Automated Essay Scoring for Brazilian Portuguese
Evidence from Cross-Prompt Evaluation of ENEM Essays

André Barbosa and Denis Deratani Mauá
Institute of Mathematics and Statistics, University of São Paulo, São Paulo, Brazil
{andre.barbosa, ddm}@ime.usp.br

Abstract

Brazil's ENEM, a high-stakes assessment determining university admission for millions of students annually, creates an immense evaluation burden where human raters process hundreds of essays daily. Automated Essay Scoring (AES) offers a potential solution, yet Portuguese-language systems remain understudied due to fragmented datasets and the complexity of ENEM's multi-trait rubric. This work investigated cross-prompt, trait-specific essay scoring using a corpus of 385 essays across 38 prompts, where models evaluated essays on unseen prompts across five traits scored on a six-point ordinal scale. We compared three model classes: feature-based methods (72 features), encoder-only transformers (109M–1.5B parameters), and decoder architectures (2.4B–671B parameters) with fine-tuned and zero-shot configurations. Experiments under varying information access and rubric conditioning revealed that no single approach serves all evaluation needs: encoder models excel at mechanical traits (fluency, cohesion) despite context limitations; decoder models achieve superior performance on argumentation (QWK 0.73) and writing style (QWK 0.60) when provided full context; and language-specific pretraining benefits only surface-level features without improving complex reasoning. Best-performing models achieved QWK scores of 0.60–0.73. Gaps to oracle bounds ranged from 0.15 (argumentation) to 0.29 (writing style), with the largest disparities in writing style and persuasiveness.

1 Introduction

Automatic Essay Scoring (AES) systems promise to release educators from the burden of grading written assignments, scaling up the ability to provide timely, consistent, and useful feedback (Page, 1966). These systems have matured significantly, evolving from feature engineering approaches (Page, 1966; Attali and Burstein, 2006; Attali, 2013) to deep neural networks (Taghipour and Ng, 2016; Dong et al., 2017; Alikaniotis et al., 2016) and, more recently, to architectures that leverage Large Language Models (Rodriguez et al., 2019; Mansour et al., 2024). However, the vast majority of this progress focuses on English corpora, leaving Portuguese-language systems understudied.

Brazil's ENEM (*Exame Nacional do Ensino Médio*) exemplifies this challenge. The high-stakes examination annually evaluates 3.9 million students, serving as the primary gateway to higher education. Human raters process 100–200 essays daily under a 20-day evaluation window, with students waiting up to 8 weeks for official results.[1] Secondary schools face compounding difficulties: class sizes of 30–50 students, limited faculty allocation for essay review, and insufficient resources for ENEM-style assessment. The consequences are predictable: delayed feedback impedes learning, and teacher burnout affects evaluation quality.

Developing AES systems to address these challenges faces its own obstacles. Existing datasets suffer from parsing artifacts and lack data provenance (Marinho et al., 2021). Prior empirical work has been limited to feature-based methods or shallow neural networks (Amorim and Veloso, 2017; Fonseca et al., 2018), with no systematic analysis of modern transformer-based architectures. Additionally, most research focuses on single-prompt scoring, which does not reflect realistic deployment where systems must generalize to unseen topics.

This work addressed four research questions. Firstly, what information do different traits require, that is, do all traits benefit equally from access to essay prompts and supporting texts? Secondly, how does information access (prompt-blind vs. prompt-aware) affect performance across the five ENEM traits. Thirdly, do Portuguese-specific models outperform multilingual alternatives, and for which

[1] As reported in https://tinyurl.com/57e5xsdv

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 2, pages 43–48

traits? Lastly, what are the practical trade-offs between model architectures regarding accuracy, computational cost, and inference latency?

We hypothesized that different traits require different computational approaches: mechanical traits such as fluency and cohesion can be evaluated with limited context, while argumentative quality requires access to the essay prompt and supporting materials. To address these questions, this work provided an extensive empirical analysis comparing 15 models spanning three paradigms: feature-based methods (72 linguistic features), encoder-only transformers (109M–1.5B parameters), and decoder architectures (2.4B–671B parameters) including fine-tuned and zero-shot configurations.

The main contributions of this dissertation can be outlined as follows:

1. A validated benchmark corpus of 385 essays across 38 prompts with expert annotations from two independent graders;
2. A systematic comparison of model architectures under varying information access and rubric conditioning;
3. A formal framework distinguishing prompt-blind and prompt-aware scoring with three rubric strategies (Student, Mixed, Grader);
4. Trait-specific analysis demonstrating that encoder models excel at mechanical traits despite context limitations while decoder models achieve superior performance on argumentation (QWK 0.73) and style (QWK 0.60) when provided full context.
5. Open research artifacts including the annotated corpus and evaluation scripts. The dataset and models generated during this research are available in `https://tinyurl.com/245mxct9`. Experiments and code used are available in `https://github.com/kamel-usp/jbcs2025`.

This research resulted in three publications. The first introduced a benchmark corpus of 385 essays across 38 different topics (also called prompts) with expert annotations from two independent graders, establishing baseline performance with encoder models (Silveira et al., 2024). The second investigated the robustness of transformer-based scorers against adversarial attacks, revealing vulnerabilities in both encoder and decoder architectures (Silveira et al., 2025). The third provided an extensive empirical analysis comparing 15 models across three paradigms, achieving state-of-the-art results with trait-specific QWK scores ranging from 0.60 to 0.73 (Barbosa et al., 2025).

The remainder of this extended abstract is organized as follows. Section 2 presents the conceptual framework including ENEM traits and information paradigms. Section 3 describes the methodology and presents experimental results. Section 4 concludes with practical implications and future directions. The full thesis is available at: `https://tinyurl.com/mvk34d98`

2 Concepts & Framework

The ENEM essay task assesses five traits that span different dimensions of writing quality. Table 1 summarizes these traits according to the official candidate guidelines. Traits C1 (Fluency) and C4 (Cohesion) evaluate surface-level linguistic features: grammar, spelling, punctuation, and the use of cohesive devices. These mechanical traits can be assessed with limited contextual information, as they depend primarily on the essay text itself. In contrast, C2 (Writing Style), C3 (Argumentation), and C5 (Persuasion) might require understanding the relationship between the essay and its prompt. Evaluating whether a student adequately addresses the topic, constructs relevant arguments, or proposes a viable intervention demands access to the prompt and supporting materials that define the task.

This work investigated cross-prompt, trait-specific scoring. In such a setting, models must evaluate essays on prompts not seen during training. Each trait is scored on a six-point ordinal scale $\{0, 40, 80, 120, 160, 200\}$, with the holistic (overall) score computed as the sum across all five traits (0–1000). The cross-prompt setting tests whether models learn transferable evaluation principles rather than prompt-specific patterns, reflecting realistic deployment scenarios where systems must generalize to new essay topics.

Two experimental dimensions systematically vary the information available to scoring models. The first dimension concerns *information access*: *prompt-blind* models receive only the essay text, while *prompt-aware* models receive the essay together with the prompt and supporting materials. This distinction is critical because mechanical traits (C1, C4) can theoretically be evaluated without prompt access, whereas argumentative traits (C2,

Code	Trait	Label	Description
C1	Trait 1	Fluency	Demonstrate command of the formal written modality of the Portuguese language.
C2	Trait 2	Writing Style	Understand the writing prompt and apply concepts from various fields of knowledge to develop the topic, within the structural limits of the argumentative-essay prose format.
C3	Trait 3	Argumentation and Relevance to Prompt	Select, relate, organize, and interpret information, facts, opinions, and arguments in defense of a point of view.
C4	Trait 4	Cohesion	Demonstrate knowledge of the linguistic mechanisms necessary for building argumentation.
C5	Trait 5	Persuasion/Intervention Proposal	Develop an intervention proposal for the issue addressed, while respecting human rights.

Table 1: Descriptions of the five ENEM essay scoring traits.

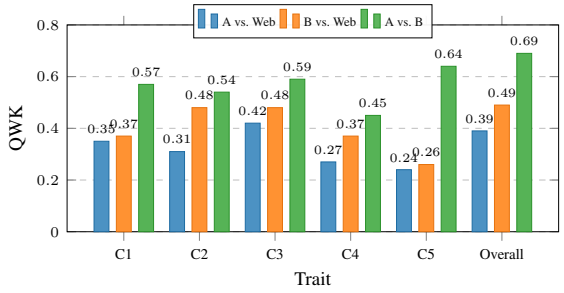

Figure 1: Human inter-rater agreement (QWK) across ENEM traits. Agreement between expert graders (A vs. B) ranged from 0.45 to 0.69 QWK.

C3, and C5) require understanding the task context. The second dimension concerns *rubric conditioning*, applicable only to decoder models operating in a zero-shot setting. Three rubric strategies were derived from official ENEM materials: Student guidelines (high-level descriptions provided to test-takers), Grader guidelines (detailed criteria from the official handbook), and Mixed guidelines (student-level descriptions with grader-level scoring rubrics).

Performance is measured using Quadratic Weighted Kappa (QWK), which quantifies ordinal agreement while penalizing predictions proportionally to their distance from ground truth (Cohen, 1960; de la Torre et al., 2018; Williamson et al., 2012; Ramnarain-Seetohul et al., 2022; Doewes et al., 2023). Standard bands classify QWK as poor (< 0.40), fair-to-good (0.40–0.74), or excellent (≥ 0.75) (Burrows et al., 2015; Fleiss et al., 2003), though these thresholds are not universally reliable (Bakeman et al., 1997). F1 Macro and F1 Weighted complement QWK by capturing class-level accuracy (Mello et al., 2024). As shown in Figure 1, human inter-rater agreement ranged from 0.45 to 0.69 QWK across traits, establishing a practical ceiling for automated systems.

3 Results and Discussion

Experiments use the dataset introduced by Silveira et al. (2024), comprising 385 essays across 38 prompts with independent annotations from two expert graders. When graders disagree, any prediction necessarily conflicts with at least one reference. Oracle baselines contextualize this ceiling: *Mean-Grade* (graders' rounded arithmetic mean) serves as the upper bound, while R_0 (most frequent training score) provides a lower bound.

The investigation systematically compared 15 models spanning three paradigms: *feature-based* methods using 72 linguistic features with Linear Regression and Random Forest classifiers; *encoder-only* transformers (109M–1.5B parameters) including BERTimbau, Albertina, and multilingual BERT; and *decoder architectures* divided into fine-tuned small language models (2.4B–14.7B parameters) such as Tucano, Phi-3, Llama3, and Phi-4, and zero-shot learners including GPT-4o, Sabiá3, and DeepSeek-R1 (up to 671B parameters). Table 2 summarizes model characteristics.

Category	Models	Params	Training
Feature-based	LR, RF	—	Full
Encoder-only	5 models	109M–1.5B	Full FT
Decoder (SLM)	4 models	2.4B–14.7B	LoRA
Decoder (ZSL)	3 models	up to 671B	Zero-shot

Table 2: Model categories evaluated. LR: Linear Regression; RF: Random Forest; FT: Fine-tuning; SLM: Small Language Models; ZSL: Zero-Shot Learners.

For decoder models, Figure 2 illustrates the prompt engineering framework exploring three dimensions: guideline source (Student, Grader, or Mixed), context inclusion (prompt-blind vs. prompt-aware), and response structuring via Chain-of-Thought (Wei et al., 2022) reasoning. This design yields six experimental conditions per trait for zero-shot evaluation.

Table 3 presents the best-performing configuration for each model class across all five traits, alongside feature-based baselines and oracle upper bounds. A detailed analysis of these interactions across all model configurations is provided in Barbosa et al. (2025).

Several key findings emerge from these results. Firstly, feature-based classifiers consistently under-

	C1: Fluency			C2: Writing Style			C3: Argument			C4: Cohesion			C5: Persuasion		
Model	M	W	Q	M	W	Q	M	W	Q	M	W	Q	M	W	Q
R_0 (Baseline)	.12	.20	.00	.11	.20	.00	.09	.17	.00	.14	.39	.00	.05	.05	.00
Linear Regressor	.41	.53	.36	.13	.23	.32	.17	.27	.26	.30	.57	.45	.15	.17	.03
Random Forest	.32	.56	.41	.14	.22	.22	.21	.29	.35	.35	.64	.48	.11	.15	.12
Best Encoder	.55	.71	.68	.33	.43	.32	.25	.36	.29	.48	.61	.60	.29	.37	.63
Best SLM	.52	.64	.67	.42	.52	.60	.37	.38	.57	.37	.58	.55	.44	.49	.59
Best ZSL	.35	.66	.69	.32	.43	.53	.44	.47	.73	.37	.60	.56	.37	.43	.60
Best	**.55**	**.71**	**.69**	**.42**	**.52**	**.60**	**.44**	**.47**	**.73**	**.48**	**.64**	**.60**	**.44**	**.49**	**.63**
MeanGrade (Oracle)	.71	.82	.85	.55	.71	.89	.62	.65	.88	.66	.83	.81	.68	.68	.90

Table 3: Test-set performance across model classes. M: macro F1, W: weighted F1, Q: QWK. SLM: Small Language Models (fine-tuned). ZSL: Zero-Shot Learners. MeanGrade represents the oracle upper bound.

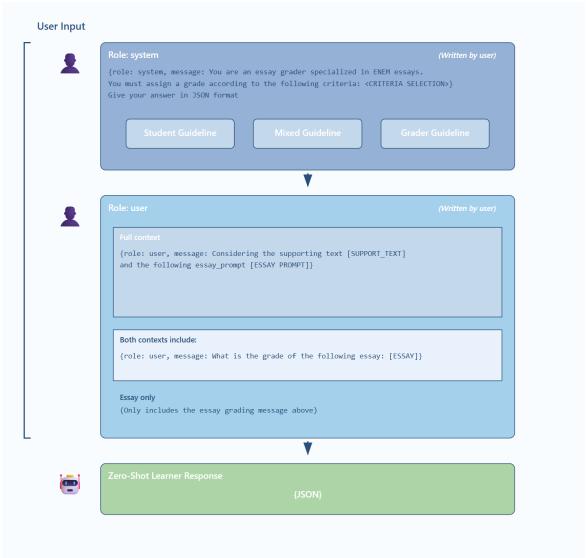

Figure 2: Prompt engineering framework for zero-shot essay scoring.

perform relative to neural approaches, with particularly poor results on Persuasion (C5). Secondly, no single architecture dominates across all traits: encoder models achieve the highest or comparable performance for Cohesion (C4: QWK 0.60) and Persuasion (C5: QWK 0.63); zero-shot learners excel at Argumentation (C3: QWK 0.73), likely due to emergent reasoning capabilities; and fine-tuned SLMs achieve the best results for Writing Style (C2: QWK 0.60). All three architectures achieve comparable results for Fluency (C1: QWK ranging from 0.67–0.69).

The impact of information access varies substantially across traits. For mechanical traits (C1, C4), additional context provides minimal benefit or even degrades performance, consistent with the "lost-in-the-middle" phenomenon where models struggle to retrieve information from extended sequences (Liu et al., 2024). In contrast, argumentative traits (C2, C3) show substantial gains when models receive the essay prompt and supporting materials.

Prompt engineering proves crucial for zero-shot performance. Notably, Sabiá3 (Abonizio et al., 2025), the Portuguese-specific model, outperforms multilingual alternatives only on Fluency (C1), suggesting that monolingual pretraining benefits surface-level linguistic assessment but confers minimal advantage for reasoning-intensive traits. A detailed analysis of model variants, context effects, and prompt engineering strategies is provided in Barbosa et al. (2025).

Comparison with oracle bounds reveals varying improvement potential across traits. Argumentation (C3: gap of 0.15) and Fluency (C1: gap of 0.16) approach human inter-rater agreement levels, while Writing Style (C2: gap of 0.29), Persuasion (C5: gap of 0.27), and Cohesion (C4: gap of 0.21) exhibit substantial headroom for advancement.

4 Conclusion

This dissertation explored how well automated systems can assess ENEM essays across five distinct traits. By systematically comparing 15 models under varying information access paradigms and rubric conditioning strategies, the research demonstrates that model performance varies significantly depending on which aspect of writing quality is being evaluated and what information is available to the model.

The findings demonstrate a striking pattern: models can almost reach human inter-rater agreement when evaluating mechanical aspects of writing, such as Fluency and Cohesion, as these features manifest as identifiable patterns within the text. However, significant disparities remain for traits that require deeper semantic understanding, including Writing Style, Argumentation, and Persuasion. This divergence indicates that current approaches are proficient at detecting linguistic indi-

cators but face challenges with tasks necessitating genuine comprehension of meaning and rhetorical effectiveness. Furthermore, no single configuration or architecture optimizes performance across all traits, which accounts for the consistent underperformance of unified approaches compared to specialized strategies.

Revisiting the research question, the evidence indicates that current AES systems occupy a middle ground between pattern matching and authentic evaluation. Evaluating whether arguments adequately respond to prompts, whether stylistic choices serve communicative goals, or whether intervention proposals present viable solutions requires reasoning capabilities beyond current systems. While these models detect surface-level quality markers, they do not engage with textual meaning the way humans do. Whether AES systems achieve genuine evaluation or merely sophisticated pattern matching remains an open challenge.

Several directions remain for future investigation. First, performance gaps for Writing Style (C2), Argumentation (C3), and Persuasion (C5) suggest these traits should receive focused attention in subsequent research. Second, encoder-only models with extended context windows, such as ModernBERT (Warner et al., 2024), could address current limitations of encoder-based approaches, though no such models existed for Portuguese at the time of this research. Third, exploring few-shot learning strategies may bridge the gap toward human inter-rater agreement levels beyond what context-limited zero-shot approaches achieve. Fourth, leveraging large language models for synthetic data generation could expand dataset size while preserving annotation consistency, enabling better performance for fine-tuned smaller models. Fifth, incorporating tool usage capabilities (Schick et al., 2023; Yao et al., 2023) could enable models to selectively retrieve trait guidelines or supporting materials only when needed, potentially mitigating the lost-in-the-middle effects observed with full-context inputs.

Rather than pursuing full automation, the field should embrace human-AI collaboration: automated systems handle mechanical evaluation while human graders focus on semantic dimensions where their understanding remains irreplaceable. The question is not whether machines can replace teachers, but how human and artificial intelligence can deliver more frequent, detailed feedback than either could provide alone.

4.1 Publications

This dissertation resulted in three peer-reviewed publications, summarized in Table 4.

Publication	Type	Year	Role
PROPOR (Silveira et al., 2024)	Conference	2024	Co-author
BRACIS (Silveira et al., 2025)	Conference	2025	Co-author
JBCS (Barbosa et al., 2025)	Journal	2025	First author

Table 4: Publications from this dissertation.

Acknowledgements

This work was partially supported by the São Paulo Research Agency (FAPESP) Grant no. 2022/02937-9, CNPq Grant no. 305136/2022-4 and CAPES Finance Code 001.

References

Hugo Abonizio, Thales Sales Almeida, Thiago Laitz, Roseval Malaquias Junior, Giovana Kerche Bonás, Rodrigo Nogueira, and Ramon Pires. 2025. Sabiá-3 technical report. *Preprint*, arXiv:2410.12049.

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 715–25.

Evelin Amorim and Adriano Veloso. 2017. A multi-aspect analysis of automatic essay scoring for Brazilian Portuguese. In *Proceedings of the Student Research Workshop at the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 94–102.

Yigal Attali. 2013. *Validity and Reliability of Automated Essay Scoring*. Routledge.

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

R. Bakeman, D. McArthur, V. Quera, and B. F. Robinson. 1997. Detecting sequential patterns and determining their reliability with fallible observers. *Psychological Methods*, 2:357–370.

André Barbosa, Igor Cataneo Silveira, and Denis Deratani Mauá. 2025. An empirical analysis of large language models for automated cross-prompt essay trait scoring in brazilian portuguese. *Journal of the Brazilian Computer Society*, 31(1):857–870.

Steven Burrows, Iryna Gurevych, and Benno Stein. 2015. The eras and trends of automatic short answer grading. *International Journal of Artificial Intelligence in Education*, 25(1):60–117.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Jordi de la Torre, Domenec Puig, and Aida Valls. 2018. Weighted kappa loss function for multi-class classification of ordinal data in deep learning. *Pattern Recognition Letters*, pages 144–154. Machine Learning and Applications in Artificial Intelligence.

Afrizal Doewes, Nughthoh Arfawi Kurdhi, and Akrati Saxena. 2023. Evaluating quadratic weighted kappa as the standard performance metric for automated essay scoring. In *Proceedings of the 16th International Conference on Educational Data Mining*, pages 103–113. International Educational Data Mining Society.

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*, pages 153–162.

Joseph L. Fleiss, Bruce Levin, and Myunghee Cho Park. 2003. *The Measurement of Interrater Agreement*, chapter 18. John Wiley & Sons, Ltd.

Erick Rocha Fonseca, Ivo Medeiros, Dayse Kamikawachi, and Alessandro Bokan. 2018. Automatically grading brazilian student essays. In *Proceedings of International Conference on Computational Processing of the Portuguese Language*, pages 170–179.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Watheq Ahmad Mansour, Salam Albatarni, Sohaila Eltanbouly, and Tamer Elsayed. 2024. Can large language models automatically score proficiency of written essays? In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2777–2786.

Jeziel Marinho, Rafael Anchiêta, and Raimundo Moura. 2021. Essay-br: a brazilian corpus of essays. In *Anais do III Dataset Showcase Workshop*, pages 53–64.

Rafael Ferreira Mello, Hilário Oliveira, Moésio Wenceslau, Hyan Batista, Thiago Cordeiro, Ig Ibert Bittencourt, and Seiji Isotanif. 2024. Propor'24 competition on automatic essay scoring of portuguese narrative essays. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese-Vol. 2*, pages 1–5.

Ellis B. Page. 1966. The imminence of... grading essays by computer. *The Phi Delta Kappan*, pages 238–243.

Vidasha Ramnarain-Seetohul, Vandana Bassoo, and Yasmine Rosunally. 2022. Similarity measures in automated essay scoring systems: A ten-year review. *Education and Information Technologies*, 27(4):5573–5604.

Pedro Uria Rodriguez, Amir Jafari, and Christopher M. Ormerod. 2019. Language models and automated essay scoring. *Preprint*, arXiv:1909.09482.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *Preprint*, arXiv:2302.04761.

Igor Cataneo Silveira, André Barbosa, Daniel Silva Lopes da Costa, and Denis Deratani Mauá. 2025. Investigating universal adversarial attacks against transformers-based automatic essay scoring systems. In *Intelligent Systems*, pages 169–183, Cham. Springer Nature Switzerland.

Igor Cataneo Silveira, André Barbosa, and Denis Deratani Mauá. 2024. A new benchmark for automatic essay scoring in Portuguese. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, pages 228–237.

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.

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. *Preprint*, arXiv:2412.13663.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

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.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. *Preprint*, arXiv:2210.03629.