Contact: e.kapetanios@herts.ac.uk

Title: Explainable Generative Assessment Systems: Toward Trustworthy AI for Complex Educational Evaluation

Background and Rationale

The assessment of complex, technical student work is a cornerstone of higher education, yet it is profoundly resource-intensive and prone to subjectivity. The emergence of Large Language Models (LLMs) like Gemini Pro and ChatGPT has shown significant potential for automating this process. A recent, internally conducted, empirical study demonstrated that while these models cannot yet replace human judgment, they show criterion-dependent strengths and can augment assessment, particularly when the evaluation task is simplified from a granular, multi-class rubric to a binary Pass/Fail classification. Our findings, however, confirm that LLMs can augment but not replace human judgment.

Two critical barriers, however, prevent their adoption in high-stakes scientific domains like **Biocomputing or Robotics**. First is the **"black box" problem**: their decision-making processes are opaque, making it difficult to trust their outputs without transparency. Second, and more dangerously, is the risk of **misinformation**. Trained on vast internet corpora, LLMs can absorb and reproduce discredited findings or flawed data, posing a systemic threat to scientific integrity. An AI that confidently presents falsehoods is worse than one that is simply incorrect.

This project directly confronts these challenges by proposing a novel, **explainable** hybrid AI framework designed not only to assist in assessment but to actively **verify the veracity** of scientific claims and make its reasoning transparent.

Contribution to Knowledge:

This project will make the following contributions:

1. Empirical: It will provide the first robust benchmark of a GenAI framework for the dual purpose of assessment and misinformation detection in a specialized scientific domain.
2. Methodological: It will pioneer a novel, interpretable hybrid RAG-LLM framework, addressing the critical need for transparency and factual grounding in scientific AI systems.
3. Explainable AI: It will contribute new methods for generating and evaluating explanations for complex, text-based reasoning and verification tasks, a key challenge in the field of XAI.
4. Practical: The framework will offer a tangible, trustworthy tool to augment the work of academic supervisors and peer reviewers, helping to manage workload while upholding rigorous standards of scientific inquiry.

Broader Impact: Transforming Research and Combating Misinformation - While focused on biocomputing, the outcomes of this research will have far-reaching implications. The principles and methodologies developed will be generalisable to other technical and scientific domains.