Long context Automated Essay Scoring with Language Models

Christopher Ormerod
Cambium Assessment Inc.
christopher.ormerod@cambiumassessment.com

Gitit Kehat
Cambium Assessment Inc.
gitit.kehat@cambiumassessment.com

Abstract

Transformer-based language models are architecturally constrained to process text of a fixed maximum length. Essays written by higher-grade students frequently exceed the maximum allowed length for many popular open-source models. A common approach to addressing this issue when using these models for Automated Essay Scoring is to truncate the input text. This raises serious validity concerns as it undermines the model's ability to fully capture and evaluate organizational elements of the scoring rubric, which requires long contexts to assess. In this study, we evaluate several models that incorporate architectural modifications of the standard transformer architecture to overcome these length limitations using the Kaggle ASAP 2.0 dataset. The models considered in this study include fine-tuned versions of XLNet, Longformer, ModernBERT, Mamba, and Llama models.

1 Introduction

Automated Essay Scoring (AES) is the application of statistical models to approximate the grading of essays by a human using a rubric. The initial models employed for AES were based on word frequencies and hand-crafted features (Page, 2003). The methods and models applied to AES have closely followed those used in more general Natural Language Processing (NLP) applications. The models employed in AES include recurrent and convolutional neural networks (Taghipour and Ng, 2016), models with attention mechanisms (Dong et al., 2017), and transformer-based large language models (LLM) (Rodriguez et al., 2019). Currently, LLMs are readily used to perform AES in research and large-scale assessment (Lottridge et al., 2023).

The first transformer-based LLM to be applied to AES was the Bidirectional Encoder-based Representations by Transformers (BERT) (Devlin et al., 2018). Since BERT arrived on the scene, the BERT model and its derivatives have readily provided state-of-the-art results in a wide range of downstream NLP tasks (Wang et al., 2019). The key to the success of these LLMs has been due to the transformer architecture (Vaswani et al., 2017) and to the ability to pretrain the model weights on a large corpus of unlabeled data on a semisupervised task such as next-token prediction (Radford et al., 2018) or masked-word prediction (Devlin et al., 2018). While we often say that the pretraining provides the model with some limited "understanding", the model weights are simply encoding enough information to encode the necessary word-probability functions.

Transformer-based models are deep feed-forward networks utilizing residual connections between layers that help stabilize training and prevent vanishing gradients (Vaswani et al., 2017). Each layer uses a multiheaded attention mechanism, similar to those used in recurrent networks (Graves et al., 2013). The input is defined by the addition of a positional embedding and a word embedding, which also defines the fixed length of the feedforward network. Since the computing power required by the attention mechanism scales quadratically with length, the length chosen for BERT was 512 (Devlin et al., 2018). This length became something of a standard for the most popular transformer-based LLMs.

The need for models that could overcome the limitations imposed by the transformer architecture became an active area of research shortly after BERT's release. We selected five different models that employ distinct approaches to addressing this challenge. These include versions of XLNet (Yang et al., 2019), Longformer (Beltagy et al., 2020), ModernBERT (Warner et al., 2024), Mamba (Gu and Dao, 2024), and a generative Llama model (AI@Meta, 2024) fine tuned for scoring using parameter-efficient methods (Xu et al., 2023). We give a brief explanation as to how each of these models addresses this limitation in

§2. The most novel of these approaches is applied in the Mamba model, which is the only pretrained language model in this study that uses the state-space model (SSM) (Gu et al., 2021). For SSMs, the computing power required scales linearly with the length of the input.

To understand the limitations of AES, researchers introduced the Automated Student Assessment Prize (ASAP) Dataset using the Kaggle platform (Shermis and Hamner, 2013). This Dataset consists of essay responses to eight prompts, some of which were assessed using trait scoring and some of which were assessed using a simpler holistic rubric. While this dataset became the definitive benchmark for AES methods, most essay responses possessed fewer than 512 tokens. This meant that, while LLMs showed superior performance with respect to traditional AES criteria (Williamson et al., 2012), the dataset did not adequately test the length issues that are often critical in the application of LLMs in large-scale assessment (Lottridge et al., 2023).

A second dataset, known as the Persuasive Essays for Rating, Selecting, and Understanding Argumentative and Discourse Elements (PERSUADE) corpus (Crossley et al., 2022), which was originally designed to evaluate the performance of models that annotate the argumentative components of essays, was later extended to the Automated Student Assessment Prize v2 (ASAP 2.0) (Crossley et al., 2025). We will describe the dataset in more detail below, but many responses in the ASAP 2.0 dataset are too long for most language models.

This article is organized as follows: We use §2 to highlight the characteristically different approaches of the models chosen for this study. This is followed by §3 in which we describe the data used and the training methods. We have two different training regimes: one regime for classification models, such as those obtained by appending a classification, and another regime for generative LLMs. This is followed by the results in §4 and a discussion in §5.

2 Models

In this section, we discuss each model used in this study and why we chose to include it. We have attempted to illustrate if and how these models circumvent the architecturally imposed length restrictions of the standard transformer architecture.

2.1 DeBERTa

The DeBERTa model has a context length of 512. It has been chosen for this study to provide a strong benchmark for models typically used for AES. It is widely regarded as one of the best-performing models in a range of tasks. The model was trained as a discriminator, similarly to the ELECTRA models (Clark et al., 2020). The DeBERTa models also deviate from the standard BERT model by disentangling the word-embedding from the positional embedding (He et al., 2021).

2.2 Longformer

The Longformer model attempts to reconcile the need for local attention with a selective form of global attention. The local attention is applied in the form of a sliding window, similar to attention using convolutional units (Wu et al., 2019) coupled with a form of global attention only applied to special tokens (Beltagy et al., 2020), such as the beginning, ending, and mask tokens. This model still possesses a length limitation, however, by only using attention selectively, the computational burden is mitigated, allowing for pretraining over larger context lengths.

2.3 XLNet

The XLNet model uses the recurrent definition of attention introduced by the Transformer-XL model (Dai et al., 2019). These models have recently been discussed for essays, where the long context was useful in accurately annotating the argumentative components of essays (Ormerod et al., 2023). Almost all masked-language models are encoder-only models; however, the XLNet model is also distinguished as one of the few decoder models that was autoregressively pretrained as a masked-language model (Yang et al., 2019).

To demonstrate the recurrence, suppose any input sequence of length L is denoted $s_\tau = [x_{\tau,1}, \dots, x_{\tau,L}]$ while the hidden state for n-th layer associated with s_τ is $h^n_\tau \in \mathbb{R}^{L\times d}$. The recurrence relation defining $h^n_{\tau+1}$ as a function of h^{n-1}_τ and $h^{n-1}_{\tau+1}$ is given as follows:

$$\begin{aligned} \tilde{h}^{n-1}_{\tau+1} &= [SG(h^{n-1}_\tau) \circ h^{n-1}_{\tau+1}], && \text{(1a)} \\ q^n_{\tau+1} &= h^{n-1}_{\tau+1} W_q, && \text{(1b)} \\ k^n_{\tau+1} &= \tilde{h}^{n-1}_{\tau+1} W_k, && \text{(1c)} \\ v^n_{\tau+1} &= \tilde{h}^{n-1}_{\tau+1} W_v, && \text{(1d)} \\ h^n_{\tau+1} &= \mathrm{MHA}(q^n_{\tau+1}, k^n_{\tau+1}, v^n_{\tau+1}), && \text{(1e)} \end{aligned}$$

where SG is the stop gradient, $[x \circ y]$ is the concatenation operation of two sequences, and MHA is an abbreviation for the typical multiheaded attention mechanism for the transformer layer. The recurrence is built into the definition of $\tilde{h}^n_\tau$, affecting the keys and values. Digging deeper into (1) tells us that while the definition allows for infinite input lengths, there is a functional limitation of the architecture in which the output of any token is only a function of at most LD of the previous tokens where D is the depth of the network. The base and large pretrained models released with (Yang et al., 2019) has $L = 512$ and $D = 12$ and $D = 24$ respectively. This effectively caps the practical length to $6,000$ and $12,000$ for these models, respectively.

2.4 ModernBERT

The ModernBERT model is an encoder-based masked language model benefiting from much of the research that has been conducted since BERTs release (Warner et al., 2024). In particular, applications of generative LLMs have pushed the context length limitations in ways that the previous models stated above have not. The key to the context length of 8196 has been the Rotational Position Embedding (RoPE) (Su et al., 2024). There is a pretraining step in which the model is trained at short lengths with a large rotational component, then further trained on a model that interleaves rotational embedding with small and large rotational values to capture contributions from close and distant tokens. This method, developed in (Fu et al., 2024), was key to extending the context length for a range of popular models such as the herd of Llama models (AI@Meta, 2024).

2.5 Llama

The Llama series is a family of open-source generative LLMs from Meta (AI@Meta, 2024). The models have become as ubiquitously associated with open-source generative models as BERT was to masked language models. These generative models use RoPE (Su et al., 2024) in combination with the methods used to extend context lengths to 128k (Fu et al., 2024). In terms of architecture, the Llama models are a variant of the decoder-only transformer-based models, utilizing RMSNorm layers and a particular activated fully connected layer. We present this architecture in Figure 1, paying particular attention to the linear layers normalizing the input into the multi-headed attention (MHA) mechanism.

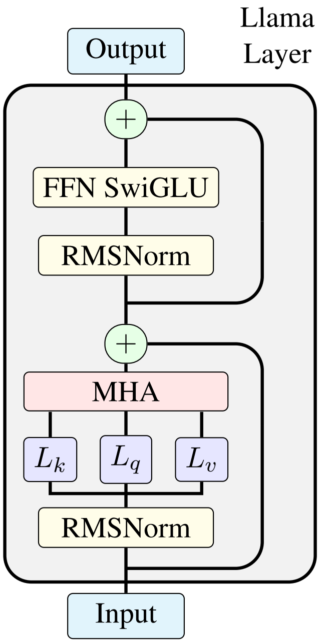

Figure 1: A layer of the Llama decoder-only architecture.

As a generative model, it was trained to predict the next token (Radford et al., 2018), followed by instruction tuning (Chung et al., 2022), followed by a reinforcement learning phase to make the models more useful (Kaufmann et al., 2024). These models come in a variety of sizes. The latest models include multi-modal capabilities; however, the models employed in this article are limited to text.

2.6 State-Space Models

This novel architecture completely replaces the transformer layer and attention with a simpler system based on discretizations of the state-space model (SSM). The SSM is a family of differential equations specified by the matrix equations

$$x'(t) = Ax(t) + Bu(t), \quad (2a)$$
$$y(t) = Cx(t) + Du(t), \quad (2b)$$

where x, u, and y are vectors and A, B, C, and D are matrices. This is a class of models broadly used in control theory. A standard discretization of (2) provides us with the recurrence relation of the form

$$h_t = Ah_{t-1} + Bx_t, \quad (3a)$$
$$x = Ch_t. \quad (3b)$$

A Mamba Layer, in contrast with the Transformer Layer, uses (3) as one component in addition to linear projections, a convolutional layer, and activation functions, as shown in Figure 2.

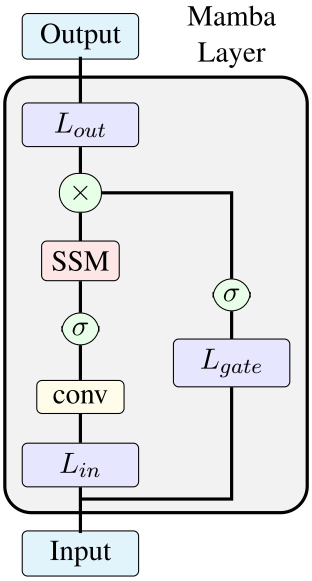

Figure 2: A single layer of the Mamba model.

The Mamba blocks can be computed with linear complexity, making them well-suited for long context tasks (Gu et al., 2021). This claim has been validated empirically by the superior performance of the Jamba models, which is an ensemble of transformer and Mamba layers (Lieber et al., 2024), on RULER benchmarks (Hsieh et al., 2024). As we seek longer and longer context lengths, models with linear complexity may be favorable from an efficiency standpoint.

2.7 Data

The reason we chose the ASAP 2.0 dataset (Crossley et al., 2025) is that this dataset provides a much-needed update of the original ASAP dataset (Shermis and Hamner, 2013), which could be considered to be saturated at this point. This dataset, derived as an extension of the PERSUADE corpus (Crossley et al., 2022), consists of essays written by students from grades 6 to 10 on a wide range of prompts.

Since a key feature of this study is our ability to handle long contexts, it is important to consider the length and grade level characteristics of the data. Because we are using a variety of LLMs, each of which has adopted different subword tokenizations (Kudo and Richardson, 2018), we have no unified notion of what defines a token. In lieu of a uniform tokenization, we will report the word count reported in the dataset. These length characteristics have been presented in Table 1.

	Train		Test	
Grade	Count	Avg. Words	Count	Avg. Words
6	2094	292.2	527	268.3
8	1648	339.9	921	295.9
9	4002	426.1	0	-
10	9563	385.8	5973	356.4
Total	17307	376.1	7421	342.7

Table 1: The size and length characteristics of the ASAP 2.0 dataset.

To evaluate the data, we use the standard metrics specified for AES (Williamson et al., 2012). The main metric used is the agreement statistic known as quadratic weighted kappa (QWK). Generally, the weighted kappa is specified by the equation

$$\kappa = 1 - \frac{\sum_{i,j} W_{i,j} O_{i,j}}{\sum_{i,j} W_{i,j} E_{i,j}} \tag{4}$$

where $O_{i,j}$ is the observed agreement between the first rater giving a score of i and the second rater a score of j, and $E_{i,j}$ is the expected agreement only assuming the two raters' general distribution. This becomes QWK under the weighting

$$W_{i,j} = \frac{(i-j)^2}{(n-1)^2}$$

where n is the number of scores. It is generally understood that this is a measure of agreement above random chance, where a QWK of 1 is perfect agreement and -1 is perfect disagreement. In practical terms, lower scores represent the level of reliability between raters (McHugh, 2012), and our models should be compared against human-human agreements (Williamson et al., 2012). The QWK between the raters is reported to be 0.745

3 Methods

In order to perform essay scoring using LLMs, we distinguish two different cases. We call the first case traditional LLM-based scoring, where the underlying LLM is a masked-language model, such as BERT (Devlin et al., 2018), or a next word predictor such as the Generative Pretrained Transformer (GPT) (Radford et al., 2018). The second class of models considered was generative, which are distinguished by typically possessing an order of magnitude more parameters, and being trained in three phases: pretaining, instruction tuning, and reinforcement (OpenAI, 2023).

3.1 Traditional LLM based scoring

The typical procedure for traditional scoring is to convert a next word or masked word prediction model into a classifier by removing the linear head that would otherwise predict a token and append, in its place, a classification head with as many targets as there are scores (Rodriguez et al., 2019). The classification head is randomly initialized.

To train each of these models, 10% of the training set was designated as a development set. The models were trained by applying the Adam optimizer with a weight decay mechanism (Loshchilov and Hutter, 2019) to the cross-entropy loss function. An initial learning rate of 10^{-6} and a linear learning rate scheduler that reduces the learning rate to 0 over 10 epochs was used with a batch size of either 4 or 1 due to the length of some essays. The QWK was optimized on the development set using an early stopping mechanism.

To fine-tune our Mamba models for classification, we appended a learnable classification head, however, we were required to effectively freeze the weights associated with the SSM, L_{gate}, and the convolutional layer (See Figure 2). Full model training seemed to readily lead to model collapse, perhaps due to the requirement that certain weights take a particular form (Gu et al., 2021). Hence, we fine-tuned the embedding layer and the associated L_{in} and L_{out} weights of every layer. This is a memory-efficient way to fine-tune that provides excellent results. We used the Adam optimizer above with a learning rate of 10^{-5} and a batch size of 8.

3.2 Generative LLM based scoring

Many attempts in the literature seek to optimize the prompting of closed-source generative models to yield higher agreement rates (Xiao et al., 2024). While this is an interesting approach, we believe fine-tuning is necessary to obtain reasonable success. Due to the large size of the models, in order to do this with reasonable computational resources, we need to employ parameter-efficient methods (Xu et al., 2023). These methods can be applied without reference to an API and, hence, can be effectively employed securely, and privately, generating a fraction of the carbon emissions (Bulut et al., 2024).

In the case of fine-tuning generative models, the dataset used mimics an instruction set the model has been trained on. This means that any element of the training set appears to be a user prompting the model to score an essay to a rubric (Ormerod and Kwako, 2024). To do this, we used the following prompt template:

User

```
Assign a **Score** to the
**Essay** using the **Rubric**
provided.

**Rubric**: {rubric}

**Essay**:
```

Assistant

```
**Score**: {score}
```

This template highlights the important aspects by using markdown, due to the formatting of the corpus the model was trained on. Given that variations in prompting can have a significant bearing on the results, we exploit this by allowing the model to summarize and rephrase the rubric in 20 different ways. We optimized the variation of the rubric by evaluating the QWK of the model before fine-tuning on a development set that consisted of 10% of the training set.

We apply the method of low-rank adapters (Hu et al., 2021) and quantization (QLoRA) by (Dettmers et al., 2023). To apply QLoRA to a model, we must specify which linear layers to apply the adapter to, the rank of the adapter, scaling factors, the usual learning rate, and batch size. Concerning Figure 1, we seek to apply low-rank adapters to L_q, L_k, and L_v in the Llama model.

4 Results

The study evaluated various long-context language models on the ASAP 2.0 dataset to assess their effectiveness in automated essay scoring (AES). Models tested included traditional encoder-only architectures like DeBERTa-Base and XLNet-Base, extended-context models such as Longformer and ModernBERT, a state-space model (Mamba-130m), and generative decoder-based models like Llama-3.2-8B.

					Grade		
Model	Reference	L	Model Size	Overall	6	8	10
Human	(Crossley et al., 2025)	inf		0.745			
DeBERTa-Base	(He et al., 2021)	512	183M	0.790	0.696	0.659	0.800
XLNet-Base	(Yang et al., 2019)	8k*	110M	0.784	0.654	0.640	0.798
Longformer	(Beltagy et al., 2020)	4k	149M	0.798	0.698	0.658	0.811
ModernBERT	(Warner et al., 2024)	8k	149M	0.790	0.639	0.658	0.804
Mamba-130m	(Gu and Dao, 2024)	8k*	130M	0.797	0.674	0.640	0.812
Llama-3.2-8B	(AI@Meta, 2024)	8k	8B	0.792	0.667	0.672	0.803

Table 2: The performance of each model in terms of QWK, given by (4). These context lengths for XLNet models and Mamba models are not specified. The value of 8k was implemented as a mechanism to bound the memory required for training.

Human-human rater agreement stood at 0.745, serving as the baseline for comparison. All models surpassed this baseline, with Longformer achieving the highest overall QWK of 0.798. Notably, Mamba-130m performed competitively despite its smaller parameter size, demonstrating that linear-complexity models can rival attention-based transformers in AES tasks. Key findings revealed that long-context models, particularly those using advanced architectural innovations like RoPE-based positional embeddings and selective state spaces, are well-suited for handling lengthy student essays. Traditional models like DeBERTa and XLNet showed strong performance but lagged slightly behind Longformer and Mamba. Despite their large parameter counts and sophisticated training methods – such as instruction tuning and reinforcement learning —- generative models did not significantly outperform encoder-based models. However, they do offer the promising capability of providing feedback (Ormerod and Kwako, 2024).

5 Discussion

Overall, the results affirm the viability of long-context models in automated scoring systems, especially when dealing with complex, lengthy texts where global coherence and argument structure are crucial. Using long context models should not be about getting higher agreement, but rather addressing a glaring flaw from a modeling perspective; it is difficult to argue that traditional language models are faithfully modeling aspects of the rubric, such as organization, when essays are being truncated at 512 tokens.

Our modeling results indicate that both the selective attention mechanism and Mamba's linear complexity architecture deliver robust AES performance on lengthy texts. The study's most notable finding is Mamba's exceptional performance despite its simplified architecture. These differences between these models also suggest a potential for ensemble approaches. Several factors position Mamba and related architectures like Jamba (Lieber et al., 2024) as compelling alternatives for large-scale assessment applications. The linear scaling relationship between computational complexity and sequence length offers significant advantages over traditional transformer architectures. Additionally, optimized implementations may achieve 2-8x speed improvements compared to transformer-based models. These efficiency gains, combined with demonstrated effectiveness on long-context tasks, make state space models like Mamba practical solutions for automated assessment and similar applications requiring efficient processing of extended sequences.

References

AI@Meta. 2024. Llama 3 Model Card.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The Long-Document Transformer. *arXiv preprint*. ArXiv:2004.05150 [cs].

Okan Bulut, Maggie Beiting-Parrish, Jodi M. Casabianca, Sharon C. Slater, Hong Jiao, Dan Song, Christopher M. Ormerod, Deborah Gbemisola Fabiyi, Rodica Ivan, Cole Walsh, Oscar Rios, Joshua Wilson, Seyma N. Yildirim-Erbasli, Tarid Wongvorachan, Joyce Xinle Liu, Bin Tan, and Polina Morilova. 2024. The Rise of Artificial Intelligence in Educational Measurement: Opportunities and Ethical Challenges. *arXiv preprint*. ArXiv:2406.18900.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi

Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, and 16 others. 2022. Scaling Instruction-Finetuned Language Models. *arXiv preprint*. ArXiv:2210.11416 [cs].

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. Technical Report arXiv:2003.10555, arXiv. ArXiv:2003.10555 [cs] type: article.

Scott A. Crossley, Perpetual Baffour, Yu Tian, Aigner Picou, Meg Benner, and Ulrich Boser. 2022. The persuasive essays for rating, selecting, and understanding argumentative and discourse elements (PERSUADE) corpus 1.0. *Assessing Writing*, 54:100667.

Scott Andrew Crossley, Perpetual Baffour, L. Burleigh, and Jules King. 2025. A Large-Scale Corpus for Assessing Source-Based Writing Quality: Asap 2.0.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V. Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context. *arXiv preprint*. ArXiv:1901.02860 [cs, stat].

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient Finetuning of Quantized LLMs. *Advances in Neural Information Processing Systems*, 36:10088–10115.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Technical Report arXiv:1810.04805, arXiv. ArXiv:1810.04805 [cs] type: article.

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.

Yao Fu, Rameswar Panda, Xinyao Niu, Xiang Yue, Hannaneh Hajishirzi, Yoon Kim, and Hao Peng. 2024. Data Engineering for Scaling Language Models to 128K Context. *arXiv preprint*. ArXiv:2402.10171 [cs].

Alex Graves, Navdeep Jaitly, and Abdel-rahman Mohamed. 2013. Hybrid speech recognition with Deep Bidirectional LSTM. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 273–278.

Albert Gu and Tri Dao. 2024. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. *arXiv preprint*. ArXiv:2312.00752 [cs].

Albert Gu, Karan Goel, and Christopher Ré. 2021. Efficiently Modeling Long Sequences with Structured State Spaces.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. *arXiv preprint*. Number: arXiv:2111.09543 arXiv:2111.09543 [cs].

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. 2024. RULER: What's the Real Context Size of Your Long-Context Language Models? *arXiv preprint*. ArXiv:2404.06654 [cs].

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv preprint*. ArXiv:2106.09685 [cs].

Timo Kaufmann, Paul Weng, Viktor Bengs, and Eyke Hüllermeier. 2024. A Survey of Reinforcement Learning from Human Feedback. *arXiv preprint*. ArXiv:2312.14925.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Opher Lieber, Barak Lenz, Hofit Bata, Gal Cohen, Jhonathan Osin, Itay Dalmedigos, Erez Safahi, Shaked Meirom, Yonatan Belinkov, Shai Shalev-Shwartz, Omri Abend, Raz Alon, Tomer Asida, Amir Bergman, Roman Glozman, Michael Gokhman, Avashalom Manevich, Nir Ratner, Noam Rozen, and 3 others. 2024. Jamba: A Hybrid Transformer-Mamba Language Model. *arXiv preprint*. ArXiv:2403.19887 [cs].

Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. *arXiv preprint*. ArXiv:1711.05101 [cs, math].

Susan Lottridge, Chris Ormerod, and Amir Jafari. 2023. Psychometric Considerations When Using Deep Learning for Automated Scoring. In *Advancing Natural Language Processing in Educational Assessment*. Routledge. Num Pages: 16.

Mary L. McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia Medica*, 22(3):276–282.

OpenAI. 2023. GPT-4 Technical Report. *arXiv preprint*. ArXiv:2303.08774 [cs].

Christopher Ormerod, Amy Burkhardt, Mackenzie Young, and Sue Lottridge. 2023. Argumentation Element Annotation Modeling using XLNet. *arXiv preprint*. ArXiv:2311.06239 [cs].

Christopher Michael Ormerod and Alexander Kwako. 2024. Automated Text Scoring in the Age of Generative AI for the GPU-poor. *arXiv preprint*. ArXiv:2407.01873 [cs].

Ellis Batten Page. 2003. Project Essay Grade: PEG. In *Automated essay scoring: A cross-disciplinary perspective*, pages 43–54. Lawrence Erlbaum Associates Publishers, Mahwah, NJ, US.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving Language Understanding by Generative Pre-training.

Pedro Uria Rodriguez, Amir Jafari, and Christopher M. Ormerod. 2019. Language models and Automated Essay Scoring. *arXiv preprint*. Number: arXiv:1909.09482 arXiv:1909.09482 [cs, stat].

Mark D. Shermis and Ben Hamner. 2013. Contrasting State-of-the-Art Automated Scoring of Essays. pages 335–368. Publisher: Routledge Handbooks Online.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. RoFormer: Enhanced transformer with Rotary Position Embedding. *Neurocomputing*, 568:127063.

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.

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 *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. Technical Report arXiv:1804.07461, arXiv. ArXiv:1804.07461 [cs] type: article.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference. *arXiv preprint*. ArXiv:2412.13663 [cs].

David M. Williamson, Xiaoming Xi, and F. Jay Breyer. 2012. A Framework for Evaluation and Use of Automated Scoring. *Educational Measurement: Issues and Practice*, 31(1):2–13. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1745-3992.2011.00223.x.

Felix Wu, Angela Fan, Alexei Baevski, Yann N. Dauphin, and Michael Auli. 2019. Pay Less Attention with Lightweight and Dynamic Convolutions. *arXiv preprint*. ArXiv:1901.10430 [cs].

Changrong Xiao, Wenxing Ma, Sean Xin Xu, Kunpeng Zhang, Yufang Wang, and Qi Fu. 2024. From Automation to Augmentation: Large Language Models Elevating Essay Scoring Landscape. *arXiv preprint*. ArXiv:2401.06431 [cs] version: 1.

Lingling Xu, Haoran Xie, Si-Zhao Joe Qin, Xiaohui Tao, and Fu Lee Wang. 2023. Parameter-Efficient Fine-Tuning Methods for Pretrained Language Models: A Critical Review and Assessment. *arXiv preprint*. ArXiv:2312.12148 [cs].

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.