Unveiling Essay Scoring Mechanisms in Large Language Models

Anonymous ACL submission

Abstract

Recent advances in Large Language Models (LLMs) have substantially transformed Automated Essay Scoring (AES), yet the internal mechanisms underlying LLM-based scoring remain poorly understood. In this work, we systematically analyze the hidden representations of eight LLMs across two English essay datasets (ASAP++, CSEE) and one Portuguese dataset (ENEM). Using linear probing, cross-prompt generalization, dimensionality reduction, and neuron-level analyses, we find consistent evidence that essay quality information is encoded in a linearly accessible form within LLM representations. These representations emerge progressively across layers, remain robust across prompting strategies, and partially transfer across essay prompts despite differences in scoring rubrics. In addition, nonlinear probes provide only marginal and inconsistent improvements over linear probes, suggesting that most essay quality information is already linearly decodable. We further identify individual "essay scoring neurons" whose activations strongly correlate with essay scores and whose behavior is sensitive to targeted intervention. Moreover, the layer-wise distribution of these neurons systematically shifts with essay length, with longer essays relying more heavily on deeper layers. Overall, our findings provide evidence that LLMs encode structured representations related to essay quality and offer new insights into the interpretability of LLM-based AES systems.

1 Introduction

Automated Essay Scoring (AES) aims to provide scalable and consistent evaluation of student writing. Traditional AES methods have largely followed two paradigms. Prompt-specific models are trained and evaluated on essays from the same essay prompt[1], achieving strong in-domain performance but often generalizing poorly to unseen prompts (Rudner and Liang, 2002; Miltsakaki and Kukich, 2004; Yannakoudakis et al., 2011; Rodriguez et al., 2019; Nadeem et al., 2019; Yang et al., 2020; Uto et al., 2020; Alikaniotis et al., 2016; Dong et al., 2017; Taghipour and Ng, 2016). Cross-prompt approaches improve transferability through domain adaptation and generalization techniques (Ridley et al., 2020; Li and Ng, 2024; Chen and Li, 2023; Wang et al., 2025; Zhang et al., 2025), but they still rely heavily on annotated data and typically underperform compared with prompt-specific systems.

Recent advances in Large Language Models (LLMs) have substantially changed this landscape. With carefully designed prompts, LLMs can perform essay scoring in zero-shot or few-shot settings (Mizumoto and Eguchi, 2023; Yancey et al., 2023; Escalante et al., 2023; Stahl et al., 2024), reducing the dependence on labeled datasets. Moreover, unlike conventional AES systems that mainly output numerical scores, LLMs can also provide diagnostic feedback and personalized comments, enabling richer forms of writing assessment. However, despite these advantages, LLM-based AES still faces important challenges. Their scoring performance often remains unstable compared with strong supervised AES systems (Lee et al., 2024). In addition, as LLMs remain inherently black-box systems, their internal decision-making processes are opaque, and their outputs are highly sensitive to LLM prompt design (Han et al., 2024). These limitations raise concerns about reliability and trustworthiness in educational applications, particularly in high-stakes assessment settings.

A central open question is therefore how LLMs internally represent essay quality. In particular, it remains unclear whether LLMs derive their scoring ability primarily from superficial statistical cues or

[1] To avoid confusion between essay prompts (writing tasks assigned to students) and LLM prompts (inputs to large language models), the unqualified term *prompt* in this paper refers to essay prompts.

whether they learn structured representations that capture higher-level aspects of writing quality. Understanding this distinction is important not only for interpretability, but also for evaluating the robustness and generalizability of LLM-based AES systems.

In this work, we investigate the internal representations underlying LLM-based AES through representation- and neuron-level analyses. We analyze eight models on two English essay datasets and one Portuguese dataset. Through linear probing, cross-prompt generalization, dimensionality reduction, and neuron intervention experiments, we study how essay quality information is represented across model layers and neurons.

Our results show that essay quality information is progressively constructed across layers and is largely linearly decodable from hidden representations. These representations remain relatively stable across prompting strategies and partially transfer across essay prompts despite differences in scoring rubrics. We further identify individual "essay-scoring neurons" that strongly correlate with essay scores and exhibit sensitivity to targeted intervention. Finally, we find that the layer-wise distribution of these neurons systematically shifts with essay length, suggesting that longer essays rely more heavily on deeper-layer computations. Together, these findings provide new insights into the internal mechanisms underlying LLM-based AES and contribute toward more interpretable and trustworthy intelligent scoring systems.

2 Related Work

2.1 Automated Essay Scoring

Early prompt-specific AES models, based on handcrafted features or neural networks, required labeled data for each new essay prompt (Miltsakaki and Kukich, 2004; Yannakoudakis et al., 2011; Alikaniotis et al., 2016; Dong et al., 2017; Rodriguez et al., 2019). To improve generalization, cross-prompt methods were later proposed (Ridley et al., 2020; Chen and Li, 2023; Li and Ng, 2024; Wang et al., 2025; Zhang et al., 2025). More recently, LLM-based zero-shot AES has emerged, enabling essay scoring without labeled data (Mizumoto and Eguchi, 2023; Yancey et al., 2023; Escalante et al., 2023; Stahl et al., 2024). Early approaches relied on simple rubric-based prompting, while later methods such as Multi-Trait Specification (Lee et al., 2024) introduced fine-grained, trait-level evaluation. However, direct scoring remains sensitive to LLM prompt design and prone to bias. RRecent work by Shibata and Miyamura (2025) addresses these issues by reformulating AES as a pairwise essay comparison task, improving robustness at the cost of greater computational overhead and reliance on unlabeled data.

2.2 Interpretability and Probing

A major direction in interpretability research concerns identifying what information is encoded in model representations and how that information supports downstream tasks. Probing methods have become one of the dominant approaches for this purpose. In probing, external classifiers are trained on hidden representations to predict linguistic, semantic, or task-related attributes, under the assumption that successful prediction indicates that the relevant information is encoded in the model (Ettinger et al., 2016; Belinkov and Glass, 2019). Probing studies have been used to analyze a wide range of properties, including syntax, morphology, factual knowledge, and reasoning abilities across pretrained language models. However, subsequent work has questioned whether probe performance alone provides reliable evidence about representation quality. In particular, expressive probes may recover task signals independently of the structure of the underlying representation, making it difficult to distinguish information genuinely encoded by the model from information introduced by the probe itself (Hewitt and Liang, 2019; Belinkov, 2022a). These limitations have motivated a broader shift toward studying the structure, geometry, and dynamics of representations rather than relying exclusively on probing accuracy.

Recent interpretability research therefore increasingly focuses on understanding how representations are organized internally and how they support model computation. Prior work has examined geometric properties of contextual embeddings (Rogers et al., 2020), investigated the emergence of linear features and feature superposition in deep networks (Elhage et al., 2022), and developed mechanistic interpretability techniques aimed at identifying neurons, attention heads, or circuits associated with particular behaviors (Olah et al., 2020; Räuker et al., 2023). Collectively, these studies move beyond the question of whether information exists in a representation toward understanding how information is distributed, transformed, and utilized during inference. Despite these advances,

interpretability research in LLM-based AES remains limited. Existing work has largely concentrated on prompting strategies that generate explanations or formative feedback for users (Xiao et al., 2025), while comparatively little attention has been paid to the internal representations underlying essay evaluation itself. While concurrent work has demonstrated that LLM activations can serve as effective features for cross-prompt scoring (Chi et al., 2025), it remains unclear how these models structurally encode and utilize essay-quality signals during inference.

3 Approach

Given a dataset of n essays $E = \{e_1, e_2, \ldots, e_n\}$ and their corresponding human-annotated target scores $Y = \{y_1, y_2, \ldots, y_n\}$ (which can be overall scores or trait scores), we feed all essays into the model and extract the hidden state activations (i.e., residual stream representations) corresponding to the final token of each essay across all layers. Let $H_i^{(l)} \in R^{L_i \times d_{\text{model}}}$ denote the hidden state matrix for the i-th essay at layer l, where L_i is the sequence length. The hidden state activation corresponding to the last token is extracted as:

$$h_{i,L_i}^{(l)} = H_i^{(l)}[-1, :] \in R^{1 \times d_{\text{model}}}$$

Collecting these representations across all essays yields the activation matrix for layer l:

$$A^{(l)} = \begin{bmatrix} h_{1,L_1}^{(l)} \\ h_{2,L_2}^{(l)} \\ \vdots \\ h_{n,L_n}^{(l)} \end{bmatrix} \in R^{n \times d_{\text{model}}}$$

To examine whether LLM representations encode essay quality information, we adopt standard probing methodologies (Alain and Bengio, 2018; Belinkov, 2022b), which aim to assess whether target labels associated with annotated inputs can be recovered from model representations using simple supervised predictors. Specifically, given an activation matrix $A^{(l)}$ and target scores Y, we train a linear ridge regression probe defined as

$$\hat{W} = \arg\min_{W} \left\| Y - A^{(l)} W \right\|_2^2 + \lambda \left\| W \right\|_2^2,$$

where λ denotes the regularization coefficient. The closed-form solution is given by

$$\hat{W} = \left(A^{(l)\top} A^{(l)} + \lambda I \right)^{-1} A^{(l)\top} Y.$$

Using the learned probe parameters, predictions are obtained as

$$\hat{Y} = A^{(l)} \hat{W}.$$

Strong generalization performance on out-of-sample data suggests that essay quality information is linearly decodable from the underlying model representations. However, consistent with prior work (Ravichander et al., 2021; Gurnee and Tegmark, 2023), successful probing does not necessarily imply that the base model itself utilizes these representations during inference. In all experiments, the regularization parameter λ is selected via efficient leave-one-out cross-validation performed on the probe training set (Hastie et al., 2009).

4 Experiments

4.1 LLMs

We evaluate eight instruction-tuned LLMs from the Llama-3.1/3.2 (Grattafiori et al., 2024), Qwen2.5/3 (Team, 2024, 2025), and Phi-4 (Microsoft et al., 2025) families: Llama-3.2-1B-Instruct, Llama-3.2-3B-Instruct, Llama-3.1-8B-Instruct, Qwen2.5-3B-Instruct, Qwen3-4B-Instruct-2507, Qwen2.5-7B-Instruct, Qwen2.5-14B-Instruct, and Phi-4-mini-instruct. Ranging from 1B to 14B parameters, these models represent diverse architectures and training strategies, supporting the generalizability of our findings.

4.2 Datasets and Evaluation Metrics

We conduct experiments on two English essay-scoring datasets, ASAP++ (Mathias and Bhattacharyya, 2018) and CSEE[2] (Xiao et al., 2025), as well as a Portuguese dataset ENEM[3] (Silveira et al., 2024a). We include CSEE to mitigate potential data leakage concerns, as the dataset was released in 2025, after the training cutoff dates of the evaluated models. Detailed descriptions of all datasets are provided in Appendix A.

Following prior AES research (Dong et al., 2017; Li and Ng, 2024), we evaluate model performance using the Quadratic Weighted Kappa (QWK) metric (Cohen, 1960). Consistent with standard practice in prompt-specific AES settings (Dong et al., 2017; Xiao et al., 2025), we split each dataset into 80% training data and 20% testing data.

[2] https://catalog.ldc.upenn.edu/LDC2014T06
[3] https://github.com/kamel-usp/aes_enem

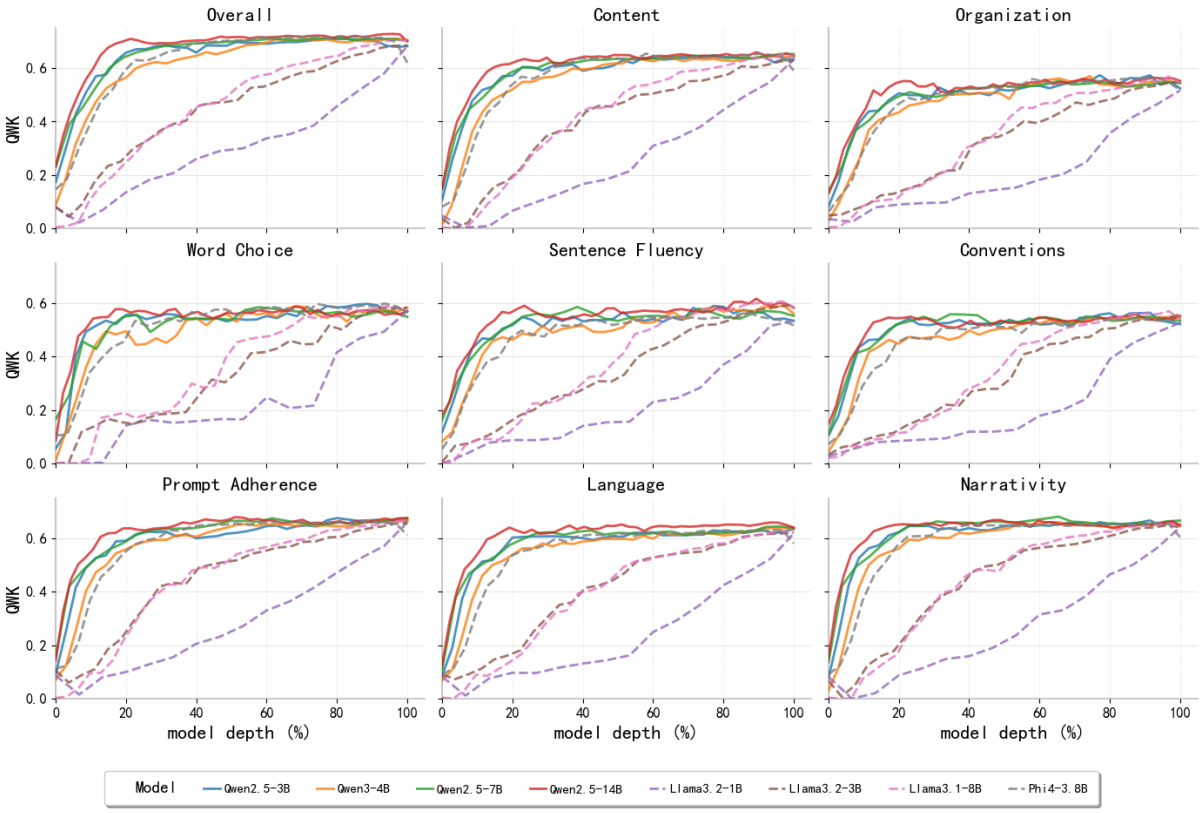

Figure 1: Average QWK scores of linear probes across all essay prompts in ASAP++. Each subplot corresponds to an essay trait and shows probe performance across layers for different models.

All experimental details (including model training and probe configurations) are provided in Appendix B.

4.3 Results

Essay Quality in Representations. As shown in Figure 1, linear probes exhibit similar trends across models and traits. Essay quality information becomes increasingly accessible in deeper layers, while final-layer performance remains broadly comparable across different models. Larger models tend to encode essay quality information more rapidly in earlier layers, leading to steeper initial performance gains, but their final peak performance differs only marginally from that of smaller models.

The results also reveal two interesting patterns. First, Llama models exhibit behavior that differs markedly from that of Qwen and Phi models. Although their final-layer QWK scores (hence representation quality) are comparable, Qwen and Phi models reach saturation at around 20% of model depth and subsequently plateau, whereas Llama models continue improving steadily all the way until the final layers. At present, we can only offer a tentative hypothesis for this phenomenon: it may stem from differences in training data quality, as prior work has shown that high-quality training data can substantially shape model capabilities, enabling smaller models to rival larger ones (Gunasekar et al., 2023; Zhang et al., 2024). Second, the probes consistently predict overall essay scores more accurately than individual trait scores. This suggests that the models capture coarse-grained essay quality representations more effectively than fine-grained trait-specific ones.

Linear Decodability. We compare linear ridge regression probes with more expressive nonlinear MLP probes of the form $W_2\mathrm{ReLU}(W_1x + b_1) + b_2$, using 256 hidden neurons (see Appendix C). Across traits, nonlinear probes provide only marginal and inconsistent improvements in QWK over linear probes. This suggests that essay quality information is largely linearly decodable from the hidden representations, i.e., a linear readout is sufficient to recover most of the task-relevant signal. This finding is consistent with prior work in interpretability research supporting the linear representation hypothesis, which proposes that features in neural networks can be recovered by projecting

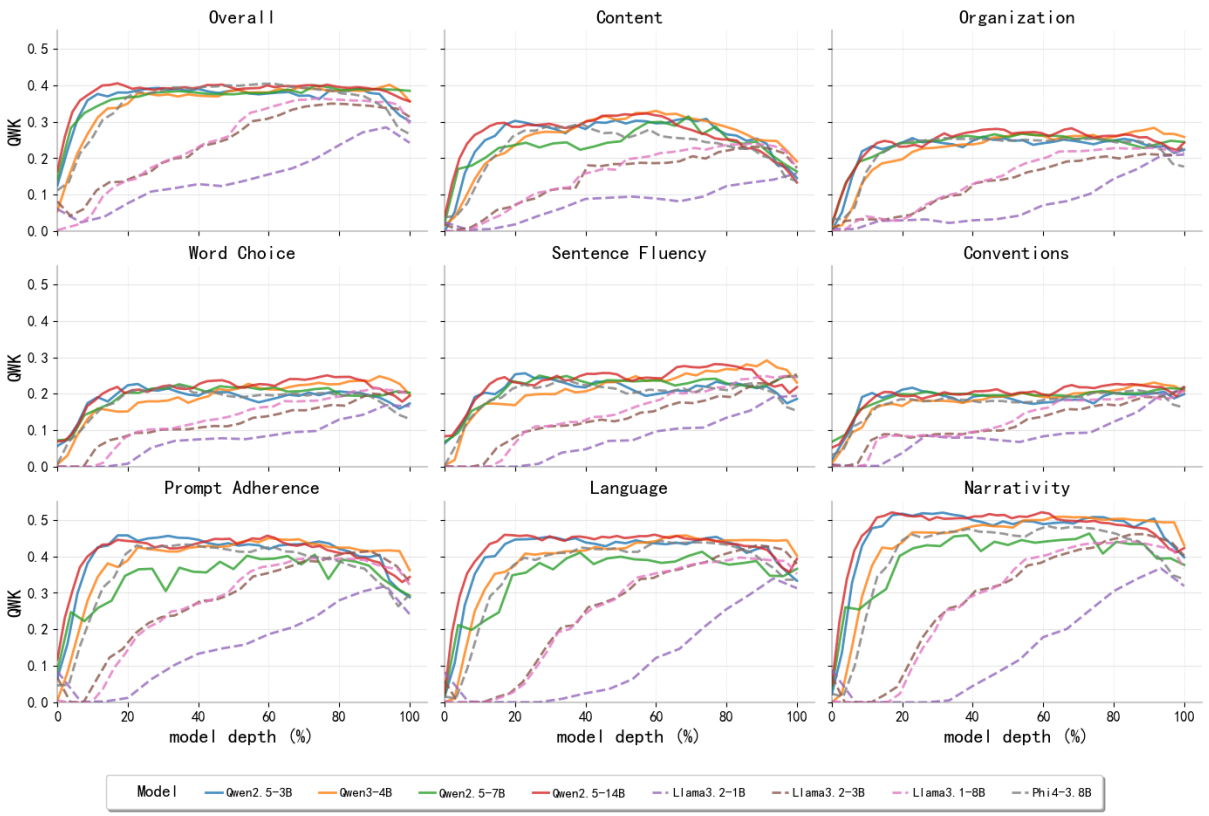

Figure 2: Average QWK scores of linear probes on ASAP++ under cross-prompt settings. Each subplot corresponds to an essay trait and shows probe performance across layers for different models.

activations onto corresponding feature directions (Mikolov et al., 2013; Olah et al., 2020; Elhage et al., 2022; Gurnee and Tegmark, 2023).

LLM Prompt Robustness. We analyze the sensitivity of essay quality representations to LLM prompt design. In practice, models are typically provided with both the essay and task instructions, and prior work has shown that prompting strategies can affect scoring performance (Stahl et al., 2024; Lee et al., 2024). We consider three prompt variants: Essay (essay only), Task (instructions + essay), and Chain-of-Thought (CoT) (instructions eliciting step-by-step reasoning + essay) (see Appendix D). We evaluate these strategies using Llama-8B on ASAP++ for overall score prediction. As shown in Figure 3, prompt variations lead to only marginal differences in essay quality representations. In most cases, CoT yields faster convergence and earlier saturation of probe performance, suggesting that explicit reasoning instructions may better elicit the model's latent scoring knowledge. However, for Prompt 8, which contains longer essays, CoT performs worst, followed by Task. This may be due to the additional instructional text introducing noise for longer inputs, which interferes with the formation of stable essay quality representations.

5 Discussion and Analysis

5.1 Cross-Prompt Generalization

The previous section demonstrated that both overall and trait-specific essay scores can be linearly reconstructed from internal activations of later LLM layers. However, this result alone does not imply that the model explicitly represents essay quality in the directions identified by the probe, as the probe may instead exploit linear combinations of more primitive features already present in the representations (Gurnee and Tegmark, 2023).

To evaluate cross-prompt generalization, we retrain the linear probe on the ASAP++ dataset using the same cross-prompt partitioning strategy as in prior work (Ridley et al., 2020; Li and Ng, 2024). Specifically, for each target prompt, all remaining prompts are used as training data, and evaluation is performed on the held-out prompt.

As shown in Figure 2, the resulting performance trends closely match those in Figure 1. The observed performance drop relative to in-prompt training is likely attributable to differences in scoring

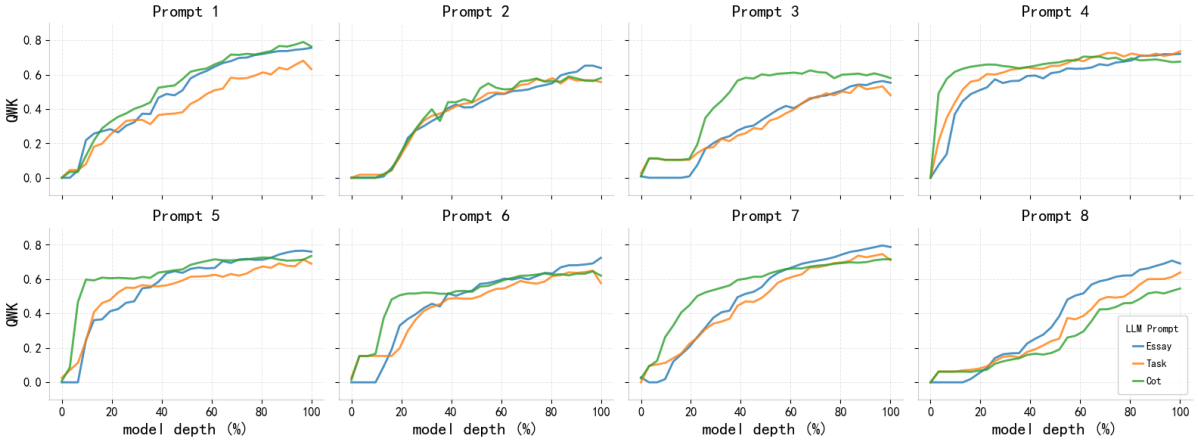

Figure 3: QWK scores of linear probes trained on overall essay scores in ASAP++ using Llama-3.1-8B-Instruct. Each subplot corresponds to an essay prompt and shows probe performance across three prompting strategies.

rubrics and prompt-specific distribution shifts. Importantly, probe performance remains substantially above chance (QWK = 0), indicating that a non-trivial portion of the essay quality signal is shared and linearly accessible across diverse prompts, despite prompt-dependent variation in how it is encoded.

5.2 Dimensionality Reduction

Although the probes we employ are linear, they operate in the full hidden dimensionality d_{model} (ranging from 2048 to 5120 for models with 1B to 14B parameters), which still allows for non-trivial capacity and potential memorization. As an additional robustness check, we use Principal Component Analysis (PCA) (Shlens, 2014) to project the activation space onto its top k principal components and train linear probes in this reduced subspace, thereby reducing the number of parameters by 2–3 orders of magnitude.

Figure 4 reports performance of probes trained to predict overall essay scores on the ASAP++ dataset across varying values of k, and compares them with full-dimensional probes. Results for Spearman correlation (see Appendix F) show that these coefficients increase more rapidly with k than QWK. This difference is expected, as Spearman correlation depends only on the rank ordering of predictions, whereas QWK additionally penalizes deviations in absolute score calibration. Overall, these results suggest that low-dimensional projections already capture substantial rank-relevant information about essay quality, while higher-dimensional components appear more important for improving calibration of absolute score predictions.

5.3 Essay Scoring Neurons

While the previous experimental results are informative, they provide only indirect evidence and do not establish whether the LLMs explicitly utilize the feature directions identified by the probes. To address this more directly, we identify individual neurons whose input or output weight vectors exhibit high cosine similarity with the probe-derived feature directions. Specifically, we focus on the overall score prediction for Prompt 1 in the ASAP++ dataset and compute the Spearman correlation between ground-truth scores and neuron activation values.

As shown in Figure 5, projecting the activation data onto the weights of these most similar neurons reveals that certain individual neurons are themselves highly sensitive to overall essay scores. In other words, some neurons can serve as effective standalone feature probes. Notably, neurons identified as most relevant for Prompt 1 also transfer to other prompts, maintaining substantial Spearman correlations, which suggests a degree of cross-prompt consistency in these representations.

If feature directions learned by supervised linear probes approximate the upper bound of the model's linearly decodable essay-related information, then the performance of individual neurons can be viewed as a lower bound. It is important to note that such features are generally expected to be distributed across multiple neurons in a superpositioned manner, making single-neuron analysis inherently limited (Elhage et al., 2022). Nevertheless, the existence of individual neurons, learned solely via the next-token prediction objective, that align with essay-scoring behavior provides evidence that

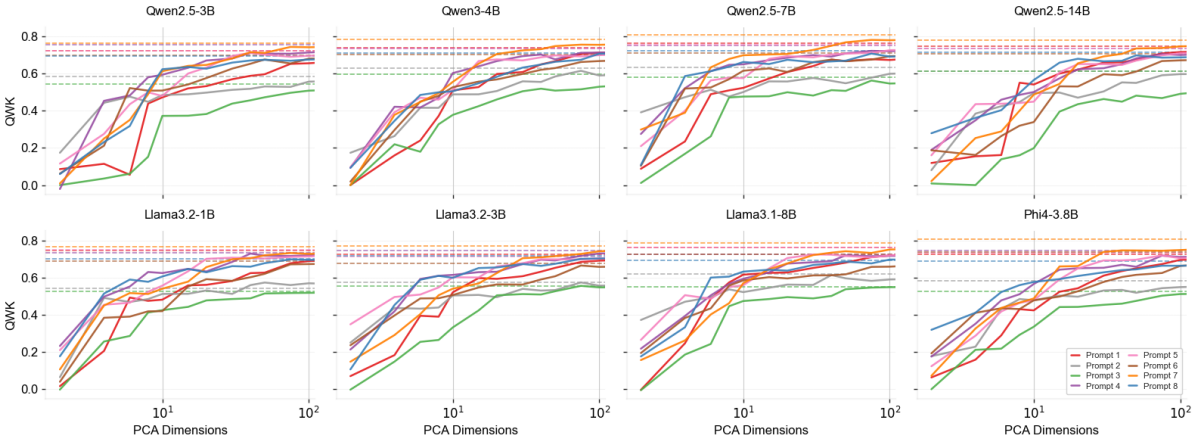

Figure 4: QWK scores across PCA dimensionality settings for each model. Dotted lines denote probes trained on full-dimensional activations.

the model encodes and utilizes features related to essay quality. We further conduct neuron intervention experiments (see Appendix G), which suggest that these neurons are more sensitive to intervention and play a more prominent role in the model's essay scoring behavior than typical neurons.

5.4 Neuron Distribution

Following the identification of neurons critical to essay scoring, we further investigate how internal processing varies across essay prompts by analyzing their distribution across network layers. On the ASAP++ dataset using Llama-3.1-8B-Instruct, Qwen2.5-7B-Instruct, and Qwen2.5-14B-Instruct, we observe a consistent distribution pattern for these neurons (see Appendix H). Specifically, for shorter essays (Prompts 3–6), the neurons are predominantly located in earlier to middle layers, whereas for longer essays (Prompts 1, 2, 7, and 8), they are more frequently found in middle to later layers. In addition, larger models tend to exhibit earlier emergence of these neurons compared to smaller models.

To further examine the point at which this distribution shifts with respect to text length, we conduct a finer-grained analysis on Prompt 8. We partition essays into 100-word bins to study how neuron distribution varies across different length ranges. To ensure sufficient samples per bin, we additionally generate essays using LLMs (see Appendix I). As shown in Figure 6, the distribution of essay scoring neurons begins to shift toward later layers once text length reaches approximately 200 words, consistently across all three models.

Interestingly, this shift appears to align with cognitive accounts of reading under increased load. For humans, longer texts introduce extended syntactic dependencies, increasing working memory demands and requiring deeper integration of information (Gibson, 1998). Similarly, in neural networks, earlier layers tend to capture local and syntactic patterns, while deeper layers are more involved in long-range and discourse-level integration (Tenney et al., 2019). From this perspective, the observed shift toward deeper layers for longer essays suggests that the model may adaptively recruit higher-layer computations to accommodate increased integration demands associated with increased essay length.

6 Conclusion

In this work, we investigate how large language models internally represent essay quality for automated essay scoring. Through extensive probing experiments, we show that both overall and trait-specific essay scores can be effectively decoded from hidden representations using simple linear probes, with essay quality information becoming increasingly accessible in deeper layers. While larger models tend to encode such information earlier in the network, final-layer performance remains broadly comparable across model families. Furthermore, nonlinear probes provide only marginal improvements over linear ones, suggesting that essay quality information is largely linearly decodable from LLM representations.

We further demonstrate that these representations are robust across prompting strategies and partially transferable across essay prompts, despite differences in scoring rubrics and prompt-specific

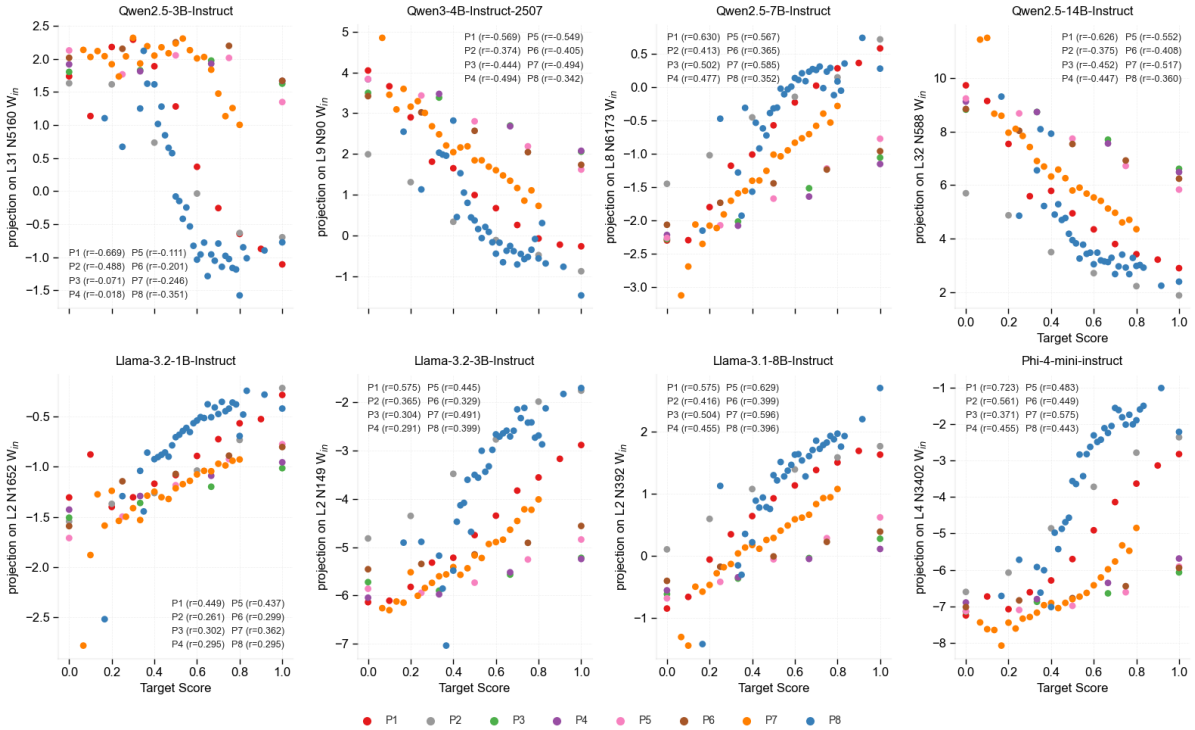

Figure 5: Essay scoring neurons in each model. Spearman correlations between neuron-weight projections and true essay scores are shown for each ASAP++ prompt. Each point denotes the average projection value for a target score.

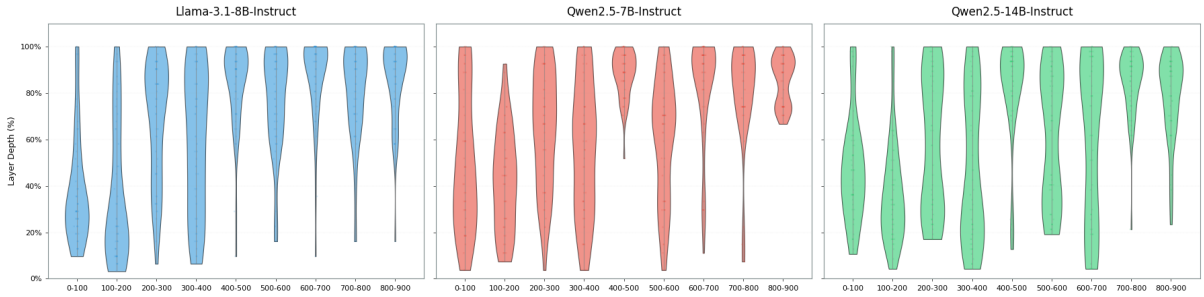

Figure 6: Distribution of the top 50 essay scoring neurons for the overall score of Prompt 8 in the ASAP++ dataset across different essay length intervals and models.

distribution shifts. Dimensionality reduction experiments additionally show that low-dimensional subspaces already preserve substantial rank-relevant information about essay quality, indicating that these signals are not solely dependent on high-capacity probe parameterization.

Beyond representation-level analysis, we identify individual neurons whose activations strongly correlate with essay scores and whose weight vectors align with probe-derived feature directions. Neuron intervention experiments further suggest that these neurons play a more prominent role in the model's essay scoring behavior than typical neurons. Moreover, we observe systematic shifts in the layer-wise distribution of essay scoring neurons as essay length increases, with longer essays relying more heavily on deeper layers. This pattern suggests that LLMs may recruit deeper computations to accommodate the increased integration demands associated with longer text inputs.

Overall, our findings provide evidence that LLMs encode structured and linearly accessible representations related to essay quality, extending beyond superficial statistical cues. More broadly, this work contributes toward bridging black-box performance and mechanistic interpretability in AES. Future work may explore how these representations and neurons can be leveraged to improve scoring robustness, controllability, and interpretability in educational applications.

Limitations

While this study provides insights into the internal mechanisms of LLM-based AES, several limitations remain.

(i) **Limited Model Scale:** Our experiments focused on open-source LLMs, ranging from 1B to 14B parameters, and did not include large-scale commercial models such as GPT-4 or Llama-3.1-70B-Instruct. Since many capabilities emerge with scale, it remains unclear whether our findings generalize to substantially larger models.

(ii) **Limited Language Coverage:** Experiments were conducted on two English datasets (ASAP++ and CSEE) and one Portuguese dataset (ENEM). Although the results suggest some degree of cross-lingual transferability, evaluating only one non-English language limits the generalizability of our conclusions. Essay scoring criteria may vary across cultural contexts, writing conventions, and educational systems, requiring broader multilingual evaluation.

(iii) **Limited Mechanistic Analysis:** Although we identified individual "essay scoring neurons", feature superposition may limit the interpretability of single-neuron analysis. In addition, our intervention experiments were restricted to individual neurons and therefore do not capture potential interactions among multiple neurons or higher-level scoring circuits.

References

Guillaume Alain and Yoshua Bengio. 2018. Understanding intermediate layers using linear classifier probes, 2018. *URL https://arxiv. org/abs/1610.01644*, 1610.

Dimitrios Alikaniotis, Helen Yannakoudakis, and Marek Rei. 2016. Automatic text scoring using neural networks. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 715–725, Berlin, Germany. Association for Computational Linguistics.

Yonatan Belinkov. 2022a. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Yonatan Belinkov. 2022b. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

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

Yuan Chen and Xia Li. 2023. PMAES: Prompt-mapping contrastive learning for cross-prompt automated essay scoring. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1503, Toronto, Canada. Association for Computational Linguistics.

Jinwei Chi, Ke Wang, Yu Chen, Xuanye Lin, and Qiang Xu. 2025. Activations as features: Probing llms for generalizable essay scoring representations. *Preprint*, arXiv:2512.19456.

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

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.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. 2022. Toy models of superposition. *Preprint*, arXiv:2209.10652.

Juan Escalante, Austin Pack, and Alex Barrett. 2023. Ai-generated feedback on writing: Insights into efficacy and enl student preference. *International Journal of Educational Technology in Higher Education*, 20(1):57.

Allyson Ettinger, Ahmed Elgohary, and Philip Resnik. 2016. Probing for semantic evidence of composition by means of simple classification tasks. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 134–139, Berlin, Germany. Association for Computational Linguistics.

Edward Gibson. 1998. Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68(1):1–76.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuanzhi Li. 2023. Textbooks are all you need. *Preprint*, arXiv:2306.11644.

Wes Gurnee and Max Tegmark. 2023. Language models represent space and time. *arXiv preprint arXiv:2310.02207*.

Jieun Han, Haneul Yoo, Junho Myung, Minsun Kim, Hyunseung Lim, Yoonsu Kim, Tak Yeon Lee, Hwajung Hong, Juho Kim, So-Yeon Ahn, and 1 others. 2024. Llm-as-a-tutor in efl writing education: Focusing on evaluation of student-llm interaction. In *Proceedings of the 1st Workshop on Customizable NLP: Progress and Challenges in Customizing NLP for a Domain, Application, Group, or Individual (CustomNLP4U)*, pages 284–293.

Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. 2009. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743, Hong Kong, China. Association for Computational Linguistics.

Sanwoo Lee, Yida Cai, Desong Meng, Ziyang Wang, and Yunfang Wu. 2024. Unleashing large language models' proficiency in zero-shot essay scoring. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 181–198, Miami, Florida, USA. Association for Computational Linguistics.

Shengjie Li and Vincent Ng. 2024. Conundrums in cross-prompt automated essay scoring: Making sense of the state of the art. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7661–7681, Bangkok, Thailand. Association for Computational Linguistics.

Sandeep Mathias and Pushpak Bhattacharyya. 2018. Asap++: Enriching the asap automated essay grading dataset with essay attribute scores. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

Microsoft, :, Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, Dong Chen, Dongdong Chen, Junkun Chen, Weizhu Chen, Yen-Chun Chen, Yi ling Chen, Qi Dai, and 57 others. 2025. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *Preprint*, arXiv:2503.01743.

Tomáš Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 746–751.

Eleni Miltsakaki and Karen Kukich. 2004. Evaluation of text coherence for electronic essay scoring systems. *Natural Language Engineering*, 10(1):25–55.

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.

Farah Nadeem, Huy Nguyen, Yang Liu, and Mari Ostendorf. 2019. Automated essay scoring with discourse-aware neural models. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 484–493, Florence, Italy. Association for Computational Linguistics.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. 2020. Zoom in: An introduction to circuits. *Distill*, 5(3):e00024–001.

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. 2021. Probing the probing paradigm: Does probing accuracy entail task relevance? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3363–3377.

R Ridley, L He, X Dai, S Huang, and J Chen. 2020. Prompt agnostic essay scorer: a domain generalization approach to cross-prompt automated essay scoring. arxiv. *arXiv preprint arXiv:2008.01441*.

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

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in bertology: What we know about how bert works. *Preprint*, arXiv:2002.12327.

Lawrence M Rudner and Tahung Liang. 2002. Automated essay scoring using bayes' theorem. *The Journal of Technology, Learning and Assessment*, 1(2).

Tilman Räuker, Anson Ho, Stephen Casper, and Dylan Hadfield-Menell. 2023. Toward transparent ai: A survey on interpreting the inner structures of deep neural networks. *Preprint*, arXiv:2207.13243.

Takumi Shibata and Yuichi Miyamura. 2025. Lces: Zero-shot automated essay scoring via pairwise comparisons using large language models. *arXiv preprint arXiv:2505.08498*.

Jonathon Shlens. 2014. A tutorial on principal component analysis. *Preprint*, arXiv:1404.1100.

Igor Cataneo Silveira, André Barbosa, and Denis Deratani Mauá. 2024a. 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.

Igor Cataneo Silveira, André Barbosa, and Denis Deratani Mauá. 2024b. 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, Santiago de Compostela, Galicia/Spain. Association for Computational Linguistics.

M Stahl, L Biermann, A Nehring, and H Wachsmuth. 2024. Exploring llm prompting strategies for joint essay scoring and feedback generation. arxiv.

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.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. Bert rediscovers the classical nlp pipeline. *Preprint*, arXiv:1905.05950.

Masaki Uto, Yikuan Xie, and Maomi Ueno. 2020. Neural automated essay scoring incorporating handcrafted features. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6077–6088, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Jiong Wang, Qing Zhang, Jie Liu, Xiaoyi Wang, Mingying Xu, Liguang Yang, and Jianshe Zhou. 2025. Making meta-learning solve cross-prompt automatic essay scoring. *Expert Systems with Applications*, page 126710.

Changrong Xiao, Wenxing Ma, Qingping Song, Sean Xin Xu, Kunpeng Zhang, Yufang Wang, and Qi Fu. 2025. Human-ai collaborative essay scoring: A dual-process framework with llms. In *Proceedings of the 15th international learning analytics and knowledge conference*, pages 293–305.

Kevin P Yancey, Geoffrey Laflair, Anthony Verardi, and Jill 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.

Ruosong Yang, Jiannong Cao, Zhiyuan Wen, Youzheng Wu, and Xiaodong He. 2020. Enhancing automated essay scoring performance via fine-tuning pre-trained language models with combination of regression and ranking. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1560–1569, Online. Association for Computational Linguistics.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A new dataset and method for automatically grading esol texts. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 180–189.

Chunyun Zhang, Jiqin Deng, Xiaolin Dong, Hongyan Zhao, Kailin Liu, and Chaoran Cui. 2025. Pairwise dual-level alignment for cross-prompt automated essay scoring. *Expert Systems with Applications*, 265:125924.

Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. 2024. Tinyllama: An open-source small language model. *Preprint*, arXiv:2401.02385.

A Datasets

ASAP++ is an extension of the ASAP[4] dataset which comprises 12,978 essays written by students in grades 7-10. These essays are produced in response to eight different prompts, which vary in genre and scoring criteria. Each essay has an overall score and 8 trait scores. The descriptive statistics of ASAP++ are outlined in Table 1.

CSEE is carefully curated in collaboration with 29 high schools in China, encompassing a total of 13,372 student essays responding to two distinct prompts used in final exams. Each essay has an overall score and 3 trait scores. The evaluation of these essays was carried out by highly experienced English teachers following the scoring guidelines of the Chinese National College Entrance Examination. Scoring was comprehensively assessed across three critical dimensions: Content, Language, and Structure, with an Overall Score ranging from 0 to 20. The descriptive statistics of CSEE are outlined in Table 2.

ENEM comprises argumentative essays written by Brazilian students in response to a variety of socially relevant prompts. Collected from public online platforms simulating the Brazilian National High School Exam (ENEM), these essays are annotated following the official ENEM scoring rubric. The rubric evaluates five aspects (C1–C5, C1: fluency, C2: writing style, C3: argumentation quality, C4: proper use of textual connectors, and C5: quality of the solution to the prompt's problem), each scored on a scale from 0 to 200 in increments of 20, resulting in a total score out of 1000.

The dataset is divided into two subsets: **Source A**, with 386 essays including full supporting texts validated by experts, serves as a high-quality benchmark; **Source B**, with 3,200 essays, is mainly used for model pretraining and augmentation (Silveira et al., 2024b). We only use source A for experiments.

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

B Experiments Settings

To ensure reproducibility, all experiments are conducted with a fixed random seed of 42. For the linear probe, we use ridge regression with built-in cross-validation (`RidgeCV`), searching the regularization strength α over 12 logarithmically spaced values in the range $[10^3, 10^{4.5}]$, while retaining the cross-validation scores.

For the nonlinear probe, we adopt a single-hidden-layer multi-layer perceptron with a hidden dimension of 256. The multi-layer perceptron is trained using AdamW with mean squared error loss, a fixed learning rate of 1×10^{-3}, and a batch size of 4096. We tune weight decay over $\{0.01, 0.03, 0.1, 0.3\}$. Training is run for up to 200 epochs with early stopping based on a 10% held-out validation split from the training set; training is stopped if the validation loss does not improve for 10 consecutive epochs.

C Linear vs. Nonlinear Probes

In Table 3, we present the average QWK scores of linear and nonlinear probes on ASAP++ essay traits, averaged across all prompts and models at full (100%) layer depth.

D LLM Prompting Templates

We present two examples based on the prompt template, corresponding to the Task prompt and the CoT prompt, respectively.

D.1 Task Prompt

> As an English teacher, your primary responsibility is to evaluate the writing quality of essays written by middle school students, with evaluation measured on a scale from {min_score} to {max_score}.
> [Essay]
> {essay}
> (end of [Essay])

D.2 CoT Prompt

> As an English teacher, your primary responsibility is to evaluate the writing quality of essays written by middle school students. During the assessment process, you will be provided with an essay. First, you should provide comprehensive and concrete feedback that is closely linked to the content of the essay. It is essential to avoid offering generic remarks that could be applied to any piece of writing. To create a compelling evaluation for both the student and fellow experts, you should reference specific content of the essay to substantiate your assessment. Next, your evaluation should culminate in assigning an overall score to the student's essay, measured on a scale from {min_score} to {max_score}, where higher score should reflect a higher level of writing quality. It's crucial to tailor your evaluation criteria to be well-suited for middle school level writing, taking into account the developmental stage and capabilities of these students.
> [Essay]
> {essay}
> (end of [Essay])

E Results on CSEE and ENEM

Figures 8 and 9 present the linear probe results on the CSEE and ENEM datasets, respectively. Both exhibit trends similar to those observed on the ASAP++ dataset (see Figure 1).

Notably, as shown in Figure 9, the probe curve exhibits greater fluctuations and a lower peak performance compared to the English-language datasets. This discrepancy may be attributed to the relatively smaller dataset size or the limited coverage of Portuguese-language training data.

F Spearman Correlation Results under Dimensionality Reduction

Figure 10 depicts the Spearman correlation between predictions of probes trained on activations projected onto the top k principal components and ground-truth scores. Each subplot shows results across different dimensionality reduction settings for each model, with the Spearman correlation of probes trained on full-dimensional activations shown as dotted lines.

G Neuron Intervention

To better understand the role of essay scoring neurons, we investigate the effect of intervening on a single essay scoring neuron ($L2.N392.W_{\text{in}}$, which exhibits a Spearman correlation of 0.574 with Prompt 1 in ASAP++) in the Llama-3.1-8B-Instruct model.

Given a prompting template $\mathcal{T}$ (see Figure 12), we fix the activation of this neuron across all tokens and sweep over a range of constant values, while tracking the prediction probabilities of the top-10 tokens (with `do_sample=False`). As shown in Figure 7, increasing the fixed activation leads to the largest increase in the weighted sum of essay-scoring neurons compared to three randomly selected neurons, while decreasing it produces the most pronounced decline. These results further suggest that this neuron is sensitive to intervention and plays a more prominent role in the model's essay scoring behavior than typical neurons.

H Neuron Distribution

Figure 11 illustrates the distribution of the top 50 essay scoring neurons across different traits and essay prompts in the ASAP++ dataset, across models of varying architectures and sizes. All models demonstrate a consistent distribution across different essay prompts.

I Data Augmentation

To enable a statistically robust analysis of essay scoring neuron distributions across different essay lengths, we augmented Prompt 8 of the ASAP++ dataset with synthetic essays. The original dataset was insufficient to support reliable analysis at 100-word intervals, so we used the `gpt-5.4-mini` model to generate additional essays until each interval contained at least 100 samples. The temperature was set to 0.7 to encourage output diversity. The detailed generation prompt is provided below.

System Prompt

You are a helpful assistant and an expert in English language assessment. You will generate essays based on a given topic and score them according to the provided rubric.

User Prompt

Essay Topic: We all understand the benefits of laughter. For example, someone once said, "Laughter is the shortest distance between two people." Many other people believe that laughter is an important part of any relationship. Tell a true story in which laughter was one element or part.

Your Task:
1. **Write the Essay:** Generate a short, true story based on the topic above. The story should be written from the perspective of a 10th-grade (Grade 10) student. The length must be between {min} and {max} words. (The average essay length for this topic is approximately 650 words.)
2. **Score the Essay:** After writing, score the essay on a scale of 0 to 60 points. Use the following four criteria, each scored from 1 to 6 points, and note that Conventions has double weight:
- **Content (1-6 points):** This category assesses the core substance and clarity of a written piece. It focuses on how clear, focused, and well-supported the main ideas are.
- **Organization (1-6 points):** This category assesses the structure and flow of a piece of writing. It focuses on how logically and smoothly the ideas are ordered and connected for the reader.
- **Sentence Fluency (1-6 points):** This category assesses the rhythm, flow, and craftsmanship of sentences. It focuses on how smoothly and pleasantly the writing reads aloud, and the variety in sentence structure.
- **Conventions (1-6 points, double weight):** This category assesses the technical correctness of the writing, including grammar, punctuation, spelling, and capitalization. It focuses on how well the writer controls standard language rules to ensure clear communication.

Scoring Formula: Total = Content + Organization + Sentence Fluency + (2 × Conventions)

Important: The essay is short ({min}-{max} words), so scores should not be too high, and typically scores below {max_score} points.

Output Format Requirements:
You must output **ONLY** a valid JSON object, nothing else. The JSON must have exactly two keys:
{{
"essay": "Your generated essay text here..."
"score": 12
}}

J Base Model vs Instruction Tuned Model

We investigate whether instruction tuning the base model enhances its capability to construct essay quality representations. As illustrated in Figure 13, the performance of probes trained on the Base model and the Instruction tuned model exhibits negligible differences. This finding demonstrates that the model's ability to construct representations of essay originates from the pretraining stage.

Prompt ID	No. of Essays	Avg. Len.	Genre	Attributes	Score Range	
					Overall	Attribute
1	1,783	418	ARG	Cont, Org, WC, SF, Conv	2 - 12	1 - 6
2	1,800	427	ARG	Cont, Org, WC, SF, Conv	0 - 6	1 - 6
3	1,726	123	RES	Cont, PA, Lan, Nar	0 - 3	0 - 3
4	1,772	105	RES	Cont, PA, Lan, Nar	0 - 3	0 - 3
5	1,805	140	RES	Cont, PA, Lan, Nar	0 - 4	0 - 4
6	1,800	172	RES	Cont, PA, Lan, Nar	0 - 4	0 - 4
7	1,569	199	NAR	Cont, Org, Conv	0 - 30	0 - 6
8	723	701	NAR	Cont, Org, WC, SF, Conv	0 - 60	2 - 12

Table 1: Statistics of ASAP++. Abbreviations: Cont (Content), Org (Organization), WC (Word Choice), SF (Sentence Fluency), Conv (Conventions), PA (Prompt Adherence), Lan (Language), Nar (Narrativity). 'Avg. Len.' refers to the average essay length in tokens, calculated using the NLTK toolkit (https://www.nltk.org/).

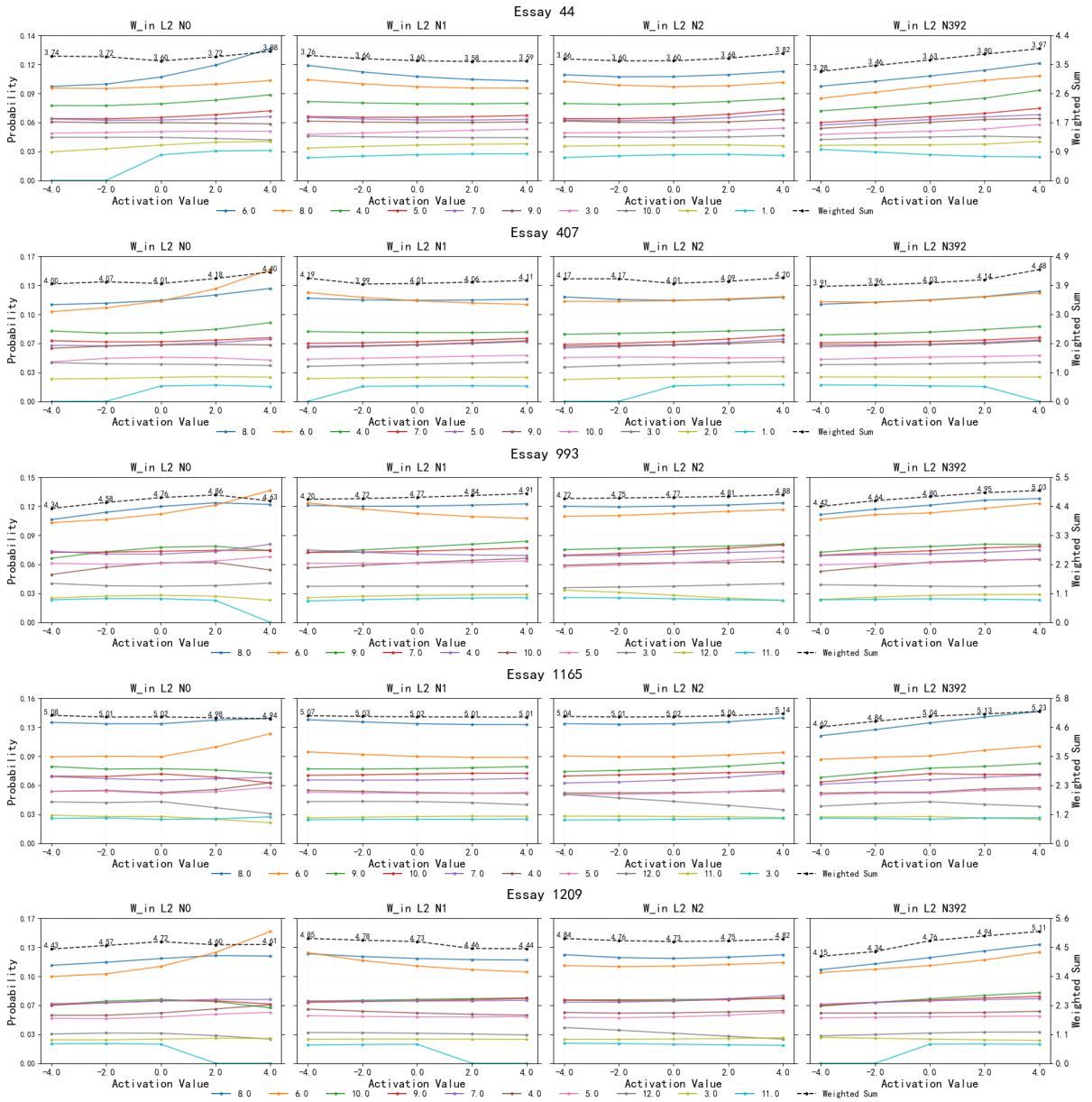

Figure 7: When the essay scoring neuron ($L2.N392.W_{in}$) is fixed to specific values, the prediction results for five different essays from prompt 1 of the ASAP dataset are compared with the prediction results from three random neurons in the same layer (L2.[0-2]) of the Llama-3.1-8B-Instruct model. We also calculate the weighted sum of top 10 tokens when the essay scoring neuron is fixed to different specific values.

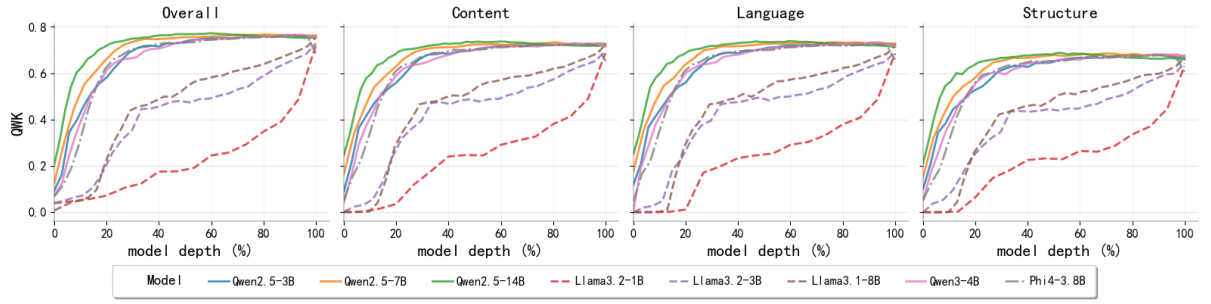

Figure 8: Average QWK scores of linear probes trained on CSEE across all essay prompts. Each subplot corresponds to a essay trait and shows probe performance across layers for different models.

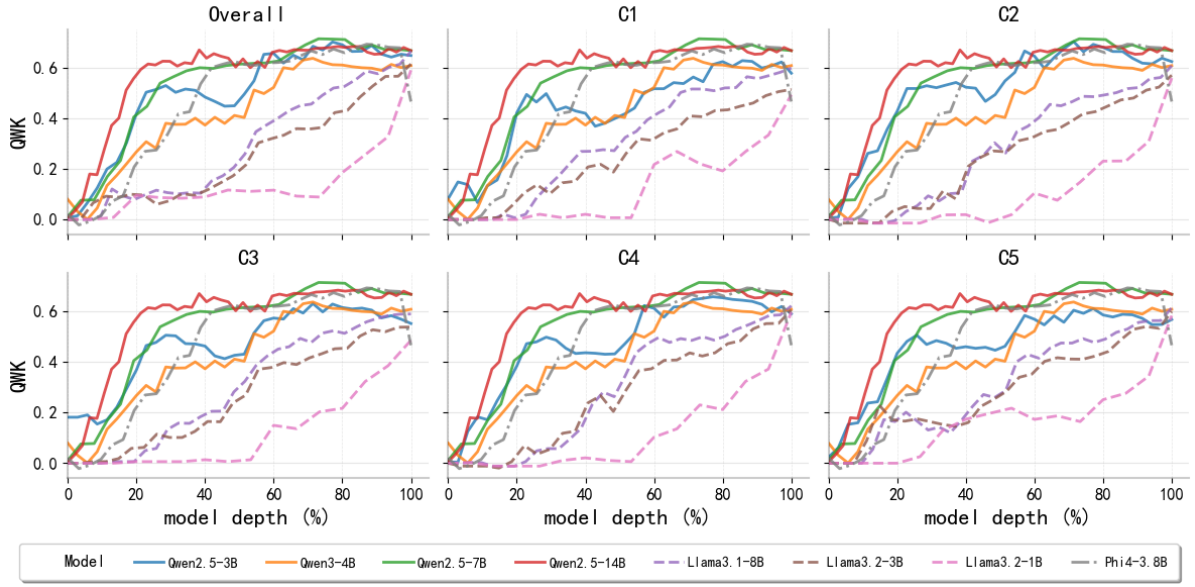

Figure 9: Average QWK scores of linear probes trained on ENEM across all essay prompts. Each subplot corresponds to a essay trait and shows probe performance across layers for different models.

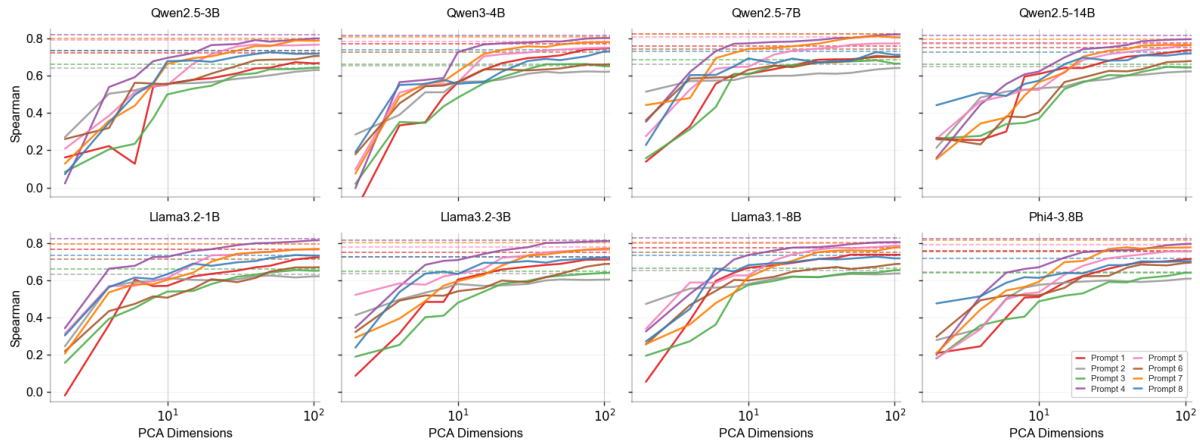

Figure 10: Spearman correlation between predictions of probes trained on activations projected onto the top k principal components and ground-truth scores. Each subplot shows results across different dimensionality reduction settings for each model, with the Spearman correlation of probes trained on full-dimensional activations shown as dotted lines.

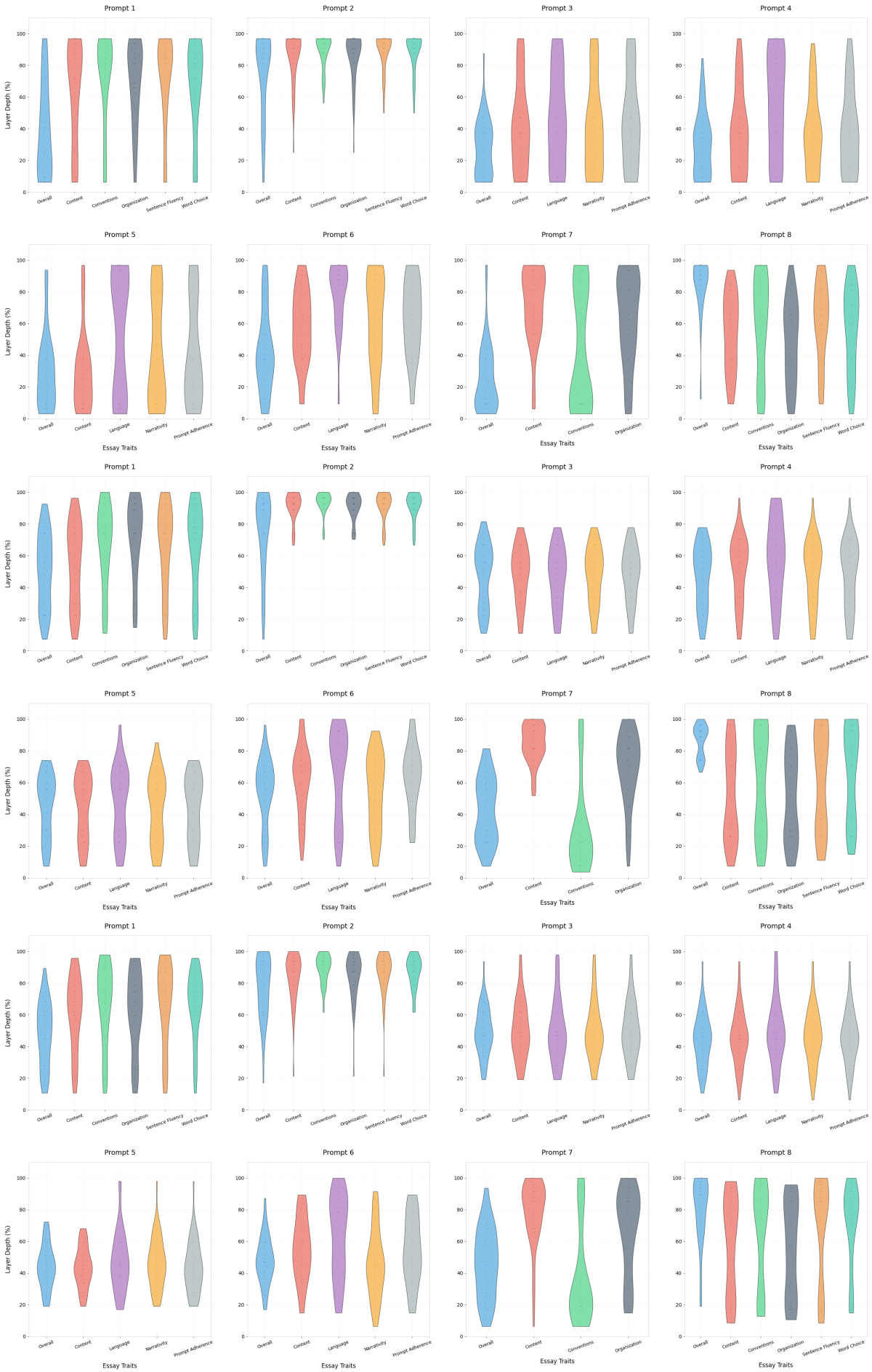

Figure 11: Distribution of the top 50 key neurons in the ASAP++ dataset, shown for different traits, essay prompts and models. Every two rows represent the results of a model, corresponding to Llama-3.1-8B-Instruct, Qwen2.5-7B-Instruct, and Qwen2.5-14B-Instruct, respectively.

Statistics of CSEE	
# of schools	29
# of essay prompts	2
# of student essays	13,372
avg. essay length	124.74
avg. Overall score	10.72
avg. Content score	4.13
avg. Language score	4.05
avg. Structure score	2.55

Table 2: Descriptive statistics of Chinese Student English Essay (CSEE) dataset (Xiao et al., 2025).

Prompt $\mathcal{T}$
Please score the following essay between 2 and 12 points, You only need to output the score. [Essay] {essay} (end of [Essay]) The Score is

Figure 12: Prompt template $\mathcal{T}$ used for neuron intervention experiments. Curly brackets { } denote placeholders to be completed.

Model	Probe	Overall	Cont	Org	WC	SF	Conv	PA	Lan	Nar	Avg.
Llama3.2-1B	Linear	0.682	0.632	0.516	0.571	0.534	0.527	0.659	0.617	0.641	0.598
	Nonlinear	0.691	0.637	0.570	0.541	0.544	0.540	0.663	0.639	0.657	0.609
Llama3.2-3B	Linear	0.685	0.627	0.549	0.583	0.593	0.558	0.657	0.622	0.641	0.613
	Nonlinear	0.653	0.620	0.544	0.536	0.576	0.555	0.631	0.616	0.634	0.596
Llama3.1-8B	Linear	0.704	0.644	0.545	0.572	0.585	0.541	0.668	0.621	0.651	0.615
	Nonlinear	0.667	0.631	0.529	0.520	0.564	0.539	0.586	0.588	0.646	0.585
Phi4-3.8B	Linear	0.620	0.591	0.518	0.565	0.517	0.528	0.612	0.581	0.604	0.571
	Nonlinear	0.644	0.617	0.552	0.569	0.468	0.496	0.613	0.565	0.574	0.567
Qwen2.5-3B	Linear	0.682	0.633	0.524	0.568	0.534	0.522	0.666	0.635	0.652	0.602
	Nonlinear	0.684	0.623	0.520	0.562	0.527	0.527	0.666	0.626	0.660	0.599
Qwen3-4B	Linear	0.705	0.644	0.544	0.583	0.561	0.544	0.663	0.635	0.650	0.614
	Nonlinear	0.692	0.643	0.579	0.542	0.602	0.594	0.653	0.636	0.635	0.620
Qwen2.5-7B	Linear	0.703	0.652	0.544	0.549	0.553	0.534	0.676	0.639	0.666	0.613
	Nonlinear	0.691	0.638	0.543	0.527	0.562	0.544	0.682	0.606	0.659	0.606
Qwen2.5-14B	Linear	0.700	0.646	0.552	0.570	0.583	0.552	0.674	0.640	0.646	0.618
	Nonlinear	0.643	0.602	0.521	0.594	0.590	0.566	0.652	0.613	0.615	0.599

Table 3: Average QWK scores of linear and nonlinear probes on ASAP++ essay traits, averaged across all prompts and models at full (100%) layer depth. Abbreviations: Cont (Content), Org (Organization), WC (Word Choice), SF (Sentence Fluency), Conv (Conventions), PA (Prompt Adherence), Lan (Language), Nar (Narrativity).

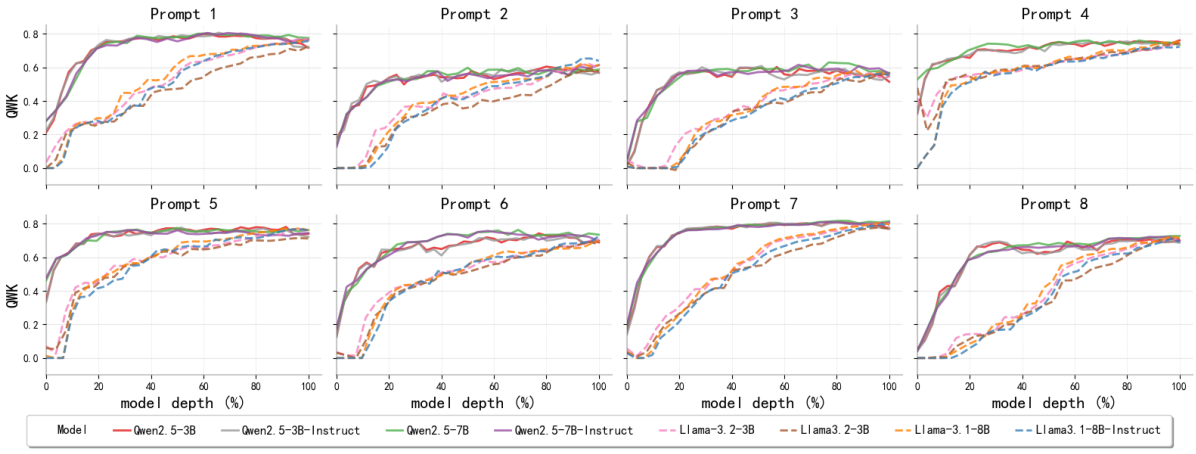

Figure 13: QWK scores of linear probes trained on the overall score of the ASAP++ dataset on each essay prompt and model.