SCORECRAFT: A LARGE LANGUAGE MODEL-BASED SYSTEM FOR AUTOMATED EXAMINATION MARKING AND FEEDBACK GENERATION

by

Dimson Ifeyinwa C[1], Nwachukwu M. M[2], Eze Ukamaka J.[3]

1,2: Department of Electronic and Computer Engineering, Nnamdi Azikiwe University, Awka, Nigeria

3: Dept of Computer Engineering, Madonna University of Nigeria, Akpugo Campus, Enugu State, Nigeria

ABSTRACT

Manual marking of examination scripts is time-consuming, prone to inconsistency, and often delays the delivery of feedback to students, particularly in large university classes. This paper presents ScoreCraft, a large language model (LLM)-based examination marking system designed to assist lecturers in evaluating student responses. The system allows lecturers to upload structured marking guides and student answer scripts, after which it generates draft scores and explanatory feedback for lecturer review and approval. The proposed system integrates marking guide ingestion, student response alignment, LLM-based scoring, feedback generation, and a human-in-the-loop moderation interface. By retaining the lecturer as the final authority over released scores, ScoreCraft is designed to support transparency, consistency, fairness, and accountability while reducing marking workload and turnaround time. The paper also presents an evaluation protocol based on agreement between system-generated scores and lecturer-assigned scores, using measures such as quadratic weighted kappa and mean absolute error. The proposed system provides a foundation for efficient and scalable examination assessment, with future extensions including multilingual assessment, handwriting recognition, and improved adaptation to diverse examination formats.

Keywords: Automated marking, Large language models, Automated essay scoring, Examination assessment, Educational technology

I. INTRODUCTION

Examination marking is one of the most time-intensive responsibilities carried by lecturers, especially in departments with large cohorts and essay-type or free-text questions that resist simple key-based automation (Ramesh & Sanampudi, 2022). Delays in returning marked scripts limit the formative value of feedback, and manual marking is vulnerable to fatigue effects and inter-rater inconsistency across large batches of scripts. Early automated essay scoring systems addressed this problem with hand-engineered linguistic features and, later, neural sequence models trained end-to-end on scored essays (Taghipour & Ng, 2016). The subsequent shift toward transformer architectures (Vaswani et al., 2017) and pretrained language representations such as BERT (Devlin et al., 2019) substantially improved contextual understanding of free-text responses. Most recently, general-purpose, instruction-following LLMs capable of following rubric instructions in natural language have renewed interest in automated marking that goes beyond pattern matching to something closer to genuine reading comprehension of student answers (Kortemeyer, 2024; Kasneci et al., 2023).

This paper describes ScoreCraft, a marking-guide-driven system in which a lecturer uploads (a) a marking guide specifying expected answer points and mark allocations, and (b) a batch of student answer scripts. The system aligns each student response to the relevant rubric item, uses an LLM to generate a score and justification, and returns both to the lecturer for review before any score is finalised. The contribution of this paper is not a claim of finished empirical validation, but a description of a practical system architecture and evaluation protocol intended for deployment in a university examination workflow, together with a positioning of the design choices against the wider AES/ASAG and educational-AI literature.

II. LITERATURE SURVEY

2.1 Classical and Neural Automated Scoring

Automated essay and short-answer scoring predates LLMs by decades, with early systems relying on hand-engineered linguistic features and statistical or shallow machine-learning classifiers (Ramesh & Sanampudi, 2022). Taghipour and Ng (2016) introduced one of the first fully neural approaches to essay scoring, using recurrent networks to learn scoring functions directly from labelled essays without manually engineered features. These systems achieve reasonable agreement with human raters on narrow, well-resourced tasks but generalise poorly to new prompts or subjects without retraining.

The introduction of the Transformer architecture (Vaswani et al., 2017) and pretrained bidirectional encoders such as BERT (Devlin et al., 2019) improved on this by learning contextual representations of text. Sung et al. (2019) showed that transformer-based pretraining materially improves short-answer grading performance relative to earlier feature-based and recurrent models, and fine-tuned scorers built on this foundation have reported strong agreement with human raters on in-

domain data (Rodrigues et al., 2024). However, fine-tuning still requires a labelled dataset per subject or rubric, which most individual lecturers do not have.

2.2 LLM-Based Grading

The introduction of instruction-following LLMs such as GPT-4 shifted attention toward zero- and few-shot grading, where a rubric is supplied directly in the prompt rather than learned from labelled examples. Kortemeyer (2024) examined GPT-4 on short-answer physics grading and found performance approaching, but not consistently matching, human raters. Yancey et al. (2023) reported that GPT-4's quadratic weighted kappa (QWK) on CEFR-scale essay rating trailed both a strong baseline AES system and human raters, particularly without calibration examples. Mizumoto and Eguchi (2023) similarly found that an early GPT model could approximate holistic essay scores but with systematic deviations from human graders. Grévisse (2024) applied LLM-based short-answer grading in an undergraduate medical education setting and reported it as a feasible aid rather than a replacement for instructor grading. Flodén (2025) evaluated ChatGPT grading of several hundred Master's-level examination responses and likewise recommended it as a support tool subject to instructor oversight. Yavuz et al. (2025) compared ChatGPT and Bard on rubric-based EFL essay grading and found reliability below that of trained human raters for both models. Naismith et al. (2023) further showed that GPT-4 can evaluate discourse coherence in ways that broadly track human judgement, suggesting LLMs are useful for qualitative, rationale-bearing feedback even where their numeric scoring remains imperfect.

2.3 Trustworthy and Explainable AI in Education

Beyond scoring accuracy, a parallel literature has examined what it takes for AI-based assessment tools to be trusted and adopted responsibly in educational settings. Kasneci et al. (2023) argue that large language models bring genuine pedagogical opportunities but require deliberate attention to their limitations and to the competencies teachers and learners need to use them responsibly. Khosravi et al. (2022) argue that explainability is a precondition for trust in AI-
based educational tools, since stakeholders are unlikely to accept opaque scoring decisions that affect grades. Khosravi et al. (2024) extend this into a roadmap for trustworthy AI in education, emphasising human oversight, transparency, and contestability of AI-generated decisions. These recommendations align closely with, and motivate, ScoreCraft's decision to expose LLM justifications and confidence indicators to lecturers rather than only a numeric score.

Taken together, this literature converges on three points relevant to ScoreCraft's design: first, LLM-based grading can meaningfully reduce marking effort and produce plausible, rubric-aligned justifications; second, none of the reviewed grading studies recommend releasing LLM-assigned scores to students without human review; and third, the broader educational-AI literature treats transparency and human oversight as preconditions for trust rather than optional add-ons. ScoreCraft's architecture is built around these points, treating the LLM as a drafting assistant rather than an autonomous marker.

III. SYSTEM OVERVIEW AND DESIGN RATIONALE

Table 1 summarises the main families of automated grading approaches identified in the literature review and the trade-offs that motivated ScoreCraft's specific design choice of rubric-grounded LLM grading with mandatory lecturer moderation, rather than either a purely statistical model or an unmoderated LLM pipeline. The full comparison appears in Table 1.

Table 1:*Comparison of Automated Marking Approaches*

Approach	Representative Studies	Strengths	Limitations
Feature-engineered / statistical AES	Ramesh & Sanampudi (2022)	Fast, interpretable, low compute cost	Poor generalisation across prompts and subjects
Fine-tuned transformer (BERT-style) scorers	Rodrigues et al. (2024)	Strong agreement with raters on trained domains	Requires large labelled datasets per subject
Zero/few-shot LLM prompting (GPT-4 class)	Kortemeyer (2024); Yancey et al. (2023)	No task-specific training data needed; generates rationale/feedback	Variable reliability; sensitive to prompt design
Rubric-grounded LLM grading with calibration	Flodén (2025); Grévisse (2024)	Aligns closely with instructor rubrics; supports free-text answers	Needs careful rubric formatting; occasional score drift

3.1 ScoreCraft Architecture

ScoreCraft is organised as a pipeline of seven modules, shown as a block diagram in Figure 1 and

summarised in Table 2. A lecturer first uploads a marking guide, which the Guide Ingestion module parses into a structured rubric: each question is broken into discrete, mark-bearing points. Student scripts are then uploaded in batch; the Script Ingestion module extracts text (including OCR for scanned scripts) and segments each script by question number.

The Alignment Engine matches each segmented response to its corresponding rubric entry, using question numbering supplemented by semantic-similarity checks to catch mislabelled or out-of-order answers. Matched pairs are passed to the LLM Scoring Core, which is prompted with the rubric item, its allocated marks, and the student's response, and asked to return a numeric score bounded by the allocated marks, a short justification referencing which rubric points were or were not addressed, and a confidence indicator. The Feedback Generator rewrites this justification into feedback phrased for the student rather than the lecturer.

Critically, none of these draft scores are released automatically. The Moderation Interface presents each question's draft score, justification, and confidence indicator to the lecturer alongside the original response, allowing rapid accept, edit, or override actions. Only lecturer-approved scores reach the Export Layer, which compiles a results sheet in a format compatible with existing departmental record-keeping. Table 2 summarises these seven modules.

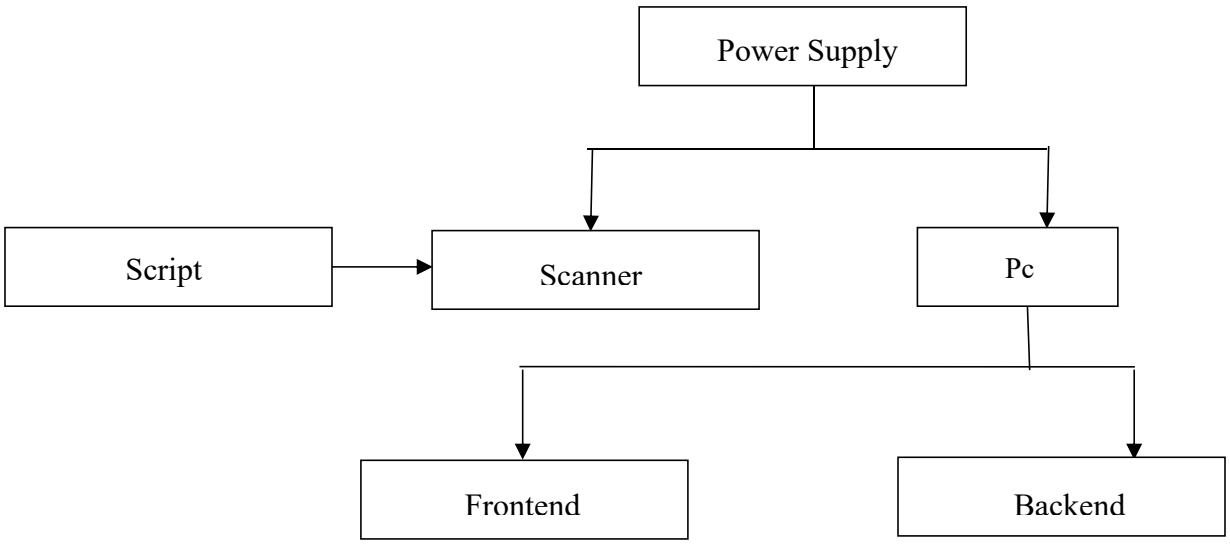

Figure 1. Block Diagram of the ScoreCraft System Pipeline

Table 2: ScriptMark System Modules

Module	Function
Guide Ingestion	Accepts lecturer-uploaded marking guides (question stems, expected points, mark allocation) and parses them into a structured rubric object.
Script Ingestion	Accepts scanned or typed student answer scripts, performs text extraction/OCR where needed, and segments responses by question number.
Alignment Engine	Maps each extracted student response to its corresponding rubric item using question-number and semantic-similarity matching.
LLM Scoring Core	Sends the rubric item and matched student response to the language model with a structured prompt requesting a numeric score, justification, and confidence indicator.
Feedback Generator	Converts the model's justification into concise, student-facing feedback highlighting missed rubric points.
Moderation Interface	Presents LLM-assigned scores and justifications to the lecturer for review, override, and final approval before scores are released.
Export Layer	Compiles final, lecturer-approved scores into a downloadable results sheet for record-keeping.

3.2 Proposed Evaluation Protocol

Because grading quality is the central concern raised throughout the literature reviewed in Section 2, we propose evaluating ScoreCraft against lecturer-assigned scores on held-out scripts using two complementary metrics. Quadratic weighted kappa (QWK) will measure ordinal agreement between LLM-drafted scores (prior to lecturer edits) and final

lecturer scores, allowing direct comparison with QWK figures reported for comparable GPT-4-class systems in the literature (Kortemeyer, 2024; Yancey et al., 2023). Mean absolute error (MAE) in raw marks will complement QWK by quantifying the average size of scoring discrepancies in terms meaningful to lecturers and students. A secondary metric, edit rate, defined as the proportion of questions where the lecturer changes the draft score during moderation, will indicate how much marking effort the system genuinely removes rather than merely relocates.

This protocol is deliberately conservative: it treats lecturer judgement as the reference standard rather than assuming the LLM's score is correct, consistent with the human-in-the-loop position argued for by Flodén (2025) and Grévisse (2024).

IV. DISCUSSION

The main design tension in systems of this kind is between marking-time savings and grading reliability. Prior work suggests LLM-only grading, without calibration or oversight, does not yet reliably match human raters, particularly for longer, more open-ended responses (Yancey et al., 2023; Yavuz et al., 2025). ScriptMark's response to this tension is architectural rather than purely algorithmic: by making the lecturer the final approver of every score, the system can absorb LLM scoring errors without those errors reaching students, while still automating the most time-consuming part of marking, namely producing a first-pass score and justification for every response.

A second consideration is subject and rubric variability. Because ScriptMark grades from a lecturer-supplied rubric at inference time rather than a model fine-tuned on historical scripts, it avoids the cold-start problem that limits fine-tuned transformer scorers (Rodrigues et al., 2024) to subjects with existing labelled data. This makes the approach more readily transferable across departments, though it depends on the lecturer producing a sufficiently explicit marking guide, which is itself a design constraint worth stating plainly rather than assuming away.

V. CONCLUSION and FUTURE WORK

This paper presented ScriptMark, an LLM-based examination marking system designed around mandatory lecturer moderation rather than autonomous scoring, and positioned its design choices against recent automated essay and short-answer grading literature. The proposed evaluation protocol, based on QWK, MAE, and edit rate against lecturer scores, offers a concrete next step for empirically validating the system in a live examination setting. Future work includes extending script ingestion to better handle handwritten scripts, evaluating performance across subjects with differing rubric styles, and investigating whether providing the LLM with a small number of lecturer-marked calibration scripts per cohort, in the spirit of Yancey et al. (2023), improves agreement without requiring full fine-tuning.

REFERENCES

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Flodén, J. (2025). Evaluating ChatGPT as a tool for grading Master's-level exam responses. *British Educational Research Journal*. Advance online publication.

Grévisse, C. (2024). LLM-based automatic short answer grading in undergraduate medical education. *BMC Medical Education, 24*(1), Article 1060. https://doi.org/10.1186/s12909-024-06026-x

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., … Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103*, Article 102274. https://doi.org/10.1016/j.lindif.2023.102274

Khosravi, H., Shum, S. B., Chen, G., Conati, C., Tsai, Y.-S., Kay, J., Knight, S., Martinez-Maldonado, R., Sadiq, S., & Gašević, D. (2022). Explainable artificial intelligence in education. *Computers and Education: Artificial Intelligence, 3*, Article 100074. https://doi.org/10.1016/j.caeai.2022.100074

Khosravi, H., Darvishi, A., Cooper, G., & Pardo, A. (2024). Trustworthy AI in education: A roadmap for ethical and effective implementation. In *Proceedings of the 13th Hellenic Conference on Artificial Intelligence* (Article 49). Association for Computing Machinery. https://doi.org/10.1145/3688671.3688781

Kortemeyer, G. (2024). Performance of the pretrained large language model GPT-4 on automated short answer grading. *Discover Artificial Intelligence, 4*(1), Article 47. https://doi.org/10.1007/s44163-024-00147-y

Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics, 2*(2), Article 100050. https://doi.org/10.1016/j.rmal.2023.100050

Naismith, B., Mulcaire, P., & Burstein, J. (2023). Automated evaluation of written discourse coherence using GPT-4. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 394–403). Association for Computational Linguistics.

Ramesh, D., & Sanampudi, S. K. (2022). An automated essay scoring systems: A systematic literature review. *Artificial Intelligence Review, 55*(3), 2495–2527. https://doi.org/10.1007/s10462-021-10068-2

Rodrigues, L., Pereira, F. D., Cabral, L., Gašević, D., Ramalho, G., & Mello, R. F. (2024). Assessing the quality of automatic-generated short answers using GPT-4. *Computers and Education: Artificial Intelligence, 7*, Article 100248. https://doi.org/10.1016/j.caeai.2024.100248

Sung, C., Dhamecha, T. I., & Mukhi, N. (2019). Improving short answer grading using transformer-based pre-training. In *Proceedings of the 20th International Conference on Artificial Intelligence in Education* (pp. 469–481). Springer. https://doi.org/10.1007/978-3-030-23204-7_39

Taghipour, K., & Ng, H. T. (2016). A neural approach to automated essay scoring. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 1882–1891). Association for Computational Linguistics. https://doi.org/10.18653/v1/D16-1193

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* (Vol. 30, pp. 5998–6008). Curran Associates.

Yancey, K. P., Laflair, G., Verardi, A., & Burstein, J. (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* (pp. 576–584). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.bea-1.49

Yavuz, F Çelik, Ö., & Yavaş Çelik, G. (2025). Utilizing large language models for EFL essay grading: An examination of reliability and validity in rubric-based assessments. *British Journal of Educational Technology*. Advance online publication. https://doi.org/10.1111/bjet.13494

