Taibah at TAQEEM 2025: Leveraging GPT-4o for Arabic Essay Scoring

NADA ALMARWANI[1], **ALAA ALHARBI**[2], **SAMAH ALOUFI**[1]
[1]Department of AI and Data Science, CCSE, Taibah University
[2]Department of Information Systems, CCSE, Taibah University

(e-mail: nmarwani, alaharbi, slhebi@taibahu.edu.sa)

Abstract

This paper presents our system submitted to TAQEEM 2025, which designed to address two tasks: (A) holistic scoring and (B) trait-specific scoring. We propose a GPT-4o-based methodology that employs few-shot prompting to serve as a grader for both tasks. Specifically, for task A, we utilize prompt-based scoring criteria with exemplars to assess overall essay quality. For task B, we design trait-specific prompting schemes to capture fine-grained grading aspects. Our system attains substantial agreement on Task A (QWK = 0.75) and a mean QWK of 0.65 across traits for task B, outperforming the shared task baseline on both tasks.

1 Introduction

Evaluating student essays plays a critical role in assessing language proficiency and writing development, particularly in educational settings where writing is a core skill. However, traditional essay scoring is labor-intensive, costly, and liable to inter- and intra-rater inconsistencies caused by human subjectivity, bias, and rater characteristics such as severity or leniency (Uto and Okano, 2020). To address these challenges, Automated Essay Scoring (AES) systems have emerged as scalable and efficient alternatives. When effectively implemented, AES systems offer timely, objective, and consistent scoring, mitigating rater bias and supporting large-scale assessment contexts such as standardized examinations.

Recent advancements in natural language processing (NLP), particularly the emergence of generative large language models (LLMs) such as OpenAI's GPT-4 and Google's PaLM, have significantly enhanced the capabilities of AES systems. A notable advantage of LLMs is their ability to perform zero-shot and few-shot scoring with minimal supervision. Mizumoto and Eguchi (2023) demonstrated that generative models like ChatGPT can reliably assess essays using standardized rubrics, confirming their feasibility and effectiveness for AES tasks. In terms of validity and reliability, Pack et al. (2024) and Li and Liu (2024) showed that GPT-4 achieved substantial agreement with human raters on AES tasks. Moreover, LLMs can be prompted to evaluate essays either via traditional linguistic features or rubric-based criteria aligned with human judgment (Pack et al., 2024). Recent work highlights that prompting strategies play a critical role in aligning LLM-generated scores with human evaluations (Li and Liu, 2024; Liew and Tan, 2024).

The majority of studies that have exploited LLMs for essay scoring have concentrated on English-language essays (Pack et al., 2024; Liew and Tan, 2024; Yavuz et al., 2025; Katuka et al., 2024; Yang, 2024; Flodén, 2025), with limited studies exploring other languages such as Chinese (Feng et al., 2024), Japanese (Li and Liu, 2024), and Arabic (Ghazawi and Simpson, 2025). The scarcity of annotated essay datasets in Arabic, which hinders the development of effective AES systems for this language, reflects a broader challenge. To address this gap, the TAQEEM shared task[1] (Bashendy et al., 2025) invites researchers to develop automated scoring models for Arabic essays, evaluating both holistic and trait-specific performance in a cross-prompt setting. Inspired by the promising results of prior work on generative LLM-based essay scoring, we employ OpenAI's GPT-4o model to simulate expert grading of Arabic essays across both tasks. Our approach leverages carefully crafted rubric-guided prompts and few-shot exemplars to achieve consistent and interpretable scoring across diverse Arabic texts. We also conduct a concise error analysis quantifying over- and under-scoring.

[1] https://sites.google.com/view/taqeem-2025/home?authuser=0

Proceedings of The Third Arabic Natural Language Processing Conference, pages 989–997

2 Task Description

The TAQEEM benchmark aims to advance automated Arabic essay scoring under cross-prompt evaluation via two tasks.

Task A: (Holistic Scoring) requires a single score reflecting the overall essay quality. **Task B: (Trait-specific Scoring)** requires the model to produce a separate score for seven rubric traits: Relevance, Organization, Vocabulary, Style, Development, Mechanics, and Grammar. The dataset provided with the TAQEEM 2025 Shared task comprises 1,265 Arabic essays, divided into 425 essays for training and 840 for testing. Each essay was written in response to one of several prompts and annotated by human for both tasks. This setup assesses systems' ability to generalize across prompts while maintaining alignment with human judgments.

3 Methodology

The essay grading system developed in this study leverages OpenAI's GPT-4o model (Hurst et al., 2024) to simulate expert scoring of Arabic essays. A small set of human-scored examples—specifically, 20 representative training samples—is embedded directly in the prompt as exemplars to guide the model through the grading process, ensuring coverage of the full range of grades. These 20 examples are randomly selected from the training dataset across a range of score levels to ensure diversity and enhance the model's ability to generalize across varying levels of essay quality, while remaining within token constraints. Importantly, all 20 exemplars were sourced from a single training prompt. These examples are then used to evaluate the model on different prompts. This checks if the model can perform well beyond the specific prompt on which it was trained, showing its adaptability across various inputs.

For Task A, the prompt includes a rubric for evaluating essays written in Arabic. This rubric assesses six core dimensions: content clarity, linguistic correctness, structural organization, strength of arguments, stylistic quality, and adherence to word count requirements. The dimensions were derived directly from the task description to ensure relevance, rather than adopting the CAST rubric, which may not have aligned with the task's unique requirements. These evaluation criteria are expressed in natural language instructions, enabling the model to internalize the scoring logic without relying on a structured input format. The original Arabic prompt, its English translation, and the rubric structure are provided in Figure 1 in the Appendix.

For Task B, we designed a structured prompt that also guides the model in evaluating Arabic student essays, simulating the behavior of an expert Arabic language teacher. This prompt instructs the model to score essays according to a detailed, criterion-referenced rubric covering seven dimensions: Relevance (max 2 points), Organization (max 5 points), Vocabulary (max 5 points), Style (max 5 points), Development (max 5 points), Mechanics (max 5 points), and Grammar (max 5 points)[2]. Each dimension is defined in natural language to ensure interpretability and consistent application of the scoring criteria. The original Arabic version of the prompt, as well as its English translation and associated rubric, are provided in Figure 2 in the Appendix.

The grading process is executed using OpenAI's API. Each essay, along with its corresponding prompt, is submitted to the GPT-4o model with a low temperature setting (0.1) to produce consistent and deterministic output.

4 Results

This section presents the performance comparison between the baseline system, which fine-tunes AraBERTv02 (Antoun et al., 2020) for automated essay scoring[3], and the proposed Taibah system for Task A and Task B, as detailed in Table 1 and Table 2, respectively. The evaluation was conducted using three key metrics: Quadratic Weighted Kappa (QWK), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE).

4.1 Task A: Holistic Scoring Results

As shown in Table 1, the Taibah system consistently outperforms the baseline in terms of QWK. For Test Prompt 9, our system achieved a QWK of 0.717, compared to the baseline's 0.608. This performance advantage remains evident in Test Prompt 10, where the QWK reached 0.784 versus the baseline's 0.670. The average QWK across both prompts was 0.751 for our system, demonstrating a notable improvement over the baseline's average of 0.639 and indicating stronger alignment with human judgments. This +0.112 increase in QWK reflects a substantial gain in rater agreement,

[2]https://sites.google.com/view/taqeem-2025

[3]https://gitlab.com/bigirqu/taqeem2025

System	Prompt 9			Prompt 10			Average		
	QWK	MSE	RMSE	QWK	MSE	RMSE	QWK	MSE	RMSE
Baseline	0.608	33.148	5.757	0.670	24.862	4.986	0.639	29.005	5.372
Taibah	0.717	31.281	5.593	0.784	19.595	4.427	**0.751**	**25.438**	**5.010**

Table 1: Performance comparison between Baseline and Taibah system for Task A. **Bold values** indicate superior performance.

especially considering that QWK values above 0.75 are often interpreted as indicating substantial to near-perfect agreement (Landis and Koch, 1977). We attribute this improvement, in part, to the rubric-aligned prompt design and the inclusion of diverse exemplars, which helped guide the model's scoring decisions.

In terms of error metrics, where lower values indicate better performance, the Taibah system also demonstrated superior performance. For Test Prompt 9, it achieved an MSE of 31.281 and RMSE of 5.593, outperforming the baseline's MSE of 33.148 and RMSE of 5.757. The advantage was even more pronounced for Test Prompt 10, with our system achieving an MSE of 19.595 and RMSE of 4.427 compared to the baseline's 24.862 (MSE) and 4.986 (RMSE). On average, the Taibah system maintained lower error rates (MSE: 25.438; RMSE: 5.010) than the baseline (MSE: 29.005; RMSE: 5.372), further validating its enhanced performance in Task A. These consistent reductions in MSE and RMSE across both prompts suggest that the few-shot GPT-4o-based approach generalizes well across different essay topics, despite the cross-prompt evaluation setting.

4.2 Task B: Trait-specific Scoring Results

As shown in Table 2, our system consistently outperformed the baseline across all traits and both test prompts in terms of QWK, demonstrating stronger alignment with human judgments in trait-level scoring. For Prompt 9, the most notable improvements were observed in Relevance (Taibah: 0.586 vs. Baseline: 0.127) and Development (Taibah: 0.727 vs. Baseline: 0.410). Similar trends were seen for Prompt 10, with substantial gains in Relevance (Taibah: 0.538 vs. Baseline: 0.182) and Mechanics (Taibah: 0.686 vs. Baseline: 0.468). On average, our system achieved higher QWK scores across all traits, with the largest improvements in Development (Taibah: 0.703 vs. Baseline: 0.458) and Relevance (Taibah: 0.562 vs. Baseline: 0.155). These gains are particularly important for scoring dimensions that are often challenging for automated systems, such as content relevance and argument development, suggesting that the prompt structure effectively guided the model's understanding of nuanced writing features.

Our system also demonstrated superior performance in error metrics, with lower MSE and RMSE values indicating better predictive accuracy. For Prompt 9, the system achieved notable reductions in both metrics across all traits. For example, in Relevance, MSE dropped from 0.514 (Baseline) to 0.221 (Taibah), and RMSE from 0.717 to 0.471. Similar improvements were observed in Development, where MSE decreased from 1.174 to 0.717 and RMSE from 1.083 to 0.847.

For Prompt 10, the system maintained its performance advantage. In Relevance, MSE decreased from 0.340 to 0.231 and RMSE from 0.584 to 0.481. Vocabulary also saw notable reductions, with MSE dropping from 0.964 to 0.669 and RMSE from 0.982 to 0.818. These results reflect the model's ability to generalize across writing prompts, a key challenge in cross-prompt AES settings.

Overall, our system achieved consistently lower average MSE and RMSE values across all traits. The most significant reductions were found in Relevance (MSE: 0.427 Baseline vs. 0.226 Taibah; RMSE: 0.651 baseline vs. 0.476 Taibah) and Vocabulary (MSE: 1.031 Baseline vs. 0.795 Taibah; RMSE: 1.015 Baseline vs. 0.889 Taibah). These findings reinforce the system's enhanced accuracy and reliability in trait-specific scoring for Task B, particularly for dimensions that require deeper semantic understanding.

5 Error Analysis and Discussion

Figure 5 in the Appendix presents confusion-matrix heatmaps for the test set that summarize prediction errors for Task A. Across both prompts, predicted scores concentrate around a few values such as 14, 18, 24, and 28 which leads to over-scoring of low-quality essays and under-scoring of high-

System	Trait	Prompt 9			Prompt 10			Average		
		QWK	MSE	RMSE	QWK	MSE	RMSE	QWK	MSE	RMSE
Relevance	Baseline	0.127	0.514	0.717	0.182	0.340	0.584	0.155	0.427	0.651
	Taibah	0.586	0.221	0.471	0.538	0.231	0.481	**0.562**	**0.226**	**0.476**
Organization	Baseline	0.563	1.117	1.057	0.619	0.954	0.962	0.591	1.036	1.010
	Taibah	0.680	0.945	0.972	0.656	0.948	0.973	**0.668**	**0.947**	**0.973**
Vocabulary	Baseline	0.546	1.098	1.048	0.602	0.964	0.982	0.574	1.031	1.015
	Taibah	0.609	0.921	0.960	0.675	0.669	0.818	**0.642**	**0.795**	**0.889**
Style	Baseline	0.560	1.164	1.079	0.584	0.981	0.990	0.572	1.073	1.035
	Taibah	0.662	0.960	0.980	0.693	0.748	0.865	**0.678**	**0.854**	**0.923**
Development	Baseline	0.410	1.174	1.083	0.506	0.883	0.940	0.458	1.029	1.012
	Taibah	0.727	0.717	0.847	0.679	0.795	0.892	**0.703**	**0.756**	**0.870**
Mechanics	Baseline	0.421	1.345	1.160	0.468	1.212	1.101	0.445	1.279	1.131
	Taibah	0.602	1.033	1.017	0.686	0.719	0.848	**0.644**	**0.876**	**0.933**
Grammar	Baseline	0.494	1.243	1.115	0.532	1.079	1.039	0.513	1.161	1.077
	Taibah	0.629	1.036	1.018	0.699	0.721	0.849	**0.664**	**0.879**	**0.934**

Table 2: Performance comparison between Baseline and Taibah system for Task B. **Bold values** indicate superior performance.

quality essays. Predictions at the extreme values 0–2 and 30–32 are rare even when the true scores lie in those ranges. In few cases, essays with a true score of 0 receive mid-range predictions which indicate leniency toward severely deficient responses. Two factors may contribute to this: the training data may contain few or no essays labeled 0, and the scoring instruction used in the prompting specified a 1–32 range rather than 0–32 which can drive predictions away from 0. Most errors lie within ± 3 points (± 1 to ± 3 points). The Pearson correlation between human and model scores is high ($r = 0.87$), indicating overall agreement despite systematic bias. Here, we define *bias* as the signed difference between model and human scores: $\Delta = \text{model} - \text{human}$; $\Delta < 0$ indicates underestimation and $\Delta > 0$ indicates over-scoring. Essays with human scores ≥ 26 are most often underestimated. On average, the model underestimates relative to human ratings by about 0.73 points, a statistically significant difference ($t = -2.53$, $p = 0.012$). Figure 3 (Appendix) illustrates this pattern: the model is more lenient at the lower end of the scale and increasingly conservative at the upper end.

Furthermore, analysis of essay length indicates that very short essays with 0–50 words yield poor performance. Performance improves with length and peaks around 150–200 words. Beyond ≈ 300 words, the MAE increases even as QWK increases, suggesting that the system preserves ranking but tends to over-score longer texts. Very long essays that exceed 500 words show low agreement and large errors.

For Task B, the model tends to assign lower scores compared to human raters for Development, Style and Organization, with the largest mean biases in the Development (-0.254) and Style (-0.225), both highly significant ($p < 0.0001$). Vocabulary is the only trait with a small positive bias ($+0.088$, $p = 0.004$), while Mechanics shows no significant difference ($p = 0.14$). For relevance, which is scored on a scale of 0–2, the observed QWK score of 0.56 is reasonable given the narrow range. Figure 4 in the Appendix shows the mean bias ($Model - Human$) for each trait at each human score level. The model tends to over-scores the lowest-performing essays and underestimate high-scoring ones, leading to more negative bias toward the upper end of the human score scale.

6 Conclusion

This study presented our proposed system for automated Arabic essay scoring which submitted to TAQEEM 2025 shared task. The system leverages GPT-4o with a few-shot prompting methodology to evaluate the quality of Arabic essays. Our system achieved strong overall performance in both holistic scoring and trait-specific scoring tasks. For future work, we aim to enhance the system scalability and generalizability by expanding the dataset to encompass a broader range of topics and writing traits.

References

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. In *LREC 2020 Workshop Language Resources and Evaluation Conference 11–16 May 2020*, page 9.

May Bashendy, Salam Albatarni, Sohaila Eltanbouly, Walid Massoud, Houda Bouamor, and Tamer Elsayed. 2025. TAQEEM 2025: Overview of the First Shared Task for Arabic Quality Evaluation of Essays in Multi-dimensions. In *Proceedings of the Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, China.

Haiyue Feng, Sixuan Du, Gaoxia Zhu, Yan Zou, Poh Boon Phua, Yuhong Feng, Haoming Zhong, Zhiqi Shen, and Siyuan Liu. 2024. Leveraging large language models for automated chinese essay scoring. In *International Conference on Artificial Intelligence in Education*, pages 454–467. Springer.

Jonas Flodén. 2025. Grading exams using large language models: A comparison between human and ai grading of exams in higher education using chatgpt. *British educational research journal*, 51(1):201–224.

Rayed Ghazawi and Edwin Simpson. 2025. How well can llms grade essays in arabic? *arXiv preprint arXiv:2501.16516*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Gloria Ashiya Katuka, Alexander Gain, and Yen-Yun Yu. 2024. Investigating automatic scoring and feedback using large language models. *arXiv preprint arXiv:2405.00602*.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Wenchao Li and Haitao Liu. 2024. Applying large language models for automated essay scoring for non-native japanese. *Humanities and Social Sciences Communications*, 11(1):1–15.

Pei Yee Liew and Ian KT Tan. 2024. On automated essay grading using large language models. In *Proceedings of the 2024 8th International Conference on Computer Science and Artificial Intelligence*, pages 204–211.

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

Austin Pack, Alex Barrett, and Juan Escalante. 2024. Large language models and automated essay scoring of english language learner writing: Insights into validity and reliability. *Computers and Education: Artificial Intelligence*, 6:100234.

Masaki Uto and Masashi Okano. 2020. Robust neural automated essay scoring using item response theory. In *International conference on artificial intelligence in education*, pages 549–561. Springer.

Yang Yang. 2024. The reliability of using chatgpt in rating efl writings. *Shanlax International Journal of Education*, 12(4):49–59.

Fatih Yavuz, Özgür Çelik, and Gamze Yavaş Çelik. 2025. Utilizing large language models for efl essay grading: An examination of reliability and validity in rubric-based assessments. *British Journal of Educational Technology*, 56(1):150–166.

A Appendix

A.1 Structured Prompt Templates for Automated Essay Scoring

Figure 1 shows the structured prompt template used in Task A's automated essay scoring system. Figure 2 displays the structured prompt template applied in Task B's automated essay scoring system.

Prompt	English Translation
أنت معلم خبير في تقييم المقالات العربية. قيّم المقالات كما لو كنت معلّمًا على مقياس من 1 إلى 32 بناءً على المعايير التالية مع وضع النسب المناسبة لكل معيار واضف معايير اخرى تراها مناسبة: 1. **المحتوى**: وضوح وشمول الأفكار وتغطية الموضوع. 2. **اللغة**: سلامة قواعد اللغة والنحو والإملاء واستخدام المفردات. 3. **الهيكل**: وجود مقدمة وعرض وخاتمة، وترابط الأفكار باستخدام أدوات الربط. 4. **الحجج والتحليل**: قوة التفسير أو الحجج مع الأدلة حسب نوع المقال (تفسيري أو إقناعي). 5. **الأسلوب**: ملاءمة الأسلوب للغرض، وجودة علامات الترقيم، وجاذبية النص. 6. **الالتزام بعدد الكلمات**: هل عدد الكلمات يقارب 300 كلمة. {examples_text} الآن قيّم هذا المقال وفقًا للمطلب التالي: {target_prompt_text} **المقال المراد تقييمه:** {essay_text} استنادًا إلى الأمثلة السابقة والمعايير المذكورة، قدم تقييمك بصيغة JSON على النحو التالي: {{ "score": <رقم من 1 إلى 32>, "confidence": <من 0 إلى 1>, "reasoning": <شرح موجز يوضح نقاط القوة والضعف حسب المعايير مع الإشارة إلى الالتزام بعدد الكلمات> }}	You are an expert instructor in evaluating Arabic essays. Evaluate the essays as if you were a teacher on a scale from 1 to 32 based on the following criteria with appropriate weightings for each criterion, and add any other criteria you deem appropriate: **1-Content**: Clarity and comprehensiveness of ideas and topic coverage. **2-Language**: Correct grammar, syntax, spelling, and vocabulary usage. **3-Structure**: Presence of introduction, body, and conclusion, with logical flow using appropriate connectors. **4-Arguments&Analysis**: Strength of explanations/arguments with evidence, depending on essay type (expository or persuasive). **5-Style**: Appropriateness of style for the purpose, punctuation quality, and text appeal. **6-Word Count Compliance**: Whether the word count is approximately 300 words. {examples_text} Now, evaluate this essay according to the following requirement: {target_prompt_text} **Essay to be evaluated**: {essay_text} Based on the previous examples and the mentioned criteria, provide your evaluation in JSON format as follows: { "score": <number from 1 to 32>, "confidence": <0 to 1>, "reasoning": "<brief explanation highlighting strengths/weaknesses per criteria, including word count compliance>" }

Figure 1: Structured Prompt Template Applied in Task A's Automated Essay Scoring System.

Prompt	English Translation
أنت شخص خبير في تقييم المقالات العربية. قيّم المقالات كما لو كنت معلّمًا بناءً على المعايير التالية، مع الأخذ في الاعتبار أن كل معيار له وزنه الخاص: **معايير التقييم التفصيلية والدرجات القصوى لكل منها:** 1. **الملاءمة** (Relevance): (الدرجة القصوى: 2) * مدى ارتباط المقال بالموضوع المطلوب وتغطيته للجوانب الأساسية للمهمة. 2. **التنظيم** (Organization): (الدرجة القصوى: 5) * وضوح الهيكل العام للمقال، بما في ذلك وجود مقدمة واضحة، فقرات متماسكة، وخاتمة قوية. * التسلسل المنطقي للأفكار والانتقال السلس بين الفقرات. 3. **المفردات** (Vocabulary): (الدرجة القصوى: 5) * ثراء ودقة المفردات المستخدمة. * استخدام كلمات متنوعة ومناسبة للسياق، وتجنب التكرار. 4. **الأسلوب** (Style): (الدرجة القصوى: 5) * جاذبية الأسلوب ووضوحه. * استخدام علامات الترقيم بشكل صحيح وفعال. * ملاءمة النبرة والأسلوب للغرض من المقال والجمهور المستهدف. 5. **التطوير** (Development): (الدرجة القصوى: 5) * عمق الأفكار وتفصيلها. * تقديم حجج قوية وأدلة داعمة (أمثلة، شواهد، براهين) حسب نوع المقال (تفسيري أو إقناعي). * القدرة على تحليل الموضوع من جوانب متعددة. 6. **الميكانيكا** (Mechanics): (الدرجة القصوى: 5) * سلامة الإملاء وعلامات الترقيم. * صحة التنسيق العام للمقال. 7. **القواعد** (Grammar): (الدرجة القصوى: 5) * سلامة قواعد اللغة والنحو والصرف. * بناء الجمل بشكل صحيح وواضح. {examples_text} الآن قيّم هذا المقال وفقًا للمطلب التالي: {target_prompt_text} المقال المراد تقييمه: {essay_text} استنادًا إلى الأمثلة السابقة والمعايير المذكورة، قدم تقييمك بصيغة JSON على النحو التالي: {{ "relevance": <رقم من 0 إلى 2>, "organization": <رقم من 0 إلى 5>, "vocabulary": <رقم من 0 إلى 5>, "style": <رقم من 0 إلى 5>, "development": <رقم من 0 إلى 5>, "mechanics": <رقم من 0 إلى 5>, "grammar": <رقم من 0 إلى 5> }}	You are an expert in assessing Arabic essays. Evaluate essays as a teacher would, based on the following weighted criteria: **Detailed Evaluation Criteria and Maximum Scores:** **1-Relevance** (Max: 2 points) Degree to which the essay addresses the assigned topic and covers its core aspects. **2-Organization** (Max: 5 points) Clear structure (introduction, coherent paragraphs, strong conclusion). Logical flow of ideas and smooth transitions between paragraphs. **3-Vocabulary** (Max: 5 points) Richness and accuracy of word choice. Use of varied, context-appropriate terms; avoidance of repetition. **4-Style** (Max: 5 points) Engaging and clear writing style. Correct and effective punctuation. Tone/style suited to the essay's purpose and audience. **5-Development** (Max: 5 points) Depth and elaboration of ideas. Strong arguments with supporting evidence (examples, citations) based on essay type (expository/persuasive). Multifaceted analysis of the topic. **6-Mechanics** (Max: 5 points) Spelling and punctuation accuracy. Proper overall formatting. **7-Grammar** (Max: 5 points) Correct syntax, morphology, and grammar. Proper sentence structure and clarity. Review the example essays: {examples_text} Evaluate the target essay against this prompt: {target_prompt_text} Assess the following essay: {essay_text} **Submit your evaluation in JSON format:** { "relevance": <0–2>, "organization": <0–5>, "vocabulary": <0–5>, "style": <0–5>, "development": <0–5>, "mechanics": <0–5>, "grammar": <0–5>, "total_score": <sum of all criteria>, "feedback": "<concise strengths/weaknesses per criterion>" }

Figure 2: Structured Prompt Template Applied in Task B's Automated Essay Scoring System.

A.2 Bias and Performance Visualizations for Tasks A and B

Figure 3 presents a bias visualization comparing human and model holistic scores for Task A. Figure 4 shows a trait-specific bias heat map for Task B, illustrating the difference between model and human scores. Figure 5 displays confusion matrices for the testing set's holistic score prediction for Task A, with Figure 5a specifically for Prompt 9 and Figure 5b for Prompt 10.

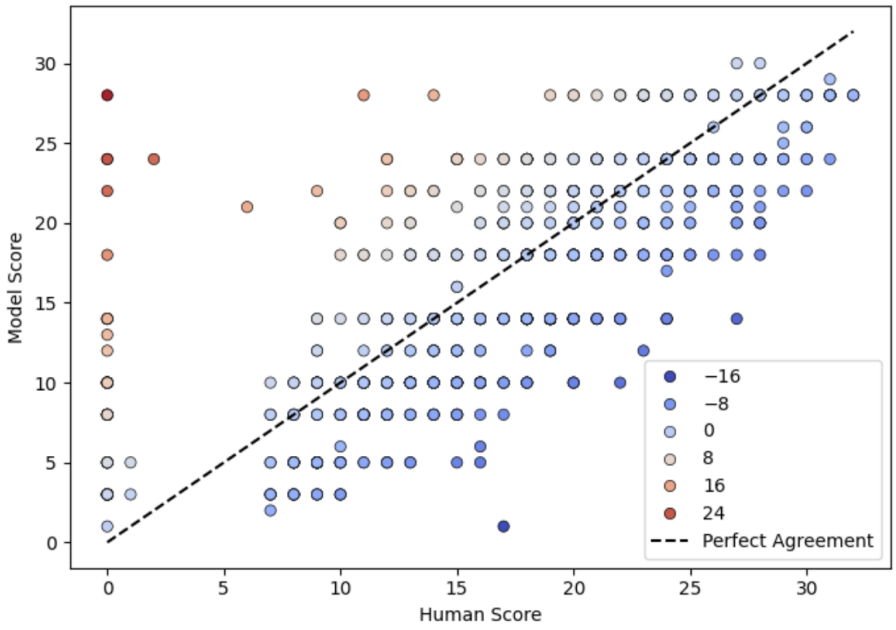


Figure 3: Bias Visualization: Human vs. Model Holistic Scores (Task A).

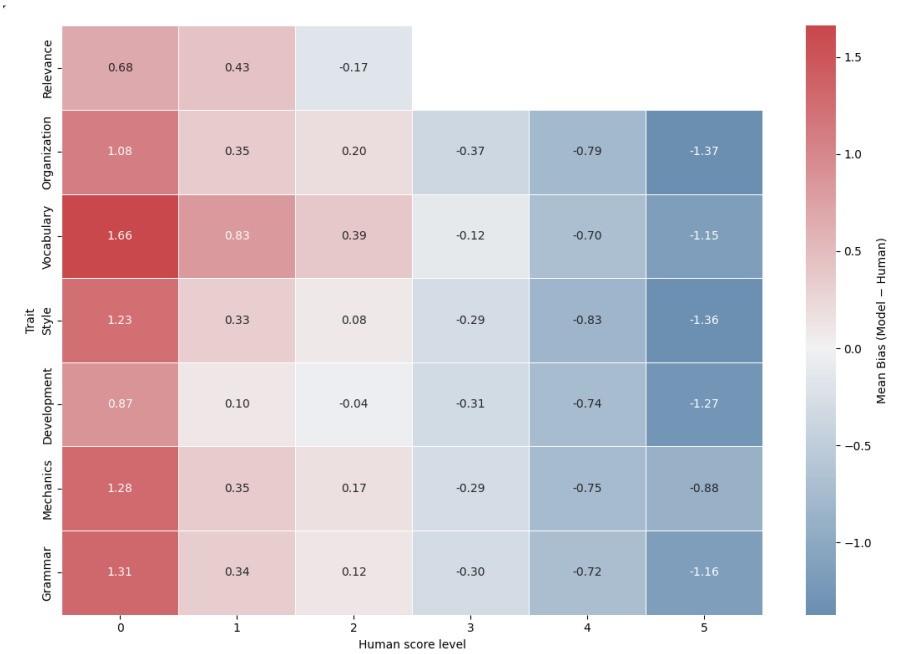
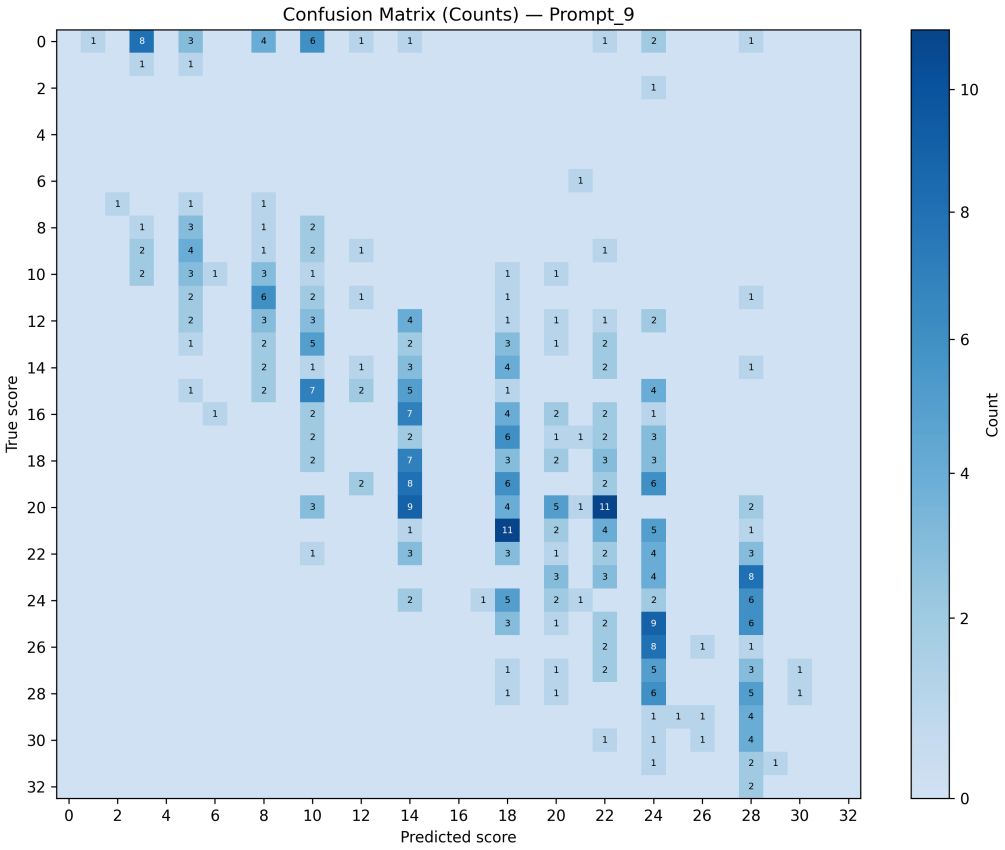
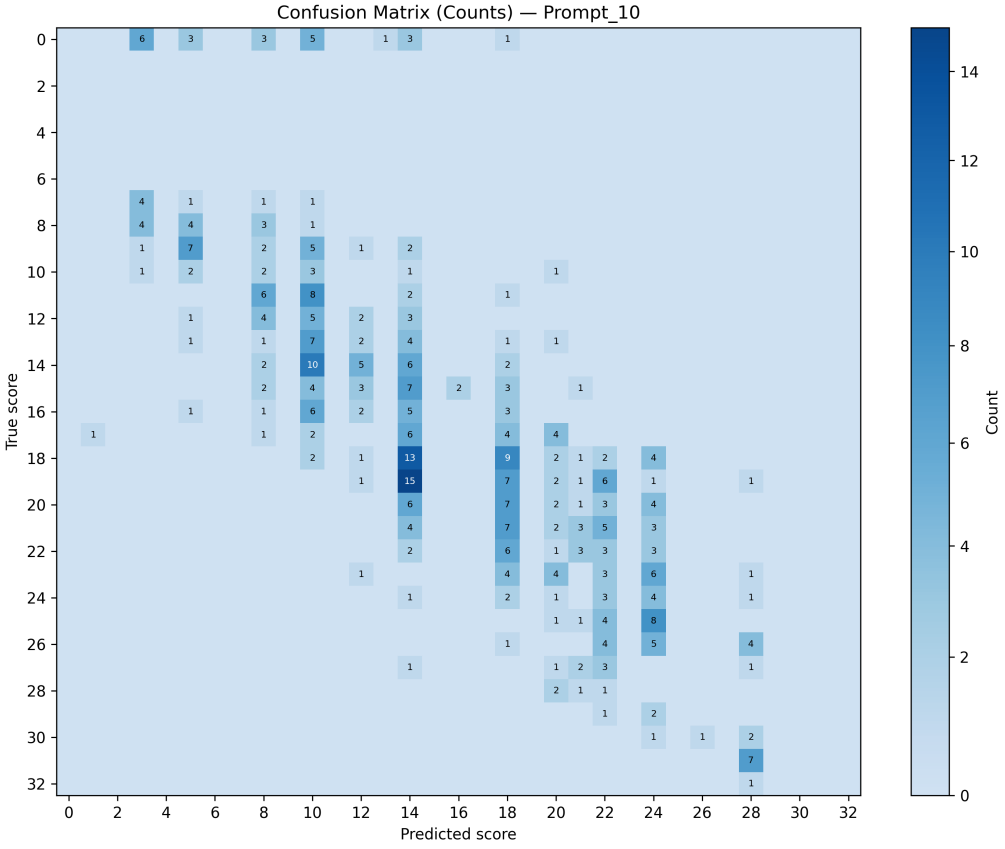


Figure 4: Task B Trait-Specific Bias Heat Map ($Model - Human$).



(a) Confusion Matrix for Testing set Holistic Score prediction For Task A: Prompt 9



(b) Confusion Matrix for Testing set Holistic Score prediction For Task A: Prompt 10

Figure 5: Confusion Matrix for Testing set Holistic Score prediction For Task A