays	#Tokens
A1	1,252	274 (110)
A2	4,723	265 (97)
B1	6,280	361 (102)
B2	3,926	483 (109)
C1	970	636 (135)
C2	156	778 (162)

(a) AES2

Level	#Essays	#Tokens
A1	25	255 (14)
A2	141	216 (71)
B1	206	282 (67)
B2	174	423 (181)
C1	97	640 (324)
C2	69	775 (239)

(b) ELG

Table 1: Distribution of essays and average token counts (mean with standard deviation in parentheses) across CEFR levels

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

[2] https://huggingface.co/datasets/UniversalCEFR/elg_cefr_en

3.2. Representation Extraction

We evaluate multiple encoder and decoder transformer models obtained from the HuggingFace library (Wolf et al., 2019). Encoder models include BERT and its variants (Devlin et al., 2018; He et al., 2020, 2021), while decoder models include Mistral (Jiang et al., 2023), Qwen (Bai et al., 2023), LLaMA (Llama Team, AI @ Meta, 2024), and DeepSeek (Bi et al., 2024). These models vary in parameter size, enabling analysis of scaling effects.

Embeddings are derived from the final hidden layer of each transformer model. For encoder models, we extract the embedding corresponding to the *[CLS]* token, which is conventionally used as a global sequence representation in classification and regression tasks (Devlin et al., 2018).[3] This representation serves as a compact essay-level embedding.

Decoder models do not employ a dedicated classification token. To obtain a fixed-size representation, we apply mean pooling over the hidden states of all tokens in the final layer. This strategy yields a length-invariant summary while preserving contextual information learned through autoregressive training. Although this approach produces a representation similar to that obtained for the decoders, one might argue that there is an asymmetry in the comparisons. To address this perspective, we also evaluated the BERT and DeBERTa models using mean pooling.

This study employed 15 transformer models, 5 encoders and 10 decoders, from Hugging Face to extract embeddings. These are small to medium models. Table 2 provides the details including the size of the embedding vectors.

All transformer models remain frozen throughout the experiments. No task-specific fine-tuning is performed in the embedding-based configurations.

3.3. Linguistic Features

To evaluate complementarity between contextual embeddings and explicit linguistic information (RQ3), we extract a set of features using the *TextDescriptives* library (Hansen et al., 2023). These features include descriptive statistics, lexical richness indicators, syntactic complexity measures, dependency distance metrics, and coherence-related indicators. Such features have been shown to correlate with writing quality and proficiency (Zesch et al., 2015; Shermis and Burstein, 2013). When combined with encoder/decoder embeddings, these features form hybrid representations that integrate

[3] The transformers library was used to extract embeddings from both encoder and decoder models. These embeddings are derived from the last hidden state of the model outputs and capture contextualized token representations.

Model	Hugging Face model	Transformer type	Vector size
BERT	google-bert/bert-base-uncased	Encoder	768
BERT Large	google-bert/bert-large-uncased	Encoder	1,024
ModernBERT	answerdotai/ModernBERT-base	Encoder	768
DeBERTa	microsoft/deberta-v3-base	Encoder	768
DeBERTa Large	microsoft/deberta-v3-large	Encoder	1,024
Mistral 7B	mistralai/Mistral-7B-Instruct-v0.3	Decoder	4,096
QWEN3 0.6B	Qwen/Qwen3-0.6B-Base	Decoder	1,024
QWEN3 1.7B	Qwen/Qwen3-1.7B-Base	Decoder	2,048
QWEN3 4B	Qwen/Qwen3-4B-Base	Decoder	2,560
QWEN3 8B	Qwen/Qwen3-8B-Base	Decoder	4,096
QWEN3 14B	Qwen/Qwen3-14B-Base	Decoder	5,120
Llama 1B	meta-llama/Llama-3.2-1B	Decoder	2,048
Llama 3B	meta-llama/Llama-3.2-3B	Decoder	3,072
Deepseek 1.5B	deepseek-ai/DeepSeek-R1-Distill-Qwen-1.5B	Decoder	1536
Deepseek 7B	deepseek-ai/DeepSeek-R1-Distill-Qwen-7B	Decoder	3584

Table 2: HuggingFace transformer models used

explicit structural signals with deep contextual encodings.

3.4. Scoring Framework

We adopt a regression-based formulation of essay scoring, enabling the prediction of fine-grained values within the ordinal scale. Treating essay scores as continuous targets avoids rigid class boundaries while allowing ordinal-sensitive evaluation through appropriate metrics.

For all embedding-based experiments, we employ LightGBM and SVM as the downstream regressor.[4] Transformer embeddings, optionally concatenated with linguistic features, are provided as input to the model. LightGBM is a histogram-based gradient boosting framework optimized for high-dimensional feature spaces, making it well suited for transformer embeddings ranging from hundreds to several thousand dimensions. In contrast, a linear SVM serves as a strong linear baseline, effectively assessing the extent to which the frozen embedding space supports linear prediction for the target task.

To address RQ1-RQ3, we evaluate a configuration in which the frozen essay-level embedding extracted from each encoder or decoder model constitutes the sole input to the LightGBM regressor. We extend the embedding-only setup by concatenating the linguistic feature vector to the transformer embedding. The combined representation is then provided as input to the same LightGBM regressor.

[4]For hyperparameter tuning, n_estimators was set to 4,000-6,000 for the AES2 dataset and 1,000-3,000 for the ELG dataset for the LightGBM, and C between 0.01 and 1 for the linear SVM.

3.5. Fine-Tuned Encoder Baseline

In addition to frozen representations, we include a fine-tuned BERT model as a supervised baseline (Rodriguez et al., 2019; Mayfield and Black, 2020). In this configuration, a regression head is added on top of the encoder, and the model is optimized end-to-end for the scoring task (i.e., all layers weights are updated). Training, validation, and evaluation follow the same cross-validation protocol used in the embedding-based experiments. The fine-tuning process was run using the following parameters: learning rate of 0.00002, per device train batch size of 8, per device evaluation batch size of 8, number of train epochs of 5, weight decay of 0.01 and AdamW optimizer.

3.6. Evaluation Protocol

Performance is evaluated using 10-fold cross-validation.[5] The primary metric is Quadratic Weighted Kappa (QWK), which measures agreement between predicted and true scores while penalizing larger discrepancies, and is prioritized as it is widely used in AES benchmarking (Taghipour and Ng, 2016). Accuracy and macro-F1 are also reported, computed after discretizing regression outputs to the nearest valid CEFR score level. Statistical significance between models is assessed using the Wilcoxon signed-rank test across cross-validation folds.

[5]In each fold, we split in 80/10/10.

4. Results

To compare frozen encoder and decoder embeddings under identical downstream conditions, we first evaluate the embedding-only configuration. The results are reported in Table 4.

Across both datasets, decoder-based embeddings consistently outperform encoder-based embeddings. On AES2, all decoder models achieve higher QWK and F1 scores than encoder models under identical regression conditions (Table 4). Statistical testing confirms that the strongest decoder variants significantly outperform the best encoder baselines (Wilcoxon signed-rank test, $p < 0.05$). A similar pattern is observed on the ELG dataset, indicating that the performance advantage of decoder embeddings is not restricted to a single corpus.

First, we compared the performance of the machine learning models (LightGBM vs. SVM), as shown in Tables 4 and 5, respectively. This analysis shows that LightGBM consistently yields better results. We further observe that CLS-based representation consistently outperforms mean pooling across both SVM and LightGBM models (Table 3). This suggests that the way in which document-level representations are constructed has a measurable impact on downstream performance, independent of the regression model. This finding reinforces the importance of representation-level design choices in AES, indicating that not only the model family (encoder vs decoder), but also the aggregation strategy, influences the quality of the resulting embedding space.

Corpus	Model	F1	Accuracy	QWK
AES2	BERT	0.53	0.56	0.69
	DeBERTa	0.58	0.61	0.75
ELG	BERT	0.58	0.59	0.84
	DeBERTa	0.46	0.47	0.76

(a) LightGBM

Corpus	Model	F1	Accuracy	QWK
AES2	BERT	0.55	0.59	0.70
	DeBERTa	0.53	0.57	0.67
ELG	BERT	0.62	0.63	0.85
	DeBERTa	0.52	0.53	0.79

(b) SVM

Table 3: BERT and DeBERTa performance using mean pooling

To evaluate robustness, we examine the consistency of relative model rankings in Table 4. Despite differences in dataset size and scoring criteria, decoder embeddings maintain their advantage across both AES2 and ELG.

We next investigate the relationship between decoder model size and performance (Table 4). While larger models tend to achieve marginally higher scores on AES2, improvements are not consistently statistically significant. On ELG, no monotonic trend is observed. These findings indicate that representational suitability for AES does not scale linearly with parameter count over the evaluated range.

We now compare embedding-only and hybrid configurations. The hybrid results are presented in Table 6. Across all models and both datasets, adding linguistic features yields consistent positive gains. This confirms that contextual embeddings do not fully subsume explicit structural information and that hybrid modeling remains beneficial. Decoder-based hybrid models remain superior to encoder-based hybrids, indicating that representational advantages persist after feature integration.

Finally, we compare frozen embeddings with a fine-tuned BERT (Table 4). Although fine-tuned models achieves competitive results, the strongest frozen decoder embeddings match or exceed its performance. This indicates that representational differences alone can rival supervised adaptation. Moreover, the best fine-tuned model depends on the corpus, likely due to the amount of training data.

In summary, the strongest encoder model (frozen DeBERTa + mean pooling, QWK = 0.74 on AES2; 0.85 on ELG) is consistently outperformed by multiple frozen decoder variants, with Mistral 7B achieving QWK = 0.79 on AES2 and QWEN3/Llama variants reaching up to 0.89 on ELG. The model fine-tuned BERT baseline (DeBERTa QWK = 0.80 on AES2; BERT QWK = 0.85 on ELG) improves over standard encoder embeddings but remains comparable to, and in some cases below, the best frozen decoder models. Overall, the ranking of best-performing approaches is consistent across datasets: frozen decoder embeddings $\geq$ fine-tuned encoder $\geq$ frozen encoder. These results indicate that representation-level differences alone are sufficient to match or exceed supervised adaptation under controlled regression conditions.

5. Discussion

This study evaluated encoder and decoder transformer embeddings for automated essay scoring under strictly controlled downstream conditions. Rather than reiterating performance differences, we discuss their implications for AES research and representation evaluation in NLP.

Recent AES research has converged on fine-tuned encoder architectures as the dominant paradigm (Rodriguez et al., 2019; Mayfield and Black, 2020). Decoder models, when considered, have largely been evaluated through prompt-based scoring frameworks (Mizumoto and Eguchi, 2023; Yancey et al., 2023; Stahl et al., 2024). This division has implicitly positioned encoders as the

Model	AES2			ELG		
	QWK	Accuracy	F1	QWK	Accuracy	F1
BERT	0.70	0.56	0.55	0.82	0.57	0.55
BERT Large	0.71	0.57	0.56	0.76	0.51	0.49
ModernBERT	0.67	0.54	0.53	0.75	0.53	0.51
DeBERTa	0.74	0.59	0.58	0.85	0.62	0.61
DeBERTa Large	0.74	0.59	0.57	0.83	0.60	0.59
Mistral 7B	**0.79**	**0.63**	**0.63**	0.88	0.67	0.66
QWEN3 0.6B	0.75	0.59	0.58	0.87	0.64	0.63
QWEN3 1.7B	0.77	0.60	0.60	0.87	0.65	0.64
QWEN3 4B	0.78	0.61	0.61	0.87	0.65	0.64
QWEN3 8B	0.78	0.61	0.61	0.88	0.67	0.66
QWEN3 14B	0.78	0.61	0.61	**0.89**	0.69	0.68
Llama 1B	0.76	0.59	0.59	0.87	0.65	0.64
Llama 3B	0.77	0.60	0.60	**0.89**	**0.70**	**0.69**
Deepseek 1.5B	0.76	0.60	0.59	0.83	0.59	0.57
Deepseek 7B	0.78	0.62	0.61	0.85	0.63	0.62
BERT-FT	0.77	0.61	0.61	0.85	0.64	0.61
DeBERTa-FT	0.80	0.66	0.65	0.74	0.50	0.43

Table 4: Performance on AES2 and ELG Using LightGBM with the CLS Token and Fine-Tuning

Model	AES2			ELG		
	QWK	Accuracy	F1	QWK	Accuracy	F1
BERT	0.68	0.58	0.54	0.82	0.59	0.58
DeBERTa	0.60	0.51	0.47	0.67	0.44	0.42
Mistral 7B	0.74	0.57	0.54	0.89	0.71	0.69
QWEN3 0.6B	0.73	0.58	0.55	0.87	0.67	0.66
QWEN3 1.7B	0.74	0.60	0.57	0.88	0.67	0.66
QWEN3 4B	0.75	0.60	0.57	0.88	0.68	0.67
QWEN3 8B	0.75	0.60	0.58	0.89	0.69	0.68
QWEN3 14B	0.76	0.60	0.58	0.89	0.71	0.69
Llama 1B	0.74	0.59	0.56	0.88	0.69	0.67
Llama 3B	0.75	0.60	0.57	0.88	0.70	0.69
Deepseek 1.5B	0.75	0.60	0.58	0.87	0.65	0.64
Deepseek 7B	0.76	0.61	0.58	0.87	0.68	0.66

Table 5: Embedding-only performance on AES2 and ELG (SVM)

appropriate representational backbone for scoring, and decoders as generative evaluators.

The present findings complicate that dichotomy. When embeddings are evaluated independently of prompting and fine-tuning, decoder representations consistently match or outperform encoder representations under identical conditions. More broadly, the results indicate that architectural preference in AES has not been systematically stress-tested under controlled representation-level comparisons. The contribution of this work is not to declare decoder superiority, but to demonstrate the usefulness of an alternative representation, which has received comparatively limited systematic evaluation in AES.

Beyond differences between model families, we observe that both representation construction and downstream modeling play an important role. Mean pooling consistently decrease performance over CLS-based representations across both SVM and LightGBM, indicating that the method used to derive document-level embeddings has a measurable impact on downstream effectiveness. Furthermore, while both regressors exhibit similar relative trends,

Model	AES2			ELG		
	QWK	Accuracy	F1	QWK	Accuracy	F1
HC + BERT	0.77 (0.01)	0.62 (0.01)	0.61 (0.01)	0.87 (0.03)	0.67 (0.05)	0.66 (0.05)
HC + BERT Large	0.77 (0.01)	0.62 (0.01)	0.60 (0.01)	0.86 (0.03)	0.64 (0.05)	0.63 (0.06)
HC + ModernBERT	0.76 (0.01)	0.61 (0.01)	0.60 (0.01)	0.86 (0.03)	0.65 (0.06)	0.64 (0.06)
HC + DeBERTa	0.77 (0.01)	0.62 (0.01)	0.61 (0.01)	0.89 (0.03)	0.69 (0.07)	0.68 (0.07)
HC + DeBERTa Large	0.77 (0.01)	0.61 (0.01)	0.60 (0.01)	0.87 (0.03)	0.65 (0.07)	0.64 (0.07)
HC + Mistral 7B	0.81 (0.01)	0.66 (0.01)	0.66 (0.01)	0.90 (0.02)	0.72 (0.05)	0.71 (0.05)
HC + QWEN3 0.6B	0.80 (0.01)	0.65 (0.02)	0.65 (0.02)	0.91 (0.02)	0.74 (0.04)	0.74 (0.04)
HC + QWEN3 1.7B	0.81 (0.01)	0.66 (0.01)	0.65 (0.01)	0.90 (0.02)	0.71 (0.06)	0.70 (0.06)
HC + QWEN3 4B	0.81 (0.01)	0.66 (0.01)	0.65 (0.02)	0.89 (0.02)	0.70 (0.04)	0.69 (0.04)
HC + QWEN3 8B	0.81 (0.01)	0.66 (0.01)	0.65 (0.01)	0.90 (0.02)	0.72 (0.03)	0.71 (0.04)
HC + QWEN3 14B	0.81 (0.01)	0.66 (0.01)	0.65 (0.02)	0.91 (0.02)	0.72 (0.03)	0.71 (0.04)
HC + Llama 1B	0.81 (0.01)	0.65 (0.01)	0.65 (0.01)	0.90 (0.02)	0.72 (0.04)	0.71 (0.04)
HC + Llama 3B	0.81 (0.01)	0.66 (0.01)	0.65 (0.01)	0.91 (0.02)	0.75 (0.04)	0.74 (0.04)
HC + Deepseek 1.5B	0.80 (0.01)	0.65 (0.01)	0.64 (0.01)	0.88 (0.03)	0.69 (0.03)	0.68 (0.04)
HC + Deepseek 7B	0.81 (0.01)	0.65 (0.01)	0.65 (0.02)	0.88 (0.02)	0.69 (0.06)	0.68 (0.06)

Table 6: Hybrid configuration (HC + embeddings). Mean and standard deviation across folds.

LightGBM consistently outperforms linear SVM, reflecting the presence of non-linear structure in the embedding space. Taken together, these findings reinforce that both representation extraction and downstream modeling choices influence how effectively transformer embeddings can be leveraged for AES.

The absence of consistent scaling effects contrasts with broader claims about generative model scaling (Kaplan et al., 2020). While scaling laws hold for language modeling objectives, their impact in the embedding space appears less straightforward. In our findings for AES, moderate-sized decoder models achieve competitive performance without clear monotonic gains from additional parameters. The findings suggest that scaling effects observed in generative language modeling do not necessarily translate into proportional gains in representation-level transfer for structured assessment tasks.

Hybrid AES architectures have consistently shown gains from integrating linguistic features (Dasgupta et al., 2018; Uto et al., 2020). While linguistic features yield consistent improvements across architectures, they do not alter the relative ranking between encoder and decoder backbones. The implication is not that linguistic features are dispensable, but that representation learning can be enriched with linguistic information to improve its representation in some contexts.

It is important to delimit the scope of inference. The study does not isolate the effect of training objective, architecture, or pre-training data composition. Encoder and decoder families differ along multiple axes, and the observed differences should be interpreted at the level of model families rather than single causal mechanisms. Consequently, the conclusions are bounded to English-language corpora and the evaluated regression task. Alternative downstream models or multilingual settings may alter relative performance.

6. Conclusion

This paper presented a controlled comparison between encoder-based and decoder-based transformer embeddings for Automated Essay Scoring in a prompt-agnostic setting. By evaluating frozen representations within a fixed regression framework across two corpora (AES2 and ELG), we isolated representation-level differences from adaptation effects.

Our findings answer the research questions as follows. First, decoder embeddings consistently match or outperform encoder embeddings under identical downstream conditions. Second, scaling effects within decoder families are modest and not systematically monotonic. Third, integrating linguistic features yields consistent gains but does not alter the relative ranking between encoder and decoder backbones. Finally, the strongest frozen decoder embeddings achieve performance competitive with a fine-tuned BERT baseline.

The main contribution of this work is a systematic, representation-focused evaluation of transformer model families for AES, disentangled from prompt-

ing and fine-tuning strategies. The results indicate that decoder-based embeddings constitute a viable and underexplored backbone for assessment tasks.

Future work should investigate multilingual settings, alternative downstream architectures, and controlled comparisons using matched model families to further clarify the sources of representational differences.

Limitations

While two datasets were used, both are English-language corpora. The generalizability of these findings to multilingual AES remains to be investigated. Also, the study focuses on frozen embeddings. Alternative downstream architectures may interact differently with transformer representations.

Plain Summary

Large Language Models are now widely used for automated essay scoring, usually either by fine-tuning encoder models or by using decoder models with prompts. However, it is still unclear how good decoder models are when used simply as fixed feature extractors (without prompting or training). In this paper, we compare encoder and decoder models in a controlled way, using them only to create essay representations. We test these representations with standard regression models on two English essay datasets. We also look at whether model size matters and whether adding linguistic features (like grammar or vocabulary measures) improves results. We find that decoder-based representations consistently perform better than encoder-based ones when used on their own, and this holds across different types of essay scoring tasks. Increasing model size only leads to small improvements. When we combine model representations with linguistic features, performance improves further. Importantly, the best decoder-based representations perform about as well as a fine-tuned BERT model, even without additional training. Overall, the results show that how models represent text (their embeddings) plays a key role in essay scoring, and that decoder models are a strong and underused option for this task.

Bibliographical References

Y. Attali and J. Burstein. 2006. Automated essay scoring with e-rater[r] v.2. *The Journal of Technology, Learning, and Assessment*, 4(3).

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, et al. 2024. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*.

M. Breuker. 2023. Cefr labelling and assessment services. In G. Rehm, editor, *European Language Grid*, Cognitive Technologies. Springer, Cham.

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.

Scott Crossley, Perpetual Baffour, Jules King, Lauryn Burleigh, Walter Reade, and Maggie Demkin. Learning agency lab-automated essay scoring 2.0, kaggle (2024). *URL https://kaggle. com/competitions/learning-agency-lab-automated-essay-scoring-2*.

T. Dasgupta, A. Naskar, L. Dey, and R. 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.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. 2018. Bert: pre-training of deep bidirectional transformers for language understanding.

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

L. Hansen, L. R. Olsen, and K. Enevoldsen. 2023. Textdescriptives: A python package for calculating a large variety of metrics from text.

P. He, J. Gao, and W. Chen. 2021. Debertav3: improving deberta using electra-style pre-training with gradient-disentangled embedding sharing.

P. He, X. Liu, J. Gao, and W. Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention.

J. M. Imperial, A. Barayan, R. Stodden, R. Wilkens, R. M. Sanchez, L. Gao, and H. T. Madabushi. 2025. Universalcefr: Enabling open multilingual research on language proficiency assessment. *arXiv preprint arXiv:2506.01419*.

A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed. 2023. Mistral 7b.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Beata Beigman Klebanov and Nitin Madnani. 2022. *Automated essay scoring*. Springer Nature.

P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. 2019. Roberta: a robustly optimized bert pretraining approach.

Llama Team, AI @ Meta. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

E. Mayfield and A. W. Black. 2020. Should you fine-tune bert for automated essay scoring? In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 151–162.

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

E. B. Page. 1966. The imminence of... grading essays by computer. *The Phi Delta Kappan*, 47(5):238–243.

L. Perelman. 2014. When "the state of the art" is counting words. *Assessing Writing*, 21:104–111.

A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever. 2018. Improving language understanding by generative pre-training.

P. U. Rodriguez, A. Jafari, and C. M. Ormerod. 2019. Language models and automated essay scoring.

A. Rogers, O. Kovaleva, and A. Rumshisky. 2020. A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

M. D. Shermis and J. Burstein. 2013. *Handbook of Automated Essay Evaluation: Current Applications and New Directions*, 1st edition. Routledge.

M. Stahl, L. Biermann, A. Nehring, and H. Wachsmuth. 2024. Exploring llm prompting strategies for joint essay scoring and feedback generation. *arXiv preprint arXiv:2404.15845*.

K. Taghipour and H. T. 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.

M. Uto, Y. Xie, and M. Ueno. 2020. Neural automated essay scoring incorporating handcrafted features. In *Proceedings of the 28th international conference on computational linguistics*, pages 6077–6088.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. 2017. Attention is all you need.

Y. Wang, Z. Wei, Y. Zhou, and X. J. Huang. 2018. Automatic essay scoring incorporating rating schema via reinforcement learning. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 791–797.

T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, and A. M. Rush. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

K. P. Yancey, G. Laflair, A. Verardi, and J. 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.

Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le. 2019. Xlnet: generalized autoregressive pretraining for language understanding.

T. Zesch, M. Wojatzki, and D. Scholten-Akoun. 2015. Task-independent features for automated essay grading. In *Proceedings of the tenth workshop on innovative use of NLP for building educational applications*, pages 224–232.